Machine-learning-accelerated Bose-Einstein condensation

Zachary Vendeiro, Joshua Ramette, Alyssa Rudelis, Michelle Chong, Josiah Sinclair, Luke Stewart, Alban Urvoy,* and Vladan Vuletić†

Department of Physics, MIT-Harvard Center for Ultracold Atoms and Research Laboratory of Electronics, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA

(Received 17 May 2022; accepted 21 November 2022; published 27 December 2022)

Machine learning is emerging as a technology that can enhance physics experiment execution and data analysis. Here, we apply machine learning to accelerate the production of a Bose-Einstein condensate (BEC) of ^{87}Rb atoms by Bayesian optimization of up to 55 control parameters. This approach enables us to prepare BECs of 2.8×10^3 optically trapped ^{87}Rb atoms from a room-temperature gas in 575 ms. The algorithm achieves the fast BEC preparation by applying highly efficient Raman cooling to near quantum degeneracy, followed by a brief final evaporation. We anticipate that many other physics experiments with complex nonlinear system dynamics can be significantly enhanced by a similar machine-learning approach.

DOI: 10.1103/PhysRevResearch.4.043216

Recently, researchers have begun applying machine-learning techniques to atomic physics experiments, e.g., to enhance data processing for imaging [1–5], determine the ground state and dynamics of many-body systems [6,7], or to identify phases and phase transitions [8–10]. One promising practical application of machine learning to atomic physics is in the optimization of control sequences with many parameters and nonlinear dynamics [11–16], and in particular to one of the workhorses of atomic physics, Bose-Einstein condensates (BECs) [11,13–16].

With few exceptions [17], experiments on BECs end with a destructive measurement, which requires repeated BEC preparation. Approaches to increase the BEC production rate, and associated signal-to-noise ratio of the experiments, have generally relied heavily on hardware improvements [18–22] or have used atomic species with narrower optical transitions [18,21,22] than offered by the most widely utilized alkali-metal atoms. For alkali-metal atoms, the tight confinement of atom-chip magnetic traps has enabled fast evaporation sequences, with a complex multilayer atom-chip achieving BEC preparation times of 850 ms for 4×10^4 atoms [19]. Non-alkali-metal atoms featuring narrow optical transitions can be used to reach lower temperatures in narrow-line magneto-optical traps (MOTs) [18,21,22]. That approach, combined with a dynamically tunable optical dipole trap, has recently been used to prepare BECs of 2×10^4 erbium atoms in under 700 ms [22].

In this article, we demonstrate a complementary approach where, in a simple experimental setup with a broad-line MOT for a standard alkali-metal atom, machine learning is leveraged to optimize a complex nonlinear laser and evaporative cooling process to quantum degeneracy. Controlling a sequence with up to 55 interdependent experimental parameters, Bayesian optimization [11,12,23] finds parameter values which cool a gas from room temperature into the quantum degenerate regime in 575 ms, creating a BEC containing $N_{\mathrm{BEC}} = 2.8 \times 10^3$ atoms.

We identify some of the physical strategies discovered by the algorithm and also investigate how the choice of cost function impacts the trade-off between final atom number and the purity of the created BEC.

Our apparatus employs only a single MOT directly loaded from a ^{87}Rb background vapor, a crossed optical dipole trap, and two Raman cooling beams as depicted in Fig. 1(a). No Zeeman slower, two-dimensional MOT, atom chip [19], dynamic trap shaping [21], or strobing [22,24] are necessary. Using Raman cooling in a crossed optical dipole trap (cODT), a method that can reach very high phase-space density and even condensation [25], the algorithm achieves a cooling slope of 16 orders of magnitude improvement in phase space density (PSD) per order of magnitude in atom loss ($\gamma = 16$) up to the threshold to quantum degeneracy. This is significantly better than the $\gamma = 7$ value we could obtain with extensive manual optimization under similar conditions [25].

I. ATOMIC PHYSICS METHODS

The Raman cooling implementation used in this work is similar to that of Ref. [25]. Cooling proceeds in a cODT formed by intersecting two noninterfering 1064-nm beams, one horizontal and one vertical [see Fig. 1(a)]. Two 795-nm beams drive the Raman cooling: the optical pumping

*Current affiliation: Laboratoire Kastler Brossel, Sorbonne Université, CNRS, ENS-Université PSL, Collège de France, 4 Place Jussieu, 75005 Paris, France.

†vuletic@mit.edu

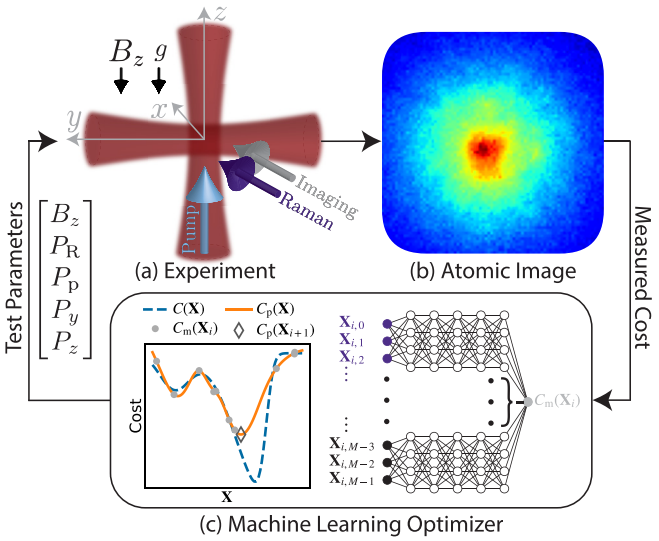

FIG. 1. (a) Setup showing 1064-nm horizontal (waist $w_h = 18\ \mu$m, beam slightly tilted downward) and vertical ($w_v = 14\ \mu$m) optical-trapping, 795-nm Raman coupling ($w_R = 500\ \mu$m) and optical pumping ($w_x = 30\ \mu$m, $w_y \approx 1$ mm), and 780-nm absorption-imaging beams. (b) Absorption image used to extract the cost function for a set of parameter values $\mathbf{X}$. (c) Bayesian optimization with a neural network. The model $C_p(\mathbf{X})$ (orange solid line) attempts to predict the actual system performance $C(\mathbf{X})$ (blue dashed line). The algorithm uses the model to predict optimal parameter values $\mathbf{X}_{i+1}$ (open diamond), tests those values, and performs a new iteration with an updated model.

beam and the Raman coupling beam. Raman cooling [26] provides sub-Doppler cooling by driving velocity-selective Raman transitions between hyperfine states, here the $|F = 2, m_F = -2\rangle$ and $|2, -1\rangle$ states of ^{87}Rb [25]. The Raman transitions are nondissipative so entropy is removed from the atomic gas in the form of spontaneously scattered photons as atoms are optically pumped back to the dark state $|2, -2\rangle$. Light-assisted collisions, which typically prohibit laser cooling at high atomic densities, are suppressed by detuning the optical pumping light 4.33 GHz to the red of the D_1 $F = 2 \rightarrow F' = 2'$ transition, where a local minimum of light-induced loss was observed [25].

The cooling dynamics are controlled via five actuators: (i) the horizontal P_y and (ii) vertical P_z trap beam powers which set the trap depth and frequencies, (iii) the Raman coupling beam power P_R which tunes the Raman rate, (iv) the power P_p of the optical pumping beam which sets the optical-pumping rate (and also Raman rate), and (v) the magnetic field B_z which adjusts the resonant velocity class for the Raman transition. The cooling procedure is divided into stages during which the controls are linearly ramped, with the endpoints of each ramp constituting the optimization parameters.

II. OPTIMIZATION SCHEME

The optimization problem can be formulated as the minimization of a cost function C, which maps a set of parameter values $\mathbf{X} \in \mathbb{R}^M$ to a corresponding cost value $C(\mathbf{X}) \in \mathbb{R}$, where M is the number of optimization parameters. The cost C quantifies the results and is generally *a priori* unknown, but can be extracted from measurements. Bayesian optimization is well-suited for this type of problem as it can tolerate noise in the measured cost and typically requires testing fewer values of $\mathbf{X}$ than other optimization methods [11–16].

Bayesian optimization begins with collecting a training dataset by experimentally measuring the cost $C_m(\mathbf{X}_i)$ for various values of sets of parameter values $\mathbf{X}_i$. The $\mathbf{X}_i$ values used to construct the training dataset are chosen by a training algorithm, which can implement another optimization algorithm or can select $\mathbf{X}_i$ randomly. A model of the cost function is then fit to the training dataset which approximates the unknown true cost function $C(\mathbf{X})$. Although Bayesian optimization typically uses a Gaussian process for its model [23], the present work uses neural networks [12,27], which were chosen for their significantly faster fitting time for our typical number of optimization parameters. Once the model is fit, a standard numerical optimization algorithm is applied to the modeled cost function $C_p(\mathbf{X})$ to determine which value $\mathbf{X}_{i+1}$ for the next iteration is predicted to yield the minimal cost, as depicted in Fig. 1(c). Optionally this numerical optimization can be constrained to a trust region (a smaller volume of parameter space centered around the $\mathbf{X}_i$ which yielded the best cost measured thus far). The predicted optimal value $\mathbf{X}_{i+1}$ is then tested by experimentally measuring the corresponding cost $C_m(\mathbf{X}_{i+1})$. The next iteration begins by retraining the model with the new result, and making a new prediction for the optimal value of $\mathbf{X}$ with the updated model. The algorithm iterates until it reaches a termination criterion, such as a set maximum number of iterations or a set number of consecutive iterations that fail to return better results. All optimization in this work was performed with the open-source packages M-LOOP [11,12] to implement the Bayesian optimization and LABSCRIPT [28] for experimental control. Additional implementation details are included in Appendix A.

III. COST FUNCTION

Since the optimization transitions the gas from the classical into the quantum degenerate regime, the final state of the gas depends strongly on how the cost function is chosen as a combination of the two experimentally accessible parameters: atom number N and temperature T. The classical phase space density $\mathrm{PSD_c}$ is defined as $\mathrm{PSD_c} \equiv n_{cp}\lambda_{dB}^3$, where λ_{dB} is the thermal de Broglie wavelength and n_{cp} is the calculated peak number density neglecting bosonic statistics (see Appendix B for calculation details). The value of $\mathrm{PSD_c}$ is nearly equal to the true PSD when $\mathrm{PSD} \ll 1$, while at the threshold to condensation $\mathrm{PSD_c} \sim 1$. Since the temperature T is more difficult to determine in the quantum degenerate regime, and also requires a fit to the data with potential convergence problems, we instead measure N and the peak optical depth (OD) in an absorption image. Generally ensembles with larger $\mathrm{PSD_c}$ have a larger atom number N and less expansion energy, which leads to a larger peak OD for a given N. Guided by this, we explored cost functions of the form

$$C(\mathbf{X}) \propto -f(N/N_1)\mathrm{OD}^3 N^{\alpha-9/5}, \tag{1}$$

where $f(N/N_1)$ is a smoothed Heaviside step function with N_1 chosen near the detection noise floor (see Appendix A).

The parameter α in the cost function tunes the trade-off between optimizing for larger atom number or lower temperature. For a pure BEC after sufficient time-of-flight (TOF) expansion, $|C/f|$ scales as $(N_{\mathrm{BEC}})^{\alpha}$ (see Appendix C). For a thermal cloud, $|C/f|$ is proportional to $\mathrm{PSD_c}$ when $\alpha = -1/5$, although that value of α is unsuitable for condensation as increasing the atom number in the BEC requires $\alpha > 0$.

IV. OPTIMIZATION PROCEDURE

The sequence begins with a separately optimized 99-ms-long MOT loading and compression period. The trap beam powers are ramped to their initial Raman cooling values during the last 10 ms of the MOT compression and then the magnetic field is adjusted to its initial Raman cooling value in 1 ms, at which point the horizontal dipole trap holds typically $N = 2.7 \times 10^5$ atoms. We then added 100-ms stages of Raman cooling one by one and optimized them individually. After five stages, the algorithm tended to turn off the Raman cooling by turning down P_{p} or P_{R} or by tuning the magnetic field B_z such that the Raman transition became off-resonant. We then added up to six shorter 30-ms-long stages in which the optical pumping and Raman coupling beams were turned off, and the algorithm performed evaporative cooling. Due to the reduced number of parameters, we were able to optimize the evaporation stages simultaneously, which produced a BEC. Subsequently we shortened the Raman cooling and evaporation stages with parameter values fixed until only a small and impure BEC was produced, and then we ran a global reoptimization. In this global optimization stage, all 42 of the Raman cooling and evaporation parameters were reoptimized simultaneously using the previous values as the initial guess for $\mathbf{X}$. Often a trust region set to one-tenth of the allowed range for each parameter was used. This kept the optimizer focused in regions of parameter space which produced a measurable signal, as adjusting even a single parameter too far would often result in the loss of all atoms. We repeated this sequence shortening and reoptimization procedure until the algorithm failed to find parameters that could produce sufficiently pure BECs.

The required beam powers generally varied over several orders of magnitude, so the logarithms of their powers were used as entries in $\mathbf{X}$, while the magnetic-field control parameter B_z was kept a linear parameter. A feedforward adjustment was included in the B_z control values to account for the light shift of the $|2, -1\rangle$ state by the optical pumping beam. We averaged over five repetitions of the experiment for each set of parameter values tested. The number of iterations per optimization varied but was typically ∼1000 (including the initial training) and required several hours, both for the single-stage optimizations and the full-sequence optimizations. A simpler optimization procedure was also attempted which did not involve optimizations of individual stages. Instead the sequence was divided into ten 100-ms stages and all 55 parameters were optimized from scratch simultaneously. That approach combined with the shortening and reoptimizing procedure successfully produced a similar BEC, albeit in slightly longer time (650 vs 575 ms), possibly due to the optimization becoming trapped in a local optimum (see Appendix A for further discussion).

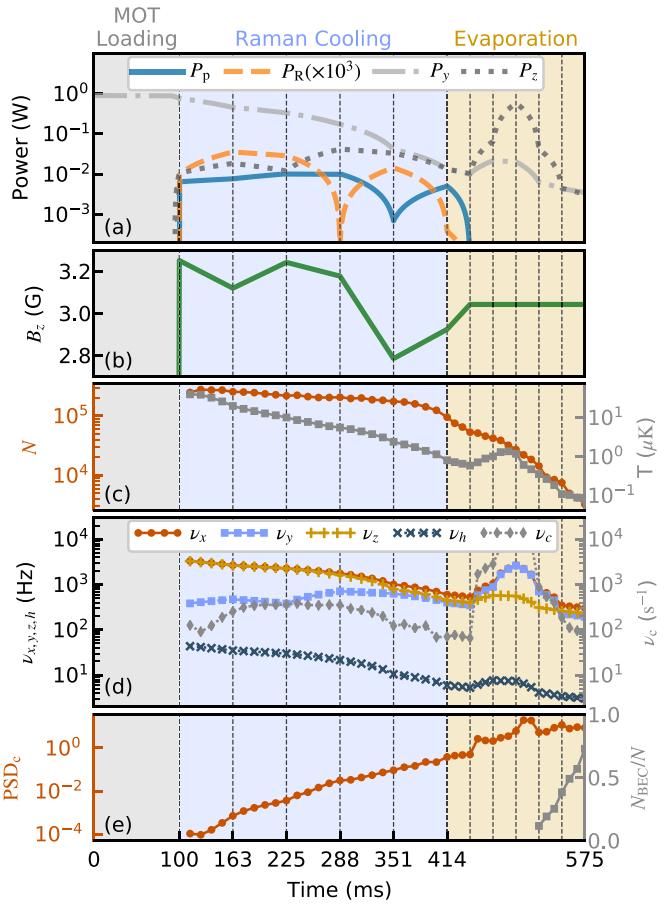

FIG. 2. Control wave forms [panels (a) and (b)] and measured trap and atomic-gas properties [panels (c)–(e)] of the optimized sequence. Gray, blue, and orange shadings mark the MOT loading, Raman cooling, and evaporation periods, respectively. The Raman beam power has been multiplied by 10^3 for better visibility. ν_x, ν_y, ν_z, and ν_h are the trap vibration frequencies in the x, y, and z directions and in the horizontal trap, respectively; ν_{c} is the atomic collision rate. $\mathrm{PSD_c}$ does not account for bosonic statistics and changes slowly while the BEC forms quickly above threshold. Calculations assume thermal equilibrium.

V. RESULTS AND PHYSICAL INTERPRETATION

The best discovered 575-ms-long control sequence and corresponding results are depicted in Figs. 2 and 3. Notably, the algorithm discovered gray molasses [29,30] in the MOT phase, which it applies at the end of the compression sequence. This outperforms the bright molasses [31,32] that was previously used in the manually optimized compression sequence, with the gray molasses loading a similar number of atoms ten times faster. After the MOT loading stage and transfer into the cODT, five ∼63-ms-long stages of Raman cooling follow, and then the optical pumping and Raman beams are ramped off, followed by six ∼27-ms-long evaporation stages. As observed in previous work [12–14], the ramps produced by Bayesian optimization are nonmonotonic and appear nonintuitive, but they outperform the routines we found by manual optimization. A reason for the nonmonotonic wave forms may be that the cost function includes many local minima. The optimization can settle into any one of these local optima

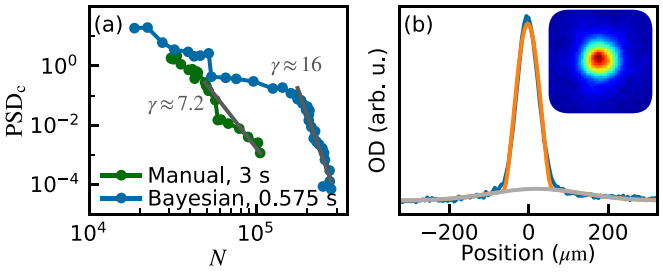

FIG. 3. Results of the 575-ms optimized sequence. (a) $\mathrm{PSD_c}$ vs atom number N. Initial cooling until $\mathrm{PSD_c} \sim 10^{-1}$ is very efficient with $\gamma \approx 16$ (gray line). The performance of the much slower (3-s-long) manually optimized sequence of Ref. [25] is shown for comparison ($\gamma \approx 7$). (b) Cross section of 24-ms TOF image (inset) shows a BEC (orange fit) with small thermal wings.

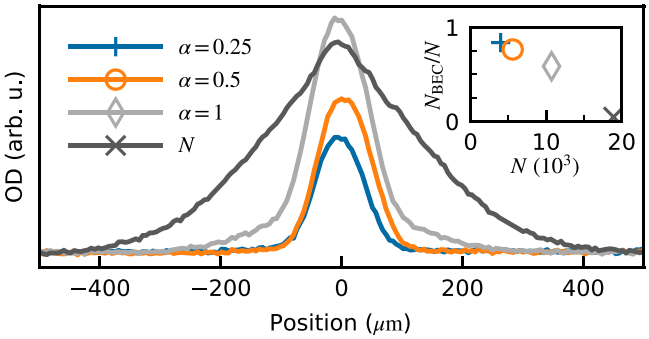

FIG. 4. Cross sections of 24-ms TOF images (200 averages) optimized for different values of the cost function parameter α (see main text) with 1-s-long sequences, demonstrating the trade-off between optimizing for atom number or temperature. Also plotted are the results of optimizing for atom number N only. Inset: Condensate fraction N_{BEC}/N vs N for different α's.

randomly and produce complex but specific wave forms, as observed in Ref. [12]. Despite the nonmonotonic ramps, $\mathrm{PSD_c}$ increases smoothly exponentially during this part of the sequence [Fig. 2(e)], due in part to the finite thermalization rate.

By shortening the sequence we are asking the algorithm to maximize the cooling speed, which is limited by the lower of the collisional rate ν_c and the trap vibration frequencies $\nu_{x,y,z}$ [33]. When the gas is still hot, we have $\nu_c \ll \nu_{x,y,z}$, and the algorithm employs Raman cooling to increase the density and collision rate [Fig. 2(d)]. However, when ν_c approaches the lowest trap vibration frequency ν_y near the time $t = 225$ ms, the algorithm starts to reduce the Raman rate, and a little later the optical pumping rate, in order to reduce light-induced collisions that scale with ν_c, rather than the trap vibration frequency. Subsequently, for times $t > 225$ ms, the cooling proceeds near optimally, with the collision rate close to, but a little below, the trap vibration frequencies. Furthermore, as the system approaches condensation near $t = 410$ ms, the collision rate is somewhat lowered to reduce light-induced atom loss [Fig. 2(c)].

Another effect limiting the cooling speed is the loading of the atoms from the single horizontal trap, in which the sample is initially prepared, into the crossed dipole trap (see the movie in the Supplemental Material [34]). Initially, the vertical-beam power P_z is held low to avoid creating a high-density dimple region which would lead to excess loss during Raman cooling. Later, P_z is ramped up to gather atoms from the horizontal trap beam into the overlap region of the cODT in order to increase the collision rate and speed up evaporative cooling. The relatively sudden ramping of the trap power up and then back down visible in Fig. 2(a) likely involves an optimal-control-like process since the trap compression and relaxation are faster than the axial period of the horizontal trap of ∼200 ms.

The optimization tended to turn off the Raman cooling after five stages because the cloud temperature T was below the effective recoil temperature [25] where Raman cooling, even with optimal parameters, becomes too slow, while leading to trap loss and heating due to light-assisted collisions [35]. The Bayesian optimization recognized this and shut down the Raman cooling at this point, with the atomic gas close to condensation. Subsequently, at higher compression, which is primarily achieved by increasing the vertical beam power, the horizontal trap power is reduced and atoms are efficiently evaporated along the direction of gravity in the tilted potential [20] (see the movie in the Supplemental Material [34]). Note also that, once the atoms have been loaded into the crossed-trap region (after $t = 350$ ms), the algorithm makes all trap vibration frequencies similar, which provides the fastest overall thermalization and hence the fastest cooling speed.

The BEC is fully prepared at the end of the evaporation stages, 575 ms after the start of the MOT loading. The final cloud contains 3.7×10^3 total atoms and is shown in Fig. 3(b). A bimodal fit of the cloud indicates that 2.8×10^3 atoms (76%) are in the BEC. Although the sequence was optimized for speed rather than efficiency, the initial cooling occurs with a logarithmic slope $\gamma = d(\log \mathrm{PSD_c})/d(\log N) \approx 16$.

VI. COST FUNCTION IMPACT

The atomic gases produced by sequences optimized for different values of α are presented in Fig. 4, as well as the results when optimizing for total atom number (N). Larger values of α result in more atoms, but at higher temperature and lower condensate fraction, while smaller values of α produce purer BECs, but with fewer atoms overall. Setting α to 0.5 was found to make a reasonable compromise (orange curve in Fig. 4); so this value was used for the final full-sequence optimization which yielded the data presented in Fig. 2.

VII. OUTLOOK

In conclusion, we have demonstrated that Raman cooling with far-detuned optical-pumping light combined with a final evaporation can rapidly produce BECs with a comparatively simple apparatus, even with a standard alkali-metal atom which lacks narrow optical transitions. Bayesian optimization greatly eased the search for a short sequence to BEC, quickly discovering initially unintuitive yet high-performing sequences. Inspection of the parameters chosen by the algorithm reveals several physical strategies, such as adjusting a collision rate close to, but below, the trap vibration frequencies to maximize the thermalization and cooling speed while minimizing density-dependent atom loss, nonadiabatic loading

into the crossed-trap dimple, and creating a nearly isotropic trap for efficient evaporation. In future applications, faster condensation can likely be achieved by including dynamical tuning of the trap size [21], while user intervention may be further reduced by factoring the sequence length into the assigned cost [14]. We anticipate that many other experimental procedures in atomic physics and beyond can be improved by machine learning.

ACKNOWLEDGMENTS

The authors would like to thank Martin Zwierlein for inspiring physics discussions and Michael Hush, Harry Slatyer, Philip Starkey, Christopher Billington, and Russell Anderson for stimulating discussions and software assistance. This work was in part supported by the NSF (Grant No. PHY-1505862), NSF CUA (Grant No. PHY-1734011), NASA (Grant No. RSA No. 1608107), DoE (Research Subcontract No. 7571809), and MURI through AFOSR (Grant No. FA9550-16-1-0323).

APPENDIX A: BAYESIAN OPTIMIZATION IMPLEMENTATION

In M-LOOP's implementation of Bayesian optimization, the training algorithm used to pick parameters and generate a training dataset is also run periodically even after the training dataset is complete [11,12]. In particular, once sufficient training data are acquired, three independent neural networks are trained. Each neural net is fully connected and consists of an input layer with one node for each optimization parameter, followed by five hidden layers with 64 nodes each and then an output layer with a single node. Once the training has completed, each neural network is used to generate a set of parameter values $\mathbf{X}$ which it predicts to be optimal, and each of those three $\mathbf{X}$'s is experimentally tested. Then another iteration of the training algorithm is performed and the $\mathbf{X}$ value it suggests is also tested. The results from all four of these measurements are included in the next training of the neural nets for the subsequent Bayesian optimization iteration. The additional iterations of the training algorithm are intended to encourage parameter space exploration and provide unbiased data [11,12].

In this work, the absorption images used to measure the cost function were generally taken after 1.5 to 8 ms of time-of-flight (TOF) expansion. We averaged over five repetitions of the experiment for each set of parameter values tested, which took $\sim$10 s accounting for experimental and analysis overhead. Simply taking the largest optical depth measured in any single pixel of an absorption image as the OD makes it prone to noise, so the OD was set to the average OD of several pixels with the largest OD to reduce noise. To compare different sequences on an equal footing during optimizations, the trap beams were always ramped to a fixed power setting before releasing the atoms for TOF imaging. This final fixed ramp is only necessary during optimizations and is omitted from the sequence once the optimizations are complete. The smoothed Heaviside step function $f(N/N_1)$ included in the cost function ensures that the cost does not diverge at low N, while having little effect when N is above the measurement noise floor. The form of $f(N/N_1)$ is inspired by the expression for the excited state population of a two-level system in thermal equilibrium and it is defined as

$$f(N/N_1) = \begin{cases} \left(\frac{2}{e^{N_1/N}+1}\right) & N > 0, \\ 0 & N \leqslant 0. \end{cases} \tag{A1}$$

For many of the optimizations in this work, particularly those with tens of parameters, the cost function landscape is "sparse" in the sense that most sets of parameter values yield poor results with a signal below the measurement noise floor. Thus the actual performance for such $\mathbf{X}$ cannot be measured, and testing them provides little information to the model. This leads to large regions of parameter space where there is no measurable signal and the direction towards better values cannot be inferred. There are two notable consequences of this. First, for such optimizations it is generally necessary to provide initial values to the optimization which give a nonzero signal. Without a good starting point, the training dataset will often only include measurements dominated by noise, making it exceedingly unlikely for the Bayesian optimization to succeed. Second, for such optimizations it is generally helpful to specify a trust region. This limits the extent of excursions as the optimizer explores parameter space, making it more likely to test parameter values which yield a measurable signal. However, this does come at the cost that it makes it less likely for the optimizer to jump from one local minimum to another better minimum. We often performed the same optimization with and without a trust region in parallel. This could be done without significantly extending the duration of optimizations because the analysis for each iteration typically took longer than the time required to perform the experiment. Thus one optimization could run experiments while the other analyzed its most recent results. For optimizations with many parameters, the results with a trust region were typically as good as or better than those without. This is likely a consequence of that fact that, given the sparsity of the cost function landscape, it is unlikely for the optimizer to discover another local optimum. Thus, it is better for the optimizer to focus on modeling the region of parameter space around the local optimum rather than fruitlessly searching for another local optimum.

The sparsity of the cost landscape and necessity for providing initial parameters which produce measurable results posed a difficulty when we optimized an entire sequence from scratch at once (rather than initially adding one cooling stage at a time). We resolved this by reducing the time of flight to 1.5 ms for the first optimization. With such a short time of flight, even poor parameter values could produce clouds with a peak optical depth above the measurement noise floor. Due to the finite dynamic range of the absorption imaging, the results of this first optimization produced a cloud which saturated the measurement and thus made it impossible to accurately quantify performance for the best-performing values. The next optimizations were performed with the same sequence duration, but the time of flight increased to 5 ms and then to 8 ms. This made it possible to better discern differences between high-performing sets of parameter values at the cost of increasing the performance required to produce a signal above the noise floor and thus increasing the sparsity of the cost function landscape. The procedure of shortening and then reoptimizing the sequence was then applied, resulting in the

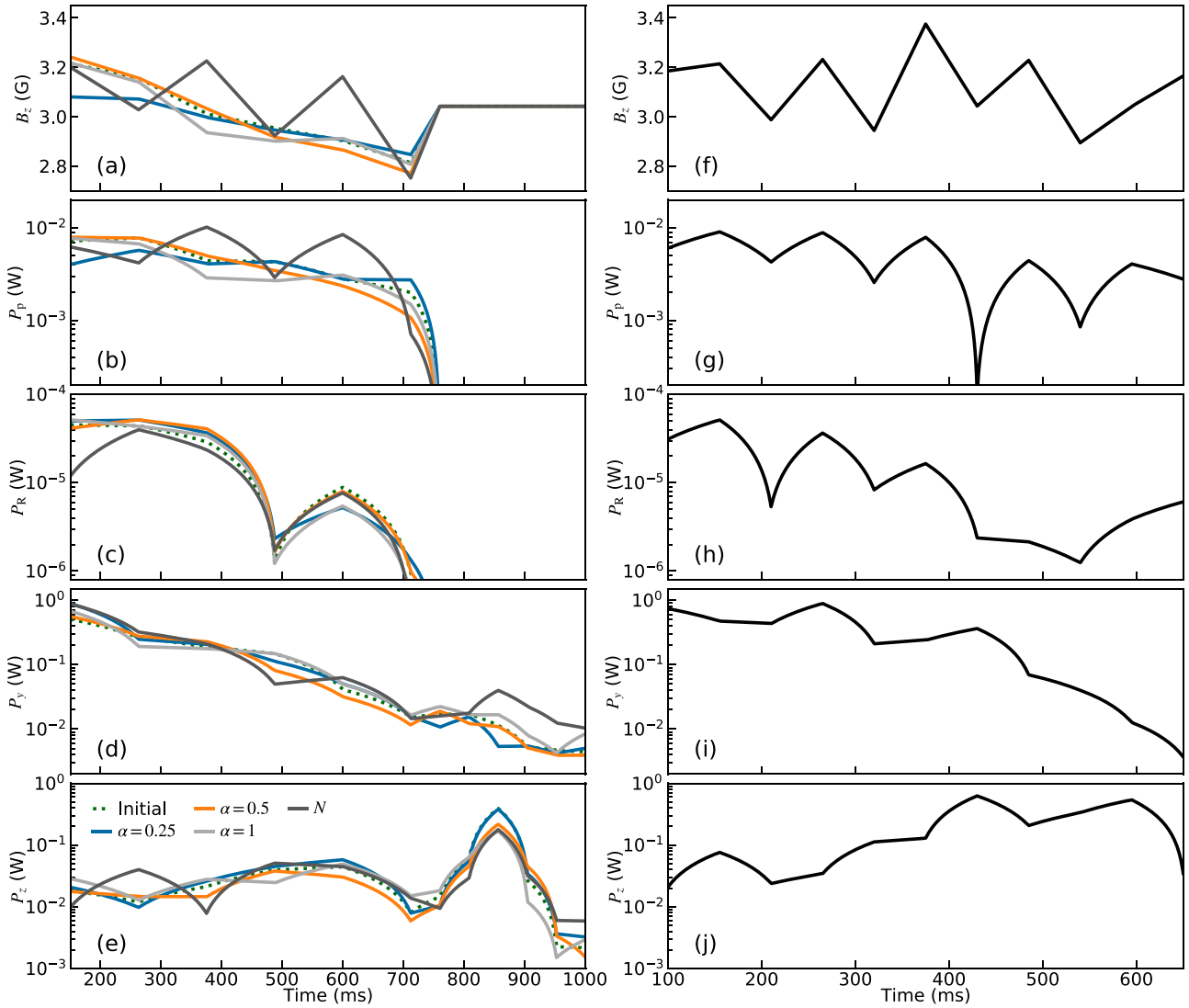

FIG. 5. (a) The control wave forms for the 1-s sequences corresponding to the results presented in Fig. 4, as well as the initial wave form used as the starting point for each of those optimizations. (b) The control wave forms for the 650-ms ten-stage sequence optimized from scratch rather than stage-by-stage initially, which was optimized with $\alpha = 0.5$. Note the differing limits for the x axes between panels (a) and (b). The sequences in panel (a) include a 50-ms magnetic coil ramp duration which was later reduced to 1 ms. The MOT sections of the sequences have been omitted for simplicity. Note that the wave forms in panel (a) are mostly qualitatively similar despite being optimized for different cost functions. The wave forms in panel (b) are more qualitatively distinct from those in panel (a), even for the orange curves which were also optimized with $\alpha = 0.5$. This suggests that the independent optimizations likely become trapped around disparate local minima. On the other hand, tuning the cost function while providing the same initial parameter values each time typically causes only smaller deviations around the initial values.

sequence presented in Fig. 5(b), which produced a BEC in 650 ms. The parameter α was set to 0.5 throughout this procedure.

Although it is not strictly fair to do so due to the differing parametrizations, it is still informative to compare the control wave forms of the independently optimized 650-ms sequence to those from Fig. 4. These wave forms are presented in Fig. 5. The sequences of Fig. 4 optimized for different α's all had fairly similar wave forms. On the other hand, the 650-ms sequence had a qualitatively different wave form. For example, it lacks the sudden rise and drop in vertical trap power towards the end of the sequence present in the other wave forms. This suggests that it has converged to a qualitatively different local optimum. On the other hand, the sequences of Fig. 4 were all optimized with a trust region and the same initial $\mathbf{X}$. Thus, those optimizations primarily performed a local search, only slightly tuning $\mathbf{X}$ to tailor the sequence for their particular value of α. Although there are small differences in parametrization, the fact that two different optimizations with the same value of α produce sequences that differ more than optimizations with the same initial $\mathbf{X}$ but different α's supports the notion that the cost function landscape includes multiple local minima, as suggested in the main text.

APPENDIX B: CALCULATIONS OF ATOMIC GAS PROPERTIES

The classical phase-space density is defined as $\mathrm{PSD_c} = n_{\mathrm{cp}}\lambda_{\mathrm{dB}}^3$, where n_{cp} is the peak number density calculated for a classical gas (i.e., neglecting Bosonic statistics) and $\lambda_{\mathrm{dB}} = h/\sqrt{2\pi m k_{\mathrm{B}} T}$ is the de Broglie wavelength. Here h is the Planck constant, m is the mass of an atom, and k_{B} is the Boltzmann constant. To calculate $\mathrm{PSD_c}$ for a cloud, its atom number N and temperature T are measured and it is assumed to be in thermal equilibrium. The value of λ_{dB} is easily calculated from the measured temperature. The partition function $Z = \int f_{\mathrm{B}}(\mathbf{x})dV$ is then calculated by numerically integrating the Boltzmann factor $f_{\mathrm{B}}(\mathbf{x}) = \exp[-U(\mathbf{x})/(k_{\mathrm{B}}T)]$ over the trap volume, where $U(\mathbf{x})$ is the trap potential at position $\mathbf{x}$. The $U(\mathbf{x})$ is taken to be the sum of two Gaussian beams, one for each cODT beam, and gravity is neglected for simplicity. Each Gaussian beam, with peak depth $U_{i.0}$ and waist $w_{i,0}$, contributes a potential of the form

$$U_i(\mathbf{x}) = U_{i,0}\left(\frac{w_{i,0}}{w_i(z')}\right)^2 \exp\left(\frac{-2(r')^2}{w_i(z')^2}\right), \tag{B1}$$

where $w_i(z') = w_{i,0}\sqrt{1+(z'/z_{\mathrm{R}})^2}$ is the spatially varying beam width and $z_{\mathrm{R}} = \pi w_{i,0}^2/\lambda$ is the Rayleigh range. The primed coordinates z' and r' are taken to be along and perpendicular to the beam's propagation direction, respectively. The value of n_{cp} can be calculated as $N f_{\mathrm{B}}(\mathbf{x}_0)/Z$, where $\mathbf{x}_0$ is the position of the bottom of the trap. Finally $\mathrm{PSD_c}$ is evaluated from its definition in terms of n_{cp} and λ_{dB}. Notably, for much of the sequence the atomic cloud extends out of the cODT region and into the wings of the horizontal ODT, in which case the trap potential seen by the cloud is not harmonic. Thus the well-known result $\mathrm{PSD_c} = N(\hbar\bar{\omega})^3/(k_{\mathrm{B}}T)^3$ for a harmonic trap with geometric mean trap frequency $\bar{\omega}$ cannot be used for most of the sequence.

Calculation of the mean collision rate ν_{c} requires averaging the collision rate $n_{\mathrm{c}}\sigma v_{\mathrm{rms}}$ over the cloud, where σ is the atomic collision cross section and v_{rms} is the root-mean-square relative velocity of atoms in the cloud. The value of n_{c} varies over the trap and obeys $n_{\mathrm{c}}(x) = N f_{\mathrm{B}}(x)/Z$, again neglecting Bosonic statistics. From equipartition for a three-dimensional gas, $(1/2)\mu v_{\mathrm{rms}}^2 = (3/2)k_{\mathrm{B}}T$, where $\mu = m/2$ is the reduced mass for two atoms. Thus the value of v_{rms} is given by $\sqrt{6k_{\mathrm{B}}T/m}$. The local collision rate is averaged by integrating $n_{\mathrm{c}}\sigma v_{\mathrm{rms}}$ over the cloud, weighted by the one-atom number density n_{c}/N, yielding

$$\nu_{\mathrm{c}} = N\sigma\sqrt{\frac{6k_{\mathrm{B}}T}{m}}\int\left(\frac{f_{\mathrm{B}}(x)}{Z}\right)^2 dV. \tag{B2}$$

The above calculations assume that the cloud is in thermal equilibrium, which is often a good approximation. However, after about 440 ms of the final optimized 575-ms sequence, the power in the vertical trapping beam P_z is rapidly increased, as can be seen in Fig. 2(a). This change is likely nonadiabatic for atoms in the wings of the horizontal ODT and the cloud may no longer be in thermal equilibrium. This is likely why the calculated $\mathrm{PSD_c}$ appears to increase beyond ∼1 before the appearance of a BEC. Notably this nonadiabatic portion of the sequence occurs only after $\mathrm{PSD_c}$ has reached 0.4, and thus it does not affect the cooling efficiency estimate of $\gamma \approx 16$ for the cooling up to $\mathrm{PSD_c} = 0.1$.

The peak trap depth $U_{i,0}$ for each beam was determined from the beam waist $w_{i,0}$ and the radial trap frequency $\omega_{i,\mathrm{r}}$ measured for each beam. The beam waists, defined as the radius at which the intensity falls to $1/e^2$ of its peak value, were measured by profiling the trap beams on a separate test setup which focused the light outside of the vacuum chamber. The trap frequencies were directly measured by carefully perturbing the position of a cloud in the cODT and observing its oscillations. Before perturbing the cloud, it was first cooled sufficiently to make it well confined to the central region of the cODT so that the potential was approximately harmonic. The peak trap depth for each beam could then be calculated as $U_{i,0} = m\omega_{i,\mathrm{r}}^2 w_{i,0}^2/4$. This expression can be derived by equating the spring constant for the trap in the radial direction $k = d^2U_i(\mathbf{x})/(dr')^2|_{\mathbf{x}=\mathbf{x}_0}$ to its value for a harmonic oscillator $k = m\omega_{i,\mathrm{r}}^2$.

APPENDIX C: COST SCALING

The peak optical depth (OD) of a pure BEC after sufficient time-of-flight expansion scales as $\mathrm{OD} \propto N_{\mathrm{BEC}}/A$, where A is the area of the cloud in the image. The area scales in proportion to $\bar{v}^2$, where $\bar{v}$ is the expansion velocity, which is related to the BEC chemical potential via $(1/2)m\bar{v}^2 = (2/7)\mu$ in a harmonic trap [36]. Thus, $A \propto \mu$. Furthermore, the chemical potential for a harmonically trapped BEC scales as $\mu \propto N_{\mathrm{BEC}}^{2/5}$ [36], so $A \propto N_{\mathrm{BEC}}^{2/5}$ and $\mathrm{OD} \propto N_{\mathrm{BEC}}^{3/5}$. The expression $\mathrm{OD}^3N^{\alpha-9/5}$ then scales as $(N_{\mathrm{BEC}})^{\alpha}$. Notably this scaling also applies to a harmonically trapped BEC when imaged *in situ*. There, the BEC radius R scales as $R \propto N_{\mathrm{BEC}}^{1/5}$ [36]. In that case, $A \propto R^2 \propto N_{\mathrm{BEC}}^{2/5}$ as before. The same arguments then apply again, indicating that $\mathrm{OD}^3N^{\alpha-9/5}$ scales as $(N_{\mathrm{BEC}})^{\alpha}$ for a harmonically trapped BEC *in situ* just as it does for a BEC after a long time-of-flight expansion.

The scaling of $\mathrm{OD}^3N^{\alpha-9/5}$ for a purely thermal cloud is also of note. For a harmonically trapped thermal cloud, the RMS size in a given direction for any time of flight is proportional to $T^{1/2}$, so $A \propto T$. Thus $\mathrm{OD} \propto N/T$ and $\mathrm{OD}^3N^{\alpha-9/5}$ scales in proportion to $N^{\alpha+(6/5)}/T^3$. Clouds with smaller temperatures are favored by the cost function, and clouds with larger atom numbers are favored as long as $\alpha > -6/5$. For the case $\alpha = -1/5$, the value of $\mathrm{OD}^3N^{\alpha-9/5}$ scales in proportion to N/T^3, which is proportional to $\mathrm{PSD_c}$. That choice of α was often used when optimizing individual stages before reaching the threshold to BEC. However, note that this choice of α leads to the scaling $\mathrm{OD}^3N^{\alpha-9/5} \propto N_{\mathrm{BEC}}^{-1/5}$ for a pure BEC and is thus not a good choice when the cloud reaches condensation.

APPENDIX D: RAMAN COOLING LASER

Standard Doppler cooling requires a laser with a linewidth narrow compared to the optical transition linewidth in order to achieve optimal temperatures. This places stringent technical requirements for Doppler cooling on narrow optical transitions. By contrast, Raman cooling can achieve similar velocity resolution and associated temperatures with a comparatively broad laser. In this work, the light for the Raman coupling

and optical pumping beams, which drive the up-leg and down-leg of the Raman transition, respectively, was derived from the same laser. This ensures that any laser frequency noise is common mode between the two legs of the Raman transition and makes it possible to resolve Doppler shifts much smaller than the laser linewidth. A Distributed Bragg Reflector (DBR) laser diode (Photodigm PH795DBR180TS) without an external cavity was sufficient to generate the Raman cooling light. The forgiving laser linewidth requirements further simplify implementation of our BEC production approach compared to schemes which require Doppler cooling on narrow optical transitions. Thus our approach may be useful even for species which include narrow optical transitions.

[1] L. R. Picard, M. J. Mark, F. Ferlaino, and R. van Bijnen, Deep learning-assisted classification of site-resolved quantum gas microscope images, Meas. Sci. Technol. **31**, 025201 (2020).

[2] Z.-H. Ding, J.-M. Cui, Y.-F. Huang, C.-F. Li, T. Tu, and G.-C. Guo, Fast High-Fidelity Readout of a Single Trapped-Ion Qubit via Machine-Learning Methods, Phys. Rev. Appl. **12**, 014038 (2019).

[3] A. Seif, K. A. Landsman, N. M. Linke, C. Figgatt, C. Monroe, and M. Hafezi, Machine learning assisted readout of trapped-ion qubits, J. Phys. B: At., Mol. Opt. Phys. **51**, 174006 (2018).

[4] G. Ness, A. Vainbaum, C. Shkedrov, Y. Florshaim, and Y. Sagi, Single-Exposure Absorption Imaging of Ultracold Atoms Using Deep Learning, Phys. Rev. Appl. **14**, 014011 (2020).

[5] S. Guo, A. R. Fritsch, C. Greenberg, I. Spielman, and J. P. Zwolak, Machine-learning enhanced dark soliton detection in Bose–Einstein condensates, Machine Learn.: Science Technol. **2**, 035020 (2021).

[6] G. Carleo and M. Troyer, Solving the quantum many-body problem with artificial neural networks, Science **355**, 602 (2017).

[7] H. Saito, Solving the Bose–Hubbard model with machine learning, J. Phys. Soc. Jpn. **86**, 093001 (2017).

[8] L. Wang, Discovering phase transitions with unsupervised learning, Phys. Rev. B **94**, 195105 (2016).

[9] J. Carrasquilla and R. G. Melko, Machine learning phases of matter, Nat. Phys. **13**, 431 (2017).

[10] G. Torlai, B. Timar, E. P. L. van Nieuwenburg, H. Levine, A. Omran, A. Keesling, H. Bernien, M. Greiner, V. Vuletić, M. D. Lukin, R. G. Melko, and M. Endres, Integrating Neural Networks with a Quantum Simulator for State Reconstruction, Phys. Rev. Lett. **123**, 230504 (2019).

[11] P. B. Wigley, P. J. Everitt, A. van den Hengel, J. W. Bastian, M. A. Sooriyabandara, G. D. McDonald, K. S. Hardman, C. D. Quinlivan, P. Manju, C. C. Kuhn *et al.*, Fast machine-learning online optimization of ultra-cold-atom experiments, Sci. Rep. **6**, 25890 (2016).

[12] A. D. Tranter, H. J. Slatyer, M. R. Hush, A. C. Leung, J. L. Everett, K. V. Paul, P. Vernaz-Gris, P. K. Lam, B. C. Buchler, and G. T. Campbell, Multiparameter optimisation of a magneto-optical trap using deep learning, Nat. Commun. **9**, 4360 (2018).

[13] I. Nakamura, A. Kanemura, T. Nakaso, R. Yamamoto, and T. Fukuhara, Non-standard trajectories found by machine learning for evaporative cooling of ^{87}Rb atoms, Opt. Express **27**, 20435 (2019).

[14] A. J. Barker, H. Style, K. Luksch, S. Sunami, D. Garrick, F. Hill, C. J. Foot, and E. Bentine, Applying machine learning optimization methods to the production of a quantum gas, Machine Learn.: Sci. Technol. **1**, 015007 (2020).

[15] E. T. Davletov, V. V. Tsyganok, V. A. Khlebnikov, D. A. Pershin, D. V. Shaykin, and A. V. Akimov, Machine learning for achieving Bose-Einstein condensation of thulium atoms, Phys. Rev. A **102**, 011302(R) (2020).

[16] Y. Wu, Z. Meng, K. Wen, C. Mi, J. Zhang, and H. Zhai, Active learning approach to optimization of experimental control, Chin. Phys. Lett. **37**, 103201 (2020).

[17] C.-C. Chen, R. González Escudero, J. Minář, B. Pasquiou, S. Bennetts, and F. Schreck, Continuous Bose–Einstein condensation, Nature (London) **606**, 683 (2022).

[18] S. Stellmer, R. Grimm, and F. Schreck, Production of quantum-degenerate strontium gases, Phys. Rev. A **87**, 013611 (2013).

[19] J. Rudolph, W. Herr, C. Grzeschik, T. Sternke, A. Grote, M. Popp, D. Becker, H. Müntinga, H. Ahlers, A. Peters *et al.*, A high-flux BEC source for mobile atom interferometers, New J. Phys. **17**, 065001 (2015).

[20] C.-L. Hung, X. Zhang, N. Gemelke, and C. Chin, Accelerating evaporative cooling of atoms into Bose-Einstein condensation in optical traps, Phys. Rev. A **78**, 011604(R) (2008).

[21] R. Roy, A. Green, R. Bowler, and S. Gupta, Rapid cooling to quantum degeneracy in dynamically shaped atom traps, Phys. Rev. A **93**, 043403 (2016).

[22] G. A. Phelps, A. Hébert, A. Krahn, S. Dickerson, F. Öztürk, S. Ebadi, L. Su, and M. Greiner, Sub-second production of a quantum degenerate gas, arXiv:2007.10807

[23] B. Shahriari, K. Swersky, Z. Wang, R. P. Adams, and N. De Freitas, Taking the human out of the loop: A review of Bayesian optimization, Proc. IEEE **104**, 148 (2015).

[24] N. R. Hutzler, L. R. Liu, Y. Yu, and K.-K. Ni, Eliminating light shifts for single atom trapping, New J. Phys. **19**, 023007 (2017).

[25] A. Urvoy, Z. Vendeiro, J. Ramette, A. Adiyatullin, and V. Vuletić, Direct Laser Cooling to Bose-Einstein Condensation in a Dipole Trap, Phys. Rev. Lett. **122**, 203202 (2019).

[26] M. Kasevich and S. Chu, Laser Cooling Below a Photon Recoil with Three-Level Atoms, Phys. Rev. Lett. **69**, 1741 (1992).

[27] J. Snoek, O. Rippel, K. Swersky, R. Kiros, N. Satish, N. Sundaram, M. Patwary, M. Prabhat, and R. Adams, Scalable Bayesian optimization using deep neural networks, in *Proceedings of the 32nd International Conference on Machine Learning* (PMLR, Lille, France, 2015), JMLR: W&CP, Vol. 37, pp. 2171–2180.

[28] P. T. Starkey, C. J. Billington, S. P. Johnstone, M. Jasperse, K. Helmerson, L. D. Turner, and R. P. Anderson, A scripted control system for autonomous hardware-timed experiments, Rev. Sci. Instrum. **84**, 085111 (2013).

[29] M. Weidemüller, T. Esslinger, M. A. Ol'shanii, A. Hemmerich, and T. W. Hänsch, A novel scheme for efficient cooling below the photon recoil limit, Europhys. Lett. **27**, 109 (1994).

[30] D. Boiron, C. Triché, D. R. Meacher, P. Verkerk, and G. Grynberg, Three-dimensional cooling of cesium atoms in four-beam gray optical molasses, Phys. Rev. A **52**, R3425(R) (1995).

[31] P. D. Lett, R. N. Watts, C. I. Westbrook, W. D. Phillips, P. L. Gould, and H. J. Metcalf, Observation of Atoms Laser Cooled below the Doppler Limit, Phys. Rev. Lett. **61**, 169 (1988).

[32] J. Dalibard and C. Cohen-Tannoudji, Laser cooling below the Doppler limit by polarization gradients: simple theoretical models, J. Opt. Soc. Am. B **6**, 2023 (1989).

[33] V. Vuletić, A. J. Kerman, C. Chin, and S. Chu, Observation of Low-Field Feshbach Resonances in Collisions of Cesium Atoms, Phys. Rev. Lett. **82**, 1406 (1999).

[34] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevResearch.4.043216 for a movie of the atomic cloud during the cooling sequence.

[35] K. Burnett, P. S. Julienne, and K.-A. Suominen, Laser-Driven Collisions between Atoms in a Bose-Einstein Condensed Gas, Phys. Rev. Lett. **77**, 1416 (1996).

[36] F. Dalfovo, S. Giorgini, L. P. Pitaevskii, and S. Stringari, Theory of Bose-Einstein condensation in trapped gases, Rev. Mod. Phys. **71**, 463 (1999).