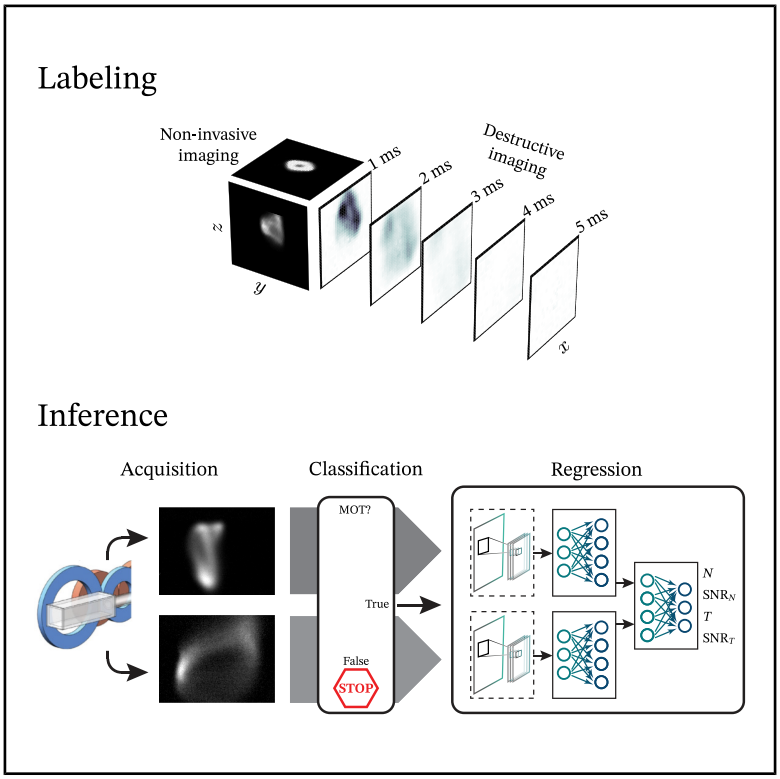
Newton

Article

Nondestructive characterization of laser-cooled atoms using machine learning

Graphical abstract



Authors

Guilherme de Sousa, Michael Doris, Dario D'Amato, Brady Egleston, Justyna P. Zwolak, Ian B. Spielman

Correspondence

jpzwolak@nist.gov (J.P.Z.), ian.spielman@nist.gov (I.B.S.)

In brief

Fluorescence images of atoms in a magneto-optical trap reveal the atoms' spatial configuration but not how cold they are. de Sousa et al. train machine learning models on paired fluorescence and time-of-flight data to show that these images contain hidden information about both atom number and temperature, pointing toward fast, nondestructive diagnostics and real-time feedback for cold-atom experiments.

Highlights

- Machine learning models infer atom number and temperature from MOT fluorescence
- Fluorescence images contain hidden information about atom temperature
- Brightness alone misses most temperature information
- Proposed method supports fast, nondestructive diagnostics for cold atoms

de Sousa et al., 2026, Newton 2, 100518
September 8, 2026 Published by Elsevier Inc.
https://doi.org/10.1016/j.newton.2026.100518

CellPress

Please cite this article in press as: de Sousa et al., Nondestructive characterization of laser-cooled atoms using machine learning, Newton (2026), https://doi.org/10.1016/j.newton.2026.100518

Newton

CellPress
OPEN ACCESS

Article

Nondestructive characterization of laser-cooled atoms using machine learning

Guilherme de Sousa,[1,2,3,7] Michael Doris,[2,4,7] Dario D'Amato,[2,4,7] Brady Egleston,[2,4] Justyna P. Zwolak,[2,5,6,8,*] and Ian B. Spielman[2,4,8,9,*]

[1]Department of Physics, University of Maryland, College Park, MD, USA
[2]Quantum Measurement Division, National Institute of Standards and Technology, Gaithersburg, MD, USA
[3]Instituto de Fisica de Sao Carlos, Universidade de Sao Paulo, Sao Carlos, SP, Brazil
[4]Joint Quantum Institute, National Institute of Standards and Technology and University of Maryland, Gaithersburg, MD, USA
[5]Technology Test and Evaluation Division, National Institute of Standards and Technology, Gaithersburg, MD, USA
[6]Joint Center for Quantum Information and Computer Science, University of Maryland, College Park, MD, USA
[7]These authors contributed equally
[8]Senior author
[9]Lead contact
*Correspondence: jpzwolak@nist.gov (J.P.Z.), ian.spielman@nist.gov (I.B.S.)
https://doi.org/10.1016/j.newton.2026.100518

ACCESSIBLE OVERVIEW The magneto-optical trap (MOT) is a foundational laser-cooling and trapping tool used in modern atomic physics, with applications including atomic clocks and sensors as well as quantum simulators and computers. Before realizing these applications, researchers must have reliable laser cooling with stable number and temperature. The usual way to measure those quantities is to release the atoms, let them fly apart for a time, and image them; it works but is destructive because once measured, the sample is destroyed. Here, we show that machine learning (ML) can recover the same information from a much gentler signal: the fluorescence that atoms naturally emit while they are confined in an MOT. To the eye, these images reveal the cloud's size and shape, but they do not obviously encode internal properties such as temperature. We built a labeled dataset by pairing fluorescence images of potassium-39 atom clouds with conventional destructive measurements, then trained ML models to infer atom number and temperature from the fluorescence images alone. Simple models that use only total brightness provide limited information, especially about temperature. By contrast, neural network models that can use spatial patterns across the images perform substantially better, showing that the fluorescence carries hidden information about the trapped atoms. In practice, this approach provides fast, nondestructive diagnostics for cold-atom experiments and supports real-time feedback in systems where repeatedly destroying the sample can be costly.

SUMMARY

Laser cooling and trapping techniques are foundational to modern atomic physics, with applications including atomic clocks and sensors as well as quantum simulators and computers. These applications require stable, well-calibrated laser cooling and accurate determination of atom number and temperature. We develop machine learning techniques for estimating physical properties of laser-cooled potassium-39 atoms in a magneto-optical trap using only the scattered light—i.e., fluorescence—that is intrinsic to the cooling process. *In situ* snapshot images of fluorescing atomic ensembles directly reveal the spatial structure of these millimeter-scale objects but contain no obvious information regarding internal properties such as the temperature. We first assembled and labeled a balanced dataset sampling 8×10^3 different experimental parameters that includes examples with large and dense atomic ensembles, a complete absence of atoms, and everything in between. We describe a range of models trained to predict atom number and temperature solely from fluorescence images. These run the gamut from a poorly performing linear regression model based only on integrated fluorescence to deep neural networks that give number and temperature with fractional uncertainties of 0.1 and 0.2, respectively. These results show that fluorescence images can provide fast, nondestructive diagnostics for cold-atom experiments and may enable real-time experimental feedback.

Please cite this article in press as: de Sousa et al., Nondestructive characterization of laser-cooled atoms using machine learning, Newton (2026), https://doi.org/10.1016/j.newton.2026.100518

INTRODUCTION

The past decade has witnessed a rapid adoption of machine learning (ML) techniques in the applied and fundamental physical sciences.[1] These approaches have been used for everything from stabilizing nuclear fusion reactors[2] and designing and controlling quantum devices[3] to imaging black hole event horizons,[4] discovering new materials,[5] and searching for physics beyond the standard model of particle physics.[6] A key use case is the identification of "hidden" information; i.e., information that is not easily accessible using the available measurement techniques. For example, in many-body quantum systems, topological order is hidden because its signatures are highly non-local[7]; nevertheless, ML tools have demonstrated the ability to identify these phases with both simulated[8] and experimental data.[9] Our focus is analogous: estimating internal properties of laser-cooled atoms from purely non-destructive fluorescence images, where only scattered light intrinsic to the cooling process is observed. While such fluorescence images directly reveal the millimeter-scale spatial distribution of the atomic cloud, they provide no obvious clues about internal properties, such as temperature, which are traditionally measured using time-of-flight (TOF) techniques. We demonstrate that deep learning can extract both the straightforward and the hidden characteristics of these ensembles with high accuracy from real experimental data.

Laser cooling is a foundational technique[10] underpinning virtually all atom-based quantum technologies, including quantum sensors, simulators, and computers. To harness quantum effects with neutral atoms, these technologies require large collections of ultracold atoms that share the same quantum state of motion and the same internal state (i.e., atomic level). For some applications, laser cooling alone is sufficient, while in others, it is followed by additional stages of cooling and state purification. The magneto-optical trap (MOT) is widely used to capture, trap, and cool neutral atom clouds[11] ranging in size from tens of billions of atoms down to the single-atom level. For example, MOTs serve as precursors to today's leading optical lattice clocks,[12] optical tweezer arrays,[13] quantum degenerate gases,[14] and much more.

Each of these applications, as well as the MOT itself, has been enhanced by ML. For example, ML has been used to automate the operation of optical atomic clocks in real-world applications.[15] Furthermore, the control parameters used when loading an MOT,[16] creating optical tweezer arrays,[17] and producing quantum gases have all been optimized using ML techniques.[18,19] Oftentimes, these ML-based optimizers discover unexpected or counterintuitive parameter sequences. Together, these applications demonstrate the breadth of ML's applicability to atom-based quantum science and technology. Here, we focus on efficiently extracting information from noninvasive images of laser cooled atoms in an MOT.

An MOT operates using an interplay of optical and magnetic forces, and atoms in an MOT constantly scatter light from illuminating lasers in all directions. Imaging this unavoidable fluorescence, therefore, affords an often-used mechanism for noninvasive monitoring of the trapped cloud. Although the exact scattering rate of each atom depends on detailed experimental parameters, it is intuitive that the overall amount of scattered light generally increases with atom number; indeed, at low enough density (such that light scattered by one atom is unlikely to be reabsorbed by another), the overall fluorescence is simply proportional to the atom number.[20]

Except for very simple atoms (internal angular momentum $J = 0$ to $J = 1$) and very small (negligible rescattering) clouds, neither the total fluorescence nor the cloud's size and shape provide any obvious indication of its temperature, thereby hiding this parameter from standard image-based analyses. Both number and temperature can be readily measured using invasive techniques such as TOF imaging, in which the atoms are released from the MOT and allowed to ballistically expand for a set time. In this way, the spatial distribution after TOF is correlated with the initial velocity distribution, from which the temperature can be estimated.[21] Performing TOF imaging immediately after acquiring fluorescence images thus allows us to generate datasets of fluorescence images, labeled by the atom number N and temperature T. While the process of compiling the dataset is destructive, the application of trained models on subsequent fluorescence images is not.

In more detail, each element of our dataset (a single "shot" of the experiment) includes two fluorescence images captured along two orthogonal axes together with a reference TOF image. These individual shots are then collected into "sets," each consisting of $M = 5$ shots that differ only in the TOF time, thereby improving the accuracy with which we determine atom number and temperature. The final dataset consists of $\approx 39 \times 10^3$ individual shots, whose $\approx 8 \times 10^3$ distinct parameters span a nominally balanced portion of parameter space, with contributions ranging from large atomic clouds down to a complete absence of atoms and everything in-between. We then train a range of regression models on these datasets to extract N and T, as well as corresponding quality metrics, directly from fluorescence images alone. In operation, new data employing the final trained models follow the overall workflow illustrated in Figure 1.

We explored a total of five regression models ranging in sophistication from a trivial model that returned a constant output (serving as our benchmark), to a convolutional neural network (CNN); these models' performance increases in line with their sophistication. The final CNN model predicts N with a typical uncertainty of $\pm 4 \times 10^6$ of 2×10^8 atoms and T with a typical fractional uncertainty of ± 0.2 (see supplemental information for all results). In the context of quantum degenerate atoms, Griffiths et al.[22] numerically demonstrate that ML models can, in principle, infer temperature from the spatial distribution of atoms in a Bose-Einstein condensate.

RESULTS

Experiment

The experiment uses laser cooling techniques to create and capture an atomic cloud of ^{39}K atoms using an MOT.[11,23,24] Our MOT, configured in the standard geometry shown in Figure 2A, relies on radiation pressure from three pairs of counterpropagating laser beams that each include contributions from both "cooling" and "repump" lasers nearly resonant with the D2 line, along with a quadrupole magnetic field. While in this work we study a range of parameters, the cooling laser would generally be

Please cite this article in press as: de Sousa et al., Nondestructive characterization of laser-cooled atoms using machine learning, Newton (2026), https://doi.org/10.1016/j.newton.2026.100518

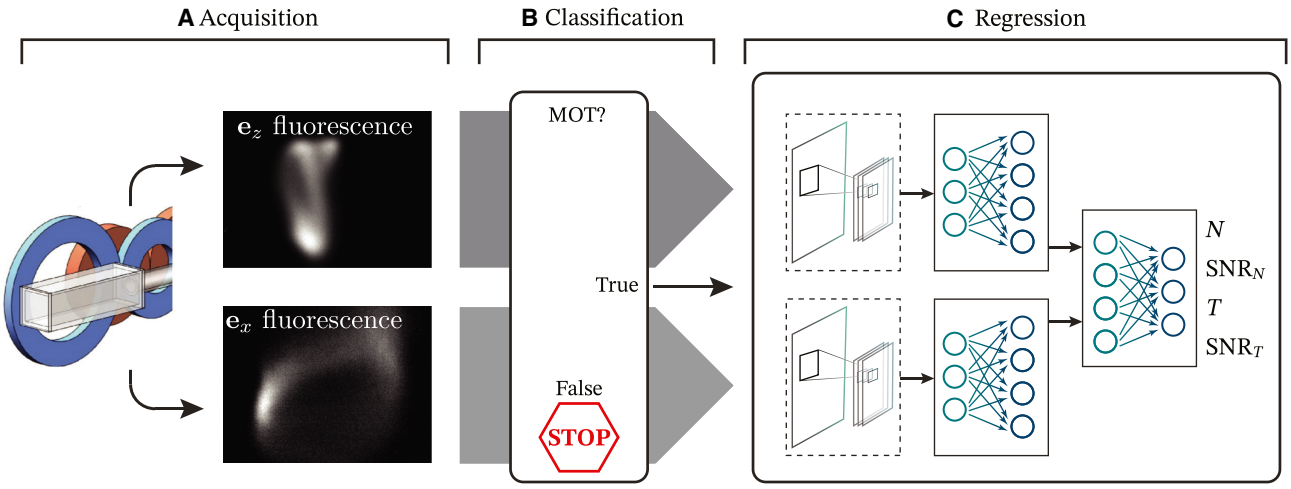


Figure 1. Workflow of classification and regression system in operation

(A) Acquisition. ^{39}K atoms are laser cooled and then fluorescence imaged along $\mathbf{e}_x$ and $\mathbf{e}_z$ as described under Experiment and Sequence. Images for optimal experimental parameters are shown.

(B) Classification. After acquisition, data are classified as having trapped atoms present or not.

The `MOT` label is correspondingly assigned `True` or `False`.

(C) Regression. The data are then passed into a regression model that first potentially processes the images independently, fuses the data, and predicts atom number N and temperature T along with corresponding signal-to-noise ratios SNR$_N$ and SNR$_T$. The ML Toolbox section discusses data pre-processing, classification, and regression.

red-detuned from the $F = 2 \rightarrow F' = 3$ transition, and the repump laser would be tuned near the $F = 1 \rightarrow F' = 2$ transition (Figure 2B).

In this configuration, the lasers' radiation pressure Doppler cools our ^{39}K atoms to a typical temperature of ≈2 mK, and the inhomogeneous detuning from the quadrupole magnetic field adds confinement. The use of cooling and repumping lasers assures that atoms do not accumulate in an unaddressed (dark) ground state, thereby leaving the cooling process.

Hardware

Our experimental apparatus, shown schematically in Figure 2A, is a standard vapor-fed MOT. This first captures atoms from the low-velocity tail (below ≈30 m/s) of the dilute room temperature ^{39}K vapor in our vacuum system. These atoms are then cooled and collected, yielding trapped atoms with velocities around 1 m/s. All of the control parameters that are varied in this study are detailed in Table 1.

The apparatus makes use of two laser systems; one generates the cooling and imaging laser beams, and the other generates the repump beams. The repump laser system is locked to a potassium reference cell using saturated absorption spectroscopy. The cooling laser system is then offset locked to the repump using a phase-locked loop circuit giving a tunable frequency offset f_{lock} between these laser systems. Each final laser beam relies on an acusto-optical modulator (AOM) to provide high-bandwidth control of the power (controlled by an external voltage) and to introduce a tunable frequency shift. In our dataset, the power of both the cooling and repump beams are tuned with control voltages V_{cool} and V_{rep}. The atomic levels relevant to Doppler laser cooling of ^{39}K are shown in Figure 2B. The vertical black line shows the nominal scale of the D2 transition; the red line shows the cooling laser, red detuned from the $F = 2 \rightarrow F' = 3$ transition; the blue line shows the repump, red detuned from $F = 1 \rightarrow F' = 2$ transition; and last, the orange line shows the absorption imaging probe frequency resonant with the $F = 2 \rightarrow F' = 3$ transition. (In the first milliseconds of TOF, the magnetic field has not reached zero, so we optimized the frequency offset at each t_{TOF}. Unlike atoms such as ^{87}Rb or ^{23}Na, in ^{39}K, the excited-state hyperfine splitting is poorly resolved, being close to the ≈6 MHz transition line width. In this configuration, the repump laser contributes significantly to the cooling and trapping forces.)

When assembling our dataset, the repump AOM provides a variable frequency f_{rep}, but the cooling AOM frequency is fixed. Instead, the offset lock provides frequency tuning to the cooling laser system and does so over a wider frequency range than is possible with an AOM.

To increase the geometric stability of our experiment, all of the laser light is injected into optical fibers after being conditioned by the AOMs. The probe laser travels in a conventional single-mode fiber, while the cooling and repump light is combined by an evanescent wave fiber splitter-combiner that distributes the optical power equally into the requisite six beams.

In addition to these optical fields, MOT operation requires a quadrupole magnetic field. We generate this field with a pair of copper coils arranged in an anti-Helmholtz configuration, each carrying the same current I_{quad}. Additional coils (data not shown) compensate for the ambient background magnetic field.

The atomic ensemble is imaged along the three Cartesian axes ($\mathbf{e}_x$, $\mathbf{e}_y$, and $\mathbf{e}_z$) using independent two-lens Keplerian microscopes. The images are captured on complementary metal oxide semiconductor (CMOS) cameras, each labeled by its imaging axes; for example, the "$\mathbf{e}_x$ camera." The important properties of these imaging systems are detailed in Table 2. Because we are

Please cite this article in press as: de Sousa et al., Nondestructive characterization of laser-cooled atoms using machine learning, Newton (2026), https://doi.org/10.1016/j.newton.2026.100518

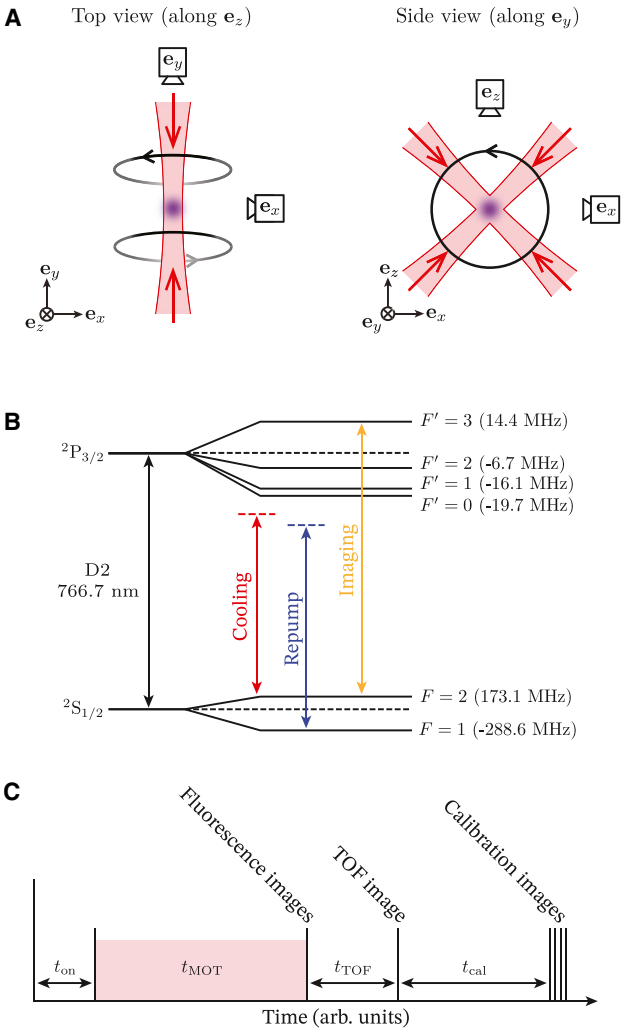


Figure 2. Experimental geometry, relevant energy levels, and MOT operation sequence

(A) Schematic of the experimental geometry including top (left) and side (right) views, with the $\mathbf{e}_z$ and $\mathbf{e}_x$ cameras marked by small boxes.

The red-shaded regions indicate the MOT laser beams with propagation direction indicated by the red arrows; the black loops denote current carrying wires, with arrows showing the direction of current flow. Last, the purple spot illustrates the position of the trapped atoms. As shown, the MOT beams render $\mathbf{e}_y$ unavailable for fluorescence imaging.

(B) Relevant energy levels for laser cooling and trapping ^{39}K using only the D2 line. In standard operation, the cooling light is red detuned by 40 MHz from the $|F = 2\rangle \rightarrow |F' = 3\rangle$ transition, and the repump light is red detuned by 30.5 MHz from the $|F = 1\rangle \rightarrow |F' = 2\rangle$ transition.

(C) Experimental sequence for MOT operation.

imaging large millimeter-scale objects, our images are demagnified with effective pixel sizes that are larger than the ≈6 μm diffraction limit of these imaging systems.

Sequence

This section outlines the time-sequence of a single experimental shot yielding an elementary unit of a dataset. This sequence (potentially) generates a cloud of laser-cooled atoms from a specific set of experimental parameters. Each such shot follows a predefined sequence of events organized into stages of initialization, MOT loading, fluorescence imaging, TOF imaging, and calibration (Figure 2C). Figure 3 shows fluorescence and TOF data for dense and compact (top) as well as more representative clouds (bottom).

Initialization. Prior to MOT loading, we allow for a period of hardware equilibration of duration t_{on} = 5 ms. During this time, the parameters `Cooling_Lock_Offset`, `Repump_AOM_Freq`, `MOT_Quad_Amps`, `Cooling_AOM_Volts`, and `Repump_AOM_Volts`) are set, with the lasers mechanically blocked by shutters just prior to the entering optical fibers.

MOT loading. Cooling and trapping is then initialized by abruptly opening the shutters. This stage has a duration t_{MOT}, equal to the `MOT_Loading_Time` parameter in Table 1. The maximum loading time of 1.8 s was selected to be double the ≈900 ms needed for the atom number to become saturated under optimal MOT parameters.

Fluorescence imaging. Immediately following MOT loading, the $\mathbf{e}_x$ and $\mathbf{e}_z$ cameras acquire the respective fluorescence images.

TOF imaging. The magnetic and optical fields are then removed, thereby freeing the atoms from the trap, after which they undergo TOF evolution for a duration t_{TOF}. The resulting 2D column density ρ_{ij} in each pixel (labeled by i) is measured via absorption imaging, a process that, in essence, detects the shadow cast by the atom cloud in a probe laser. This image is acquired by pulsing on the probe laser (traveling along $\mathbf{e}_y$) for 10 μs, and an auxiliary repump beam (traveling along $\mathbf{e}_z$) starting 20 μs prior to the probe pulse. The $\mathbf{e}_y$ camera then measures the shadowed probe.

Calibration. The raw fluorescence and absorption images require additional reference data to mitigate the effect of background light as well as calibrate the unshadowed probe profile (see the supplemental information for a description of these reference frames). After TOF imaging, these additional images are acquired, adding a time t_{cal} = 411 ms + $2t_{TOF}$ to each shot.

Data

We collected data under a wide range of experimental conditions to produce a diverse dataset for model training. This was achieved by varying the six experimental parameters in Table 1. The values of these parameters were sampled so as to generate an approximately balanced dataset.

The dataset consists of 14 batches, each ranging in size from 14 to 1,000 sets of shots. Each set contains M = 5 shots with identical MOT parameters, except that t_{TOF} samples the set {1,2,3,4,5} ms. Each shot yields two fluorescence images taken at the end of MOT loading and one absorption image taken after a subsequent t_{TOF} expansion. In total, the dataset contains 38,915 shots.[25] Each data file contains the experimental parameters used to generate the shot, as well as all image frames described under Hardware and Sequence.

Labeling strategy

Each shot in the dataset is assigned the six labels shown in Table 1; these labels are all assigned at the set level. Here, we

Please cite this article in press as: de Sousa et al., Nondestructive characterization of laser-cooled atoms using machine learning, Newton (2026), https://doi.org/10.1016/j.newton.2026.100518

Table 1. Dataset labels

Symbol	Label	Approximate range	Description
V_{cool}	Cooling_AOM_Volts	0.1–1.5 V	parameter: voltage of the cooling laser beam's AOM, controls the intensity of the cooling laser beam
V_{rep}	Repump_AOM_Volts	0.4–1.5 V	parameter: voltage of the repump laser beam's AOM, controls the intensity of the repump laser beam
f_{lock}	Cooling_Lock_Offset	85–95 MHz	parameter: controls the frequency offset of the cooling laser with respect to the repump laser
f_{rep}	Repump_AOM_Freq	74–94 MHz	parameter: controls the frequency offset of the repump laser beam compared to the repump laser source
I_{quad}	MOT_Quad_Amps	2–40 A	parameter: current of the MOT quadrupole coils, affects the strength of the magnetic field
t_{MOT}	MOT_Loading_Time	100 to 1,800 ms	parameter: duration of MOT loading time
t_{TOF}	TOF_Time	1–5 ms	set variable: time of flight
–	MOT	[True, False]	label: indicates whether there is any identifiable fluorescence signal
–	VALID_SET	[True, False]	label: indicates whether data could be assigned labels
N	NUM	8×10^6 to 2×10^8	label: number of atoms
SNR_N	NUM_REL	0.1–100	label: number reliability.
T	TEMP	0.5–30 mK	label: temperature
SNR_T	TEMP_REL	0.1–100	label: temperature reliability.

This table contains the subset of our experimental parameters that are varied between different elements in our dataset (top), the settings that change within each set (middle), and the labels derived by our classification and fitting processes (bottom). See supplemental information for a calibrated conversion between relevant control parameters and quantities in units of intensity and magnetic field gradient. The parameters broadly span possible trap conditions, including the optimal MOT settings for the apparatus.

describe our strategy for generating these labels in the order they are assigned in our labeling pipeline.

The `MOT` label identifies data with overall fluorescence in excess of the background noise level and is either `true` or `false`. Because our data are organized into sets sharing the same parameters, the `MOT` label for a specific shot is assigned `true` only if every image in its set has a detectable fluorescence signal.

We separately determine the background noise level for the $\mathbf{e}_x$ and $\mathbf{e}_z$ cameras from fluorescence frames taken with no atoms present. The pixel values in each such mock fluorescence image are summed to obtain an overall signal, shown by the black curves in Figure 4. Gaussian fits yield a root-mean-square (RMS) width of RMS = 1.22×10^5 counts for both cameras. These curves are then plotted along with corresponding histograms of the summed fluorescence signal S_x and S_z from true fluorescence images (red), showing good agreement for the nonphysical negative portion of the fluorescence signal. We estimate the true distribution by subtracting the background distributions (pink curves).

Together, these distributions allow us to compute the F1 metric (see supplemental information for any proposed threshold [inset]. In practice, we select thresholds of 1.9 × RMS and 1.5 × RMS (horizontal gray lines) for the $\mathbf{e}_x$ and $\mathbf{e}_z$ cameras, respectively, where the F1 metrics achieve their maximum values of 0.977 and 0.975.

Table 2. Imaging parameters

Camera name	Model	Resolution	Magnification	Magnified pixel size	Measurement type
$\mathbf{e}_x$ camera	Mako G-030B	644 × 484	0.5	14.8 μm	fluorescence imaging
$\mathbf{e}_y$ camera	Mako G-131B	1280 × 1024	0.333	15.9 μm	absorption imaging
$\mathbf{e}_z$ camera	Mako G-030B	644 × 484	0.375	19.7 μm	fluorescence imaging

This table describes the relevant information for our three imaging systems (any mention of equipment, instruments, software, or materials does not imply recommendation or endorsement by the National Institute of Standards and Technology). The $\mathbf{e}_x$ and $\mathbf{e}_z$ magnifications differ due to geometric constraints of the optical layout and the available lens selection.

Please cite this article in press as: de Sousa et al., Nondestructive characterization of laser-cooled atoms using machine learning, Newton (2026), https://doi.org/10.1016/j.newton.2026.100518

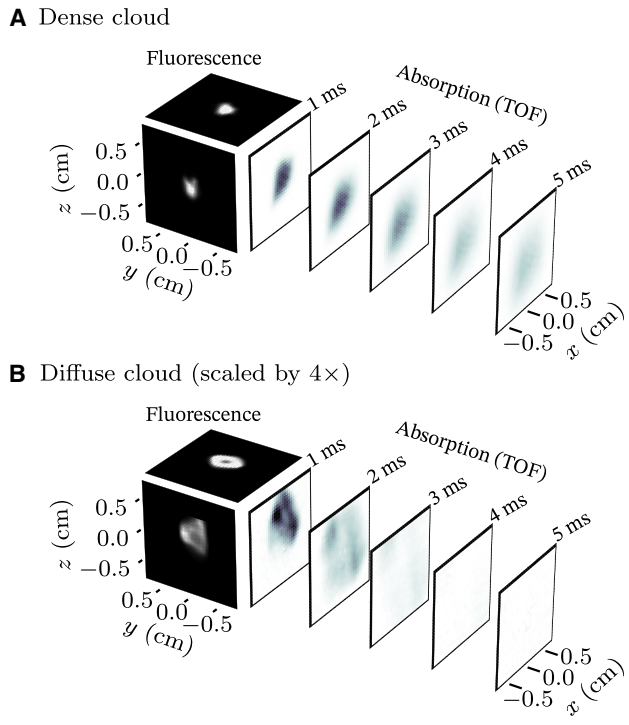


Figure 3. Comparison between compact and diffuse clouds

(A) and (B) correspond to (`BATCH 00, SET 150`) and (`BATCH 03, SET 995`), respectively. The signal in (B) is scaled by 4× to make this lower-quality MOT visible.

The `NUM` and `TEMP` labels, describing the cloud's atom number and temperature, are real valued and by default are assigned `NaN` when *MOT* = *False*. Both quantities can be obtained from the spatial distribution of atoms expanding during TOF. The integrated distribution directly yields atom number, while the evolution of the distribution over time provides access to the velocity distribution and, therefore, an effective mean thermal energy.[21]

In order to make this labeling stage fast and reliable, we model both the velocity distribution and the initial density distribution as Gaussian (the fitting process is further accelerated by downsampling the 2D column density images from 1 024 × 1 280 to 51 × 64 pixels, which, for our millimeter-scaled TOF distributions, does not impact the fit parameters). The resulting model distribution is Gaussian at every t_{TOF}, with density

$$\rho(x,z) = \frac{N}{2\pi w_x w_z} \exp\left[-\frac{1}{2}\left(\frac{x^2}{w_x^2} + \frac{z^2}{w_z^2}\right)\right]; \quad \text{(Equation 1)}$$

because our TOF data are imaged along the $\mathbf{e}_y$ axis, we cannot access the distribution along $\mathbf{e}_y$. This expression is in terms of the TOF-expanded RMS widths

$$w_{x,z}^2(t_{\mathrm{TOF}}) = w_{x,z}^2(0) + \frac{k_B T}{m} t_{\mathrm{TOF}}^2, \quad \text{(Equation 2)}$$

the initial widths $w_{x,z}(0)$, the atom number N, the temperature T, the atomic mass m, and Boltzmann's constant k_B.

For each set, we perform a joint fit to its M shots (each at a different t_{TOF}), yielding a single number and temperature. In practice, we find that the distribution expands at different rates along $\mathbf{e}_x$ and $\mathbf{e}_z$, giving separate values T_x and T_z, which we average to yield the final temperature label. This average quantifies the overall thermal energy (along the observed directions), which is a primary quantity of interest for experiments using MOTs. The supplemental information details the fitting process.

The `NUM_REL` and `TEMP_REL` labels describe the reliability of the number and temperature labels, are real valued, and by default are assigned `NaN` when *MOT* = *False*. As seen in Figure 3, the observed TOF density distributions can be far from Gaussian; therefore, the fit uncertainties, ΔN and ΔT, reported by our Gaussian model cannot be simply interpreted as statistical uncertainties. Instead, they reflect an uncalibrated combination of statistical uncertainties and systematic artifacts that we use to define heuristic reliability indices *NUM_REL* = $|N/\Delta N| \equiv \mathrm{SNR}_N$ and *TEMP_REL* = $|T/\Delta T| \equiv \mathrm{SNR}_T$ that can loosely be interpreted as signal-to-noise (SNR) ratios.

We cross-checked the number obtained from Gaussian fits against that found from directly summing over the 1 ms TOF images (before atoms have left the field of view) and found that their differences are consistent with the fit uncertainties, independent of N and almost independent of T. This indicates a typical number uncertainty of 6×10^6 for the labels.

The `VALID_SET` Boolean label is assigned `True` unless the set is rendered invalid for technical reasons, such as one or more images not being acquired or a labeling fit failing to converge.

Data organization

After labeling, we set aside all of batch 6 (684 sets, about 10% of the complete dataset) as an out-of-distribution test set and a randomly selected 10% of the remaining 13 batches for a traditional in-distribution test set. These test sets were completely excluded from the training process; their contents were never used for training, nor were they used to guide the model development process. All training and validation therefore only used the ≈80% of the data that were not part of either test set; during training, the data are further subdivided for 10-fold cross-validation (discussed under Training). The overall breakdown of the dataset was therefore approximately 72:8:10:10% for training : validation : in − distribution testing : out − of − distribution testing.

ML toolbox

Here, we describe the mechanism by which fluorescence image data from the $\mathbf{e}_x$ and $\mathbf{e}_z$ cameras are processed as they move through our ML analysis pipeline. Prior to any further processing, (1) the `MOT` label is determined as described under Labeling strategy by separately comparing the summed fluorescence counts from the $\mathbf{e}_x$ and $\mathbf{e}_z$ images to the predetermined thresholds; (2) the pixel values are divided by 4 096, normalizing them to the maximum signal of our 12-bit charge-coupled device (CCD) cameras; and (3) the resolution of both fluorescence images are reduced from 644 × 484 to 64 × 48. Processing then terminates for data with *MOT* = *False*.

Regression

This section describes the regression models employed to predict the number, temperature, and associated reliabilities using

Please cite this article in press as: de Sousa et al., Nondestructive characterization of laser-cooled atoms using machine learning, Newton (2026), https://doi.org/10.1016/j.newton.2026.100518

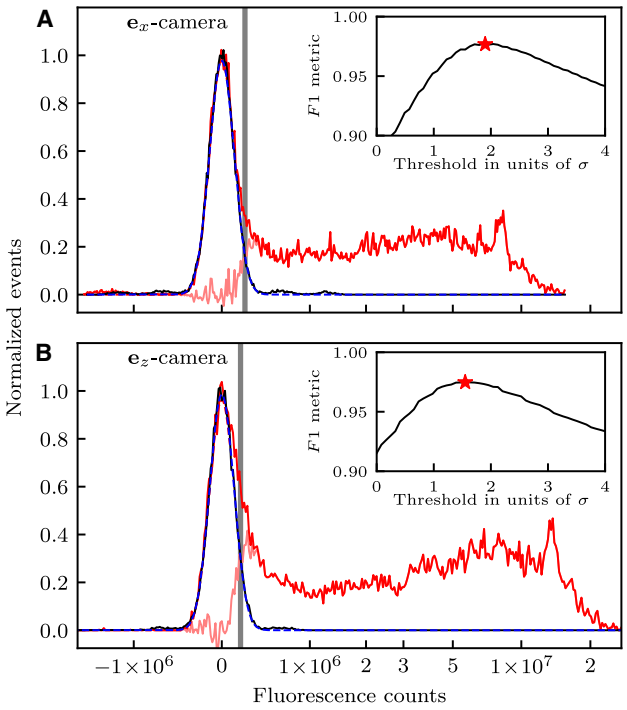


Figure 4. Normalized fluorescence histograms

(A) and (B) derive from the $\mathbf{e}_x$ and $\mathbf{e}_z$ cameras, respectively. Horizontal axes are arcsinh scaled, a symmetric scaling that interpolates between linear for small arguments and logarithmic for large arguments. Solid black curves, obtained with no atoms present, are plotted along with Gaussian fits (dashed blue curves). Red curves describe the complete dataset, while the pink curves are the difference between the complete and the no-atoms histograms. The vertical gray lines represent the threshold below which data are assigned `MOT = False`. Insets: F1 metric as a function of threshold, with the operating point marked in red.

only the fluorescence images. The least sophisticated of these "models" is a constant (CON) function that returns the same output irrespective of its input; this defines the baseline performance level to which all other models are compared. We then progress to a simple linear function (LIN) of the summed fluorescence counts, to a single-layer linear model (i.e., matrix multiplication [MM]), then to a multi-layer perceptron (MLP), and culminate with a CNN. These models approximately recover N, SNR_N, T, and SNR_T, with performance increasing in line with their sophistication (the parameter count of all models is tabulated in the supplemental information).

All of these models are optimized with respect to the overall loss function $L^2 = D^{-1}\sum \ell^2$, averaged over a K-element dataset. Each member of the dataset has an individual loss

$$\ell^2 = \mathrm{SNR}_N^2\left\{\left[1-\frac{N'}{N}\right]^2 + \left[1-\frac{\mathrm{SNR}'_N}{\mathrm{SNR}_N}\right]^2\right\} + \mathrm{SNR}_T^2\left\{\left[1-\frac{T'}{T}\right]^2 + \left[1-\frac{\mathrm{SNR}'_T}{\mathrm{SNR}_T}\right]^2\right\}, \quad \text{(Equation 3)}$$

derived from model predictions of the physical parameters N' and T' and their reliability indices SNR'_N and SNR'_T. This loss function minimizes the fractional uncertainties in each quantity, weighted by the SNR predicted by the fit; i.e., our reliability label.

This is equivalent to the standard weighted least-squares loss function, where the error terms such as $(N - N')^2$ are weighted by ΔN^2. Therefore, ℓ^2 can be interpreted as the L2 norm of the loss vector

$$\ell = \left[\frac{N-N'}{\Delta N}, \mathrm{SNR}_N - \mathrm{SNR}'_N, \frac{T-T'}{\Delta T}, \mathrm{SNR}_T - \mathrm{SNR}'_T\right]. \quad \text{(Equation 4)}$$

Because ℓ consists of ratios of like quantities, every component is of nominally comparable scale; we therefore did not require additional relative weighting hyper-parameters.

The CON model is the absolute minimal case that returns the same values irrespective of its input and therefore has 4 "trainable" parameters, one per regression variable. The simple form of the loss function allows us to express these parameters in closed form. For a dataset with K elements and arbitrary labels $\{A_k\}_{k=1}^K$ and $\{\mathrm{SNR}_{A,k}\}_{k=1}^K$, the corresponding contribution to Equation 3 is minimized by

$$A' = \frac{\sum_{k=1}^K A_k \Delta A_k^{-2}}{\sum_{k=1}^K \Delta A_k^{-2}}, \quad \text{(Equation 5)}$$

where $\Delta A_k^{-2} = \left(\mathrm{SNR}_{A,k}/A_k\right)^2$ serve as statistical weight factors.

We now incrementally increase complexity with the LIN model, a linear function of the summed fluorescence counts S_x and S_z. For example, atom number is predicted by

$$N' = a_x S_x + a_z S_z + b, \quad \text{(Equation 6)}$$

with learnable slopes a_x and a_z and offset b. Thus, the overall model for our 4 regression variables has 12 parameters.

We continue to increase complexity by turning to MM, the most general linear model, an offset matrix product (equivalent to a single fully connected layer with bias and linear activation). To do so, we address the multi-input nature of the dataset with early fusion,[26] in which we flatten both images into 1D vectors and concatenate them. In terms of the resulting data vector d_j, with dimension $D = 2\times(64 \times 48) = 6\,144$, this linear model predicts a 4D vector

$$p'_i = \sum_j A_{ij} d_j + b_i, \quad \text{(Equation 7)}$$

where the linear transform is encoded by the $4 \times D$ matrix A_{ij} and 4D offset vector b_i, yielding a total of $4D + 4$ parameters.

The MLP is a *bona fide* deep learning model; our implementation employs an intermediate fusion approach[26] to reduce the parameter count. As schematically illustrated in Figure 1, the images are individually flattened and propagated through a sequence of independent fully connected layers whose output vectors are then concatenated (fused). The fused vector is then passed through a series of fully connected layers. The MLP is distinguished from MM in that each layer consists of MM followed by a non-linear activation function (in our case, the leaky rectified linear unit, ReLU[27]). Without an activation function, an arbitrary number of linear layers can always be reduced into a single matrix product as in Equation 7.

CNN architectures are effective in identifying spatial patterns, making them well-suited for image analysis; each

Please cite this article in press as: de Sousa et al., Nondestructive characterization of laser-cooled atoms using machine learning, Newton (2026), https://doi.org/10.1016/j.newton.2026.100518

convolutional layer convolves its input with one or more learned feature kernels, applies a non-linear activation function, and pools the outputs. Convolutional layers identify features in a translationally invariant way and, in conventional implementations such as ours, successively down-sample the image resolution.

Similar to the MLP model, our CNN also employs the intermediate fusion approach, here with two parallel series of convolutional layers. The outputs of these layers are flattened, concatenated into a single vector, and passed through a series of fully connected layers, yielding a 4D output. These fully connected stages are not translationally invariant, implying that our overall CNN model can also learn information related to where features reside in the images.

Data augmentation

We employ data augmentation via geometric transformations to increase the robustness of our models to out-of-distribution data and the stability of the training. Both fluorescence images could be unaltered (U), randomly reflected (R), randomly translated (T), or both (RT). These transformations are physically realistic, as if the 3D atom cloud giving rise to the images has itself been reflected and/or translated in 3D space. Augmentations are re-randomized for each training epoch.

An ideal MOT would be symmetric under reflection about planes normal to the three Cartesian axes, generated by operators $R_{x,y,z}$. Deviations from this ideal configuration are dominated by misalignment and power imbalances between the lasers, both of which tend to drift on the month timescale; gravitational effects are negligible for ^{39}K cooled on the D2 line.

As a result, our reflection-augmentation is implemented by elements of the group generated by these (commuting) operators $\{I,R_x,R_y,R_z,R_{xy},R_{yz},R_{zx},R_{xyz}\}$, where I is the identity; each group element R obeys $R^2 = I$ and, for example, $R_{xy} \equiv R_x R_y$. We do not include other symmetry-allowed operations, such as rotations about $\mathbf{e}_z$ because there is insufficient information in our two fluorescence images to generate such data (such operations correspond to observations of the cloud along arbitrary axes in the $\mathbf{e}_x$-$\mathbf{e}_y$ plane). When this augmentation is employed, a different randomly selected operator (including the identity) is applied to each element of the training dataset.

Although our system is not translationally invariant, the observed position of the laser-cooled atoms on the sensors results from the manual alignment of the imaging systems and is prone to small changes when the system is reconfigured. In addition, variability in the alignment of the MOT lasers as well as stray magnetic fields lead to translations of the cloud. We therefore augment via 3D translations of the atomic cloud, constrained so that the integrated signal is reduced by no more than 10%; this confines translations to a cube of side 6 mm.

When this augmentation is active, we randomly select displacement vectors where each component is uniformly distributed within the allowed domain.

Last, we combine both forms of augmentation by implementing a reflection followed by a translation. Both the reflection plane and the translation vector are drawn at random from the distributions described above.

Training

All models are implemented in PyTorch[28] and are fully detailed in the supplemental information. Of these models, only the MLP and CNN have tunable hyperparameters (e.g., number of layers, layer size, and kernel size), and we selected their values heuristically to obtain performant outcomes. It is likely that careful hyperparameter optimization will improve the performance of fully trained models.

Every model is then separately trained on data augmented using each strategy discussed under Data augmentation.

We employ mini-batch gradient descent using the Adam optimizer and an adaptive learning rate reduction scheme with a base learning rate of 10^{-4}. For consistency, all models are trained for 4,000 epochs; however, this value serves only as a fixed stopping point. During training, both training and validation losses are continuously monitored, and the model achieving the minimum validation loss is retained (equivalent to a validation-based stopping condition).

We observe that the number of epochs required before saturation varies drastically depending on the model type. For example, the LIN model training saturates after about 20 epochs, while the MLP improvement slows around 1,000 epochs. In all cases, the validation data confirm that overfitting is not occurring (supplemental information).

Model performance

This section cross-compares performance for each model architecture and augmentation. Training is performed via 10-fold cross-validation (with randomly selected folds), yielding 10 trained models for every architecture/augmentation combination. Unless otherwise stated, we report the average result of all 10 models with uncertainties given by the standard deviation.

Figure 5 provides a high-level summary of the results for models trained and tested with RT augmentation. Figure 5A shows the marginal distributions of residual vectors ℓ (Equation 4), obtained by projecting the underlying 4D distribution onto all possible 1D axes. This construction yields unbiased, axis-independent 1D distributions whose widths systematically decrease as model sophistication increases from CON to CNN. We quantify the relative change in width using the fraction of variance unexplained (FVU = 1-R^2 in terms of the weighted coefficient of determination R^2), which is directly related to our loss function by $FVU = L^2/L_C^2$. Here L^2 is the total loss for a specific model, and L_C^2 is that of the trivial CON model (thus, FVU$\equiv$1 for the CON model). Thus, FVU measures the fraction of the variance that is unexplained by the model relative to the constant baseline, a quantification of the information learned by the model.

Next, Figures 5B and 5C plot the inferred number and temperature as a function of the corresponding labels, with solid lines marking the desired one-to-one correspondence. Markers are colored according to model architecture, with intensity given by the SNR. For the inferred number in Figure 5B, the deviation from the lines reduces for successive model architectures (from left to right). It is not surprising that even the LIN and MM models show a modest degree of correlation because, everything else being equal, the overall amount of light scattered during laser cooling increases with atom number. This contrasts

Please cite this article in press as: de Sousa et al., Nondestructive characterization of laser-cooled atoms using machine learning, Newton (2026), https://doi.org/10.1016/j.newton.2026.100518

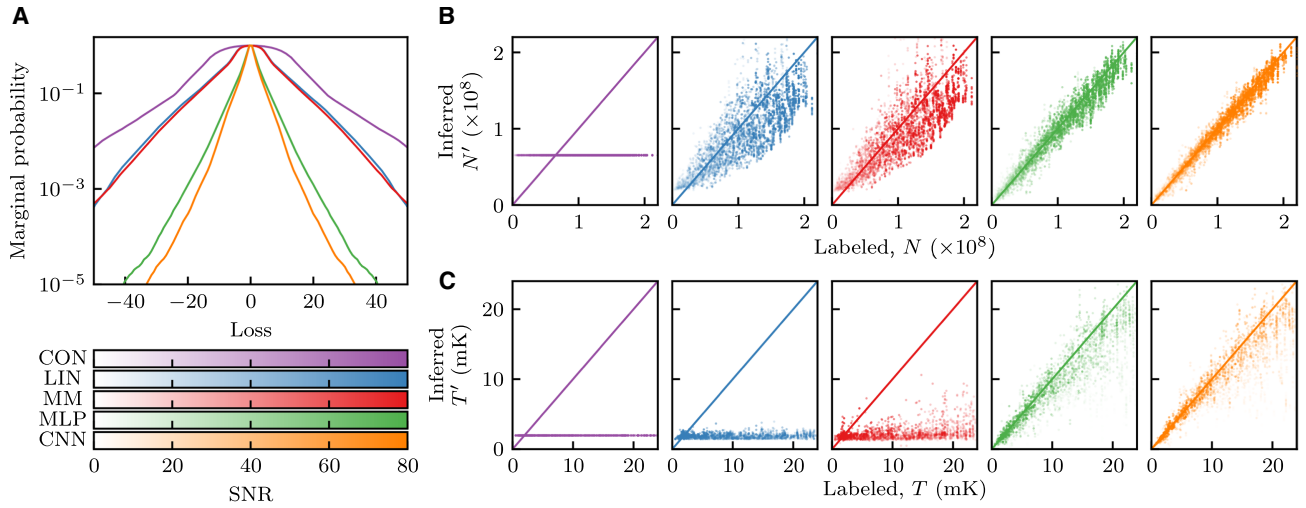


Figure 5. Inference for all models

The models are trained using RT augmentation and compared to the in-distribution test set with RT augmentation; the models are distinguished by color as in the color bar: CON (purple), LIN (blue), MM (red), MLP (green), and CNN (orange) models. The test data were augmented 10-fold, assuring that each element of the test dataset sampled a diverse set of augmented configurations.

To focus the effect of the test distribution's statistics, we evaluate a randomly selected model from those trained in the 10-fold cross-validation. (A) Marginal distributions of residual vectors ℓ.

(B and C) Number and temperature inference, respectively.

Points are shaded according to their SNR (color bars), and lines of slope 1 indicate the ideal behavior.

with temperature inference in Figure 5C, which shows that the LIN and MM models hardly improve upon the CON model. Only the MLP and CNN models have significant predictive power.

For these MLP and CNN architectures, the error in number is largely independent of N, while that of temperature is proportional to T. We therefore report performance in terms of FVU, the RMS number error ΔN, and the RMS fractional temperature error $\mathcal{F}_T = \Delta T/T$. The key numerical results are summarized in Table 3., highlighting models trained using U and RT augmentation and tested against RT-augmented data. For the brave of heart, complete results are tabulated in the supplemental information.

Impact of model architecture and augmentation

We continue with a more detailed discussion of performance as it relates to the choice of both model architecture and augmentation. Figure 6 plots FVU, ΔN, and $\mathcal{F}_T$, for which smaller values indicate improved performance. Model architectures are presented in columns, testing augmentations are indicated on horizontal axes, and training augmentations are distinguished by color. Because the three performance metrics have similar overall trends, we focus our discussion on FVU and reserve discussion of ΔN and $\mathcal{F}_T$ for cases when they exhibit noteworthy behavior.

The CON model sets the baseline performance; to facilitate comparison to other models, its outcomes are delineated by the bottom of the gray-shaded areas in Figure 6. The outcomes, FVU = 1 (by definition), $\Delta N \approx 44 \times 10^6$, and $\mathcal{F}_T \approx 0.76$ are properties of the distribution of labels and therefore independent of augmentation.

Figure 6A shows that the LIN model's performance is nearly independent of augmentation, degrading only slightly for T and RT. This is because the translation operation (and, to a much lesser degree, reflection) can shift a small fraction of the fluorescence signal out of the image. The overall FVU values are slightly below 1, an improvement compared to CON, though most of this improvement results from the reduction in ΔN. The LIN model has essentially the same accuracy as the CON model for $\mathcal{F}_T$, showing that total fluorescence has no obvious information

Table 3. Summary of results

	CON	LIN		MM		MLP		CNN	
	–	U	RT	U	RT	U	RT	U	RT
FVU	1 (exact)	0.70(2)	0.563(1)	$3.3(2)\times 10^2$	0.522(3)	$-1.2(9)\times 10^2$	0.078(4)	3.8(6)	0.052(3)
$\Delta N \times 10^{-6}$	43.791(3)	27.7(2)	23.82(5)	285(9)	23.76(7)	$5(1)\times 10^1$	10.4(4)	48(2)	7.5(2)
$\mathcal{F}_T$	0.75542(6)	0.758(3)	0.7484(4)	11.9(2)	0.7302(9)	7(3)	0.39(3)	1.4(1)	0.31(3)

Models were trained with either U or RT augmentation and evaluated on RT-augmented in-distribution test data. Augmenting the test data enlarged the dataset by $\times 10$. The quoted values reflect the average and standard deviation of the model outcomes based on a 10-fold cross-validation.

Please cite this article in press as: de Sousa et al., Nondestructive characterization of laser-cooled atoms using machine learning, Newton (2026), https://doi.org/10.1016/j.newton.2026.100518

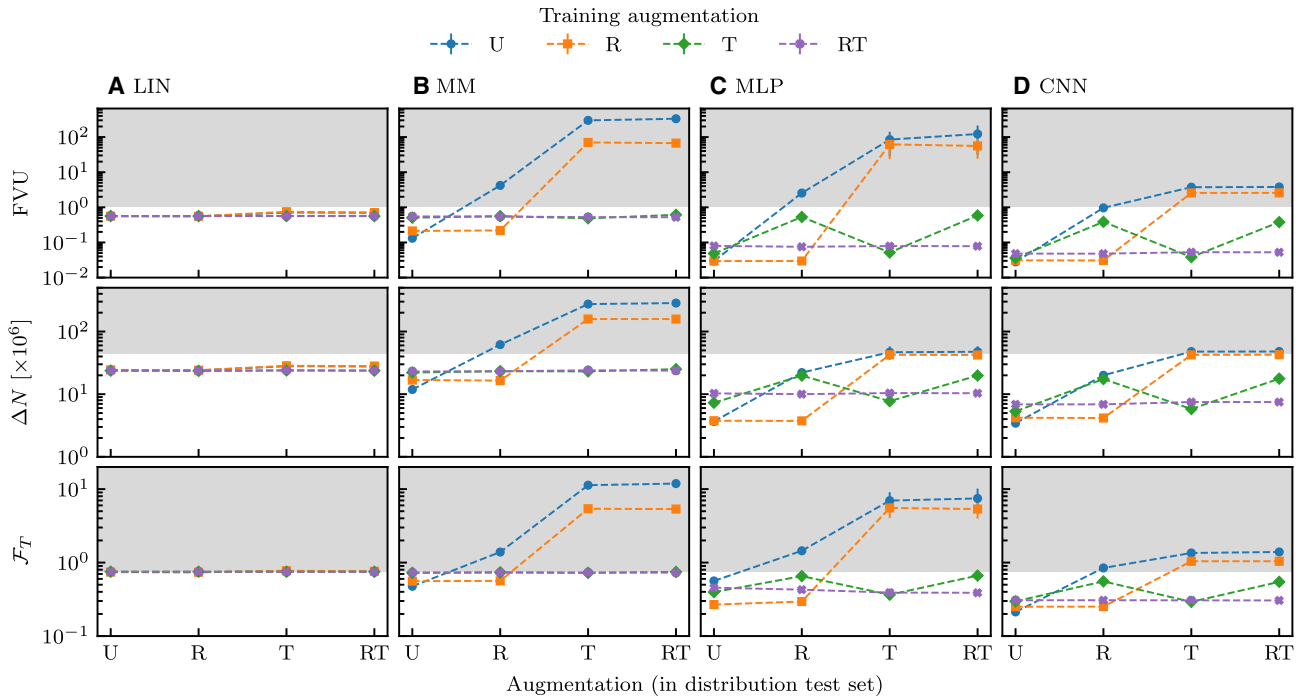


Figure 6. Model performance on in-distribution test data

Columns (A)–(D) denote results for LIN, MM, MLP, and CNN, respectively; rows plot the three performance metrics FVU, ΔN, and $\mathcal{F}_T$; and the horizontal axes mark the different testing augmentations. The gray-shaded regions indicate performance below that of the CON model. The choice of training augmentation is denoted by the marker color: U (blue), R (orange), T (green), and RT (purple). Each point reflects the average from a 10-fold cross-validation, with error bars indicating the standard deviation.

about temperature. These observations signify the slight correlation in Figure 5B or number and the lack thereof in Figure 5C for temperature.

As seen in Figure 6B for the MM model, certain combinations of training and testing augmentation yield performance exceeding that of the LIN model. Specifically, U training with U testing and R training with R or U testing are markedly improved. However, U and R training dramatically worsen performance in all other test cases, underperforming even the simple CON model. By contrast, T- and RT-trained models have indistinguishable performance that is robust across all test cases but has reduced to that of the LIN model in Figure 6A.

Together, these results demonstrate that the spatial structure present in fluorescence images (as in Figure 3) contains information relating to both atom number and temperature; we comment further on this under Discussion. Because the only translationally invariant MM kernels consist of constant entries, this added information is erased by T and RT augmentation. This furthermore suggests that spatial patterns present in MM kernels obtained for U and R training are incompatible with translated data, leading to the worsened performance discussed above.

The MLP results in Figure 6C display further improvement but with the same overall dependency on augmentation as the MM model. A key difference is that models trained with T augmentation no longer perform well on R- and RT-augmented test data, indicating that these datasets contain learnable information violating the expected reflection symmetry.

Mirroring the MM results, RT-trained MLP models are robust, with performance that is essentially independent of test augmentation. MLP models dramatically improve performance, with FVU, ΔN, and $\mathcal{F}_T$ all exceeding the best-case MM models. Thus, unlike the linear MM model, the non-linear activation functions between the MLP layers enable information regarding spatial structure to be retained even with T- and RT-augmented training.

A final noteworthy observation is that, for MLP models trained and tested without augmentation, $\mathcal{F}_T$ is the worst (largest) outcome of any training augmentation (Figure 6C, bottom). Nevertheless for U training, the overall loss ($\propto$FVU) for U testing does not exceed that of the other test configurations. While it may seem surprising, it results from our optimization of uncertainty-weighted, not absolute, residuals.

Figure 6D concludes with our CNN models; as compared to the MLP models, these have a similar augmentation dependence but improved performance. The CNN architecture consists of a set of translationally invariant convolutional input layers followed by densely connected output layers. Together, these features yield a smaller decrease in performance for U- and R-trained models evaluated on T and RT test data while still retaining some information regarding absolute position, making U and T augmentation inequivalent. The RT-trained CNN achieves a performance of FVU $\approx$ 0.05 for all augmentations: the best of our models.

Please cite this article in press as: de Sousa et al., Nondestructive characterization of laser-cooled atoms using machine learning, Newton (2026), https://doi.org/10.1016/j.newton.2026.100518

Table 4. Out-of-distribution data

	CON	LIN		MM		MLP		CNN	
	–	U	RT	U	RT	U	RT	U	RT
L_{in}	691(6)	387(4)	389(4)	91(1)	381(4)	20.5(3)	55(1)	19.4(3)	33.3(4)
L_{out}	721(6)	380(3)	388(3)	106(1)	374(4)	57.6(6)	103(2)	63(1)	72(1)
$(L_{out} - L_{in})/L_{in}$	0.04(1)	−0.02(1)	0.00(1)	0.16(2)	−0.02(2)	1.81(4)	0.87(3)	2.24(5)	1.17(3)

Models were trained with either U or RT augmentation, and we report the overall loss computed for U-augmented in- and out-of-distribution test datasets. Values reflect the average of the model outcomes based on a 10-fold cross-validation, and uncertainties represent the sample standard deviation of the $\approx 3 \times 10^3$ element test datasets evaluated across the 10 folds.

Out-of-distribution data

Our augmentation process was designed to simulate variability that could, in principle, result from drifting external experimental parameters such as ambient magnetic fields, laser power, or optical alignment. The out-of-distribution test dataset discussed under Data organization allows us to assess the impact of augmentation. Because the out-of-distribution data were unseen until final training was complete, all decisions regarding our physically motivated augmentations were uninformed by out-of-distribution performance.

The data in Table 4 compare the overall loss L of models trained with either U augmentation or RT augmentation evaluated on both in- and out-of-distribution U-augmented data. Note that degraded performance results in an increase in L. First, performance decreases by a modest but statistically meaningful amount for the CON model, implying that the out-of-distribution data are drawn from a distinguishably distinct distribution. For the LIN model, this decrease vanishes for either training augmentation, both of which show performance differences consistent with zero.

The remaining models (MM, MLP, and CNN) have degraded performance for both augmentations. In each of these cases, the RT-trained models suffer a smaller fractional reduction in their performance. From the perspective of overall loss, the MLP and CNN models are each impacted by a similar amount, independent of augmentation strategy. Indeed, from this more global perspective, the U-trained models outperform the RT-trained models both in- and out-of-distribution data. From this, we conclude that most of the variability reflected by the out-of-distribution dataset is not captured by our augmentations or the range of our parameter scans.

DISCUSSION

In this work, we explored the utility of ML techniques for extracting relevant characteristics of atoms in an MOT, such as their number and temperature, from non-destructive fluorescence images. We began by creating a labeled dataset with $\approx 39 \times 10^3$ elements from laser-cooled ^{39}K atoms in an MOT. Each element of the dataset contains a pair of fluorescence images and a follow-up destructive absorption image acquired after a short TOF. Atom number and temperature labels were obtained from the absorption images. We investigated five ML models with a range of complexities to estimate these parameters from the fluorescence images alone. The training procedure optionally included data augmentation that combined reflections and translations of the fluorescence images.

Our first model, the trivial case with constant outputs (CON), served as the baseline to which the remaining models were compared. The next two models were linear: a simple linear function (LIN) of the summed fluorescence counts and a single fully connected layer (MM). Both of these models improved upon the CON model for number inference, while only the MM model showed improvement for temperature inference. The two non-linear models, an MLP and a CNN, further improved inference for both number and temperature, with the CNN performing best and most robustly for all training and testing configurations.

Of our models, only the CNN operates on 2D data; the process of flattening images to vectors (for MM and MLP) removes implicit information regarding spatial structure; integrating (for LIN) discards yet more information. A key strength of CNN models is that they operate directly on image data (leading to effective training on smaller datasets), while dense models must learn embeddings of spatial properties in their otherwise unstructured weights (requiring larger datasets).

The best CNN model predicts number with an uncertainty of 4×10^6 and temperature with a fractional uncertainty of 0.2. Interestingly, these regression uncertainties are below the estimated uncertainties of the labels, 6×10^6 and 0.5, respectively. As noted under Data, the observed TOF density distributions used for labeling are often poorly described by our simple Gaussian model, leading to inflated uncertainty estimates. This confirms the opportunity for significant improvement in labeling.

During the final preparation of this manuscript, and after unblinding the test data, we initiated a 4×10^5 epoch training run for the fully augmented CNN model (10-fold cross-validation requires about 20 weeks on a single NVIDIA GeForce RTX 4080 16 GB GDDR6X). This shows that, while the loss first plateaus at about 300 epochs, both the training and validation loss begin dropping again after 10^4 epochs. After the full 4×10^5 epochs, the validation loss fell by an additional factor of five compared to the results presented above, and no sign of saturation was observed. This underscores the potential improvements from model optimization and improved training methodology.

The MM, MLP, and CNN models all utilized the atoms' spatial distribution for improved inference of both number and temperature. Incorporating any form of translation augmentation into the training data degraded the MM model's capability for temperature inference to the level of the LIN and CON models, therefore implying that spatial structure is the only source of temperature information.

For each output, the MM model operates by learning a kernel that multiplies fluorescence images in a pixel-by-pixel manner.

Please cite this article in press as: de Sousa et al., Nondestructive characterization of laser-cooled atoms using machine learning, Newton (2026), https://doi.org/10.1016/j.newton.2026.100518

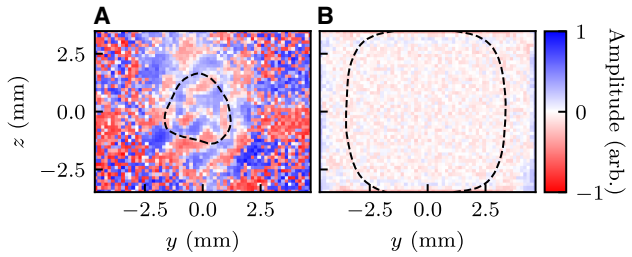


Figure 7. MM model kernels for temperature inference

The model is trained on (A) unaugmented data and (B) data with combined reflection and translation augmentations. Kernels are for $\mathbf{e}_x$ fluorescence images, and the black dashed curve encloses regions with significant fluorescence signal. Its increase in size in (B) is a consequence of the translation augmentation.

Typical kernels for temperature inference are visualized in Figure 7, both without augmentation (Figure 7A) and with combined reflection and translation augmentation (Figure 7B). The dashed curves outline the region where fluorescence images show significant signal. The kernel from unaugmented training data in Figure 7A has significant spatial structure that is erased by the use of augmentation in Figure 7B, thereby recovering the simple summation employed in the LIN model.

Although these kernels directly visualize the MM model's operation, they do not suggest an underlying physical mechanism. Furthermore, neither of the higher performing non-linear models can be interpreted even in this limited way. This suggests the importance of exploring ML techniques targeting physical dynamics or interpretability, such as symbolic regression[29] or explainable boosting machines,[30,31] respectively.

Even on a low-end graphical processing unit (GPU), the models presented here can perform inference in ≈0.5 ms. This enables real-time applications because it is far below the typical $\gtrsim$5 ms timescale of MOT dynamics. Such ML tools both provide diagnostic access to quantities that otherwise require destructive measurements and open new pathways for real-time feedback control of laser-cooled atoms operating in novel parameter regimens.[32]

The inherent complexity of quantum platforms, from system and state preparation to control and finally measurement, makes them an ideal use case for ML-based information extraction and ML-enabled optimal control. Our work is therefore a significant step in these directions, giving demonstrable access to otherwise hidden information, further motivating the use of ML methods in cold-atom-based platforms and in quantum science and technology broadly speaking.

METHODS

Imaging details

The raw fluorescence images are recorded by briefly collecting light emitted by the laser-cooled atoms onto our $\mathbf{e}_x$ and $\mathbf{e}_z$ cameras; these images inevitably include a non-negligible background of scattered laser light and ambient room light. To mitigate this, we subtract a second reference frame taken under otherwise identical conditions but without atoms.

Each TOF absorption image of 2D column density $\rho_{\mathbf{i}}$ is assembled from three image frames in raw ADUs (analog to digital units) from the $\mathbf{e}_y$ camera: an "atoms" frame $N_{\mathbf{i}}^{\mathrm{at}}$, a probe-only frame $N_{\mathbf{i}}^{\mathrm{pr}}$, and a background frame $N_{\mathbf{i}}^{\mathrm{bg}}$. Here, we label pixels by their vector coordinate $\mathbf{i} = (i,j)$. In the atoms' frame, the TOF-expanded cloud is illuminated by the probe laser for 10 μs. The probe frame is a replica of the atoms' frame but with no atoms present. Finally, we record the background frame with the probe laser turned off but all other conditions kept the same.

Together, these frames yield the column density

$$\rho_{\mathbf{i}} = -\frac{1}{\sigma_0}\left[\log\left(\frac{N_{\mathbf{i}}^{\mathrm{at}} - N_{\mathbf{i}}^{\mathrm{bg}}}{N_{\mathbf{i}}^{\mathrm{pr}} - N_{\mathbf{i}}^{\mathrm{bg}}}\right) - \frac{N_{\mathbf{i}}^{\mathrm{at}} - N_{\mathbf{i}}^{\mathrm{pr}}}{N_{\mathrm{sat}}}\right], \quad \text{(Equation 8)}$$

where $\sigma_0 = 3\lambda^2/(2\pi)$ is the resonant scattering cross-sections for the imaging transition at wavelength $\lambda = 766.7$ nm, and $N_{\mathrm{sat}} = 200$ is the number of ADUs[33,34] corresponding to the saturation intensity $I_{\mathrm{sat}} = 1.75$ mW/cm^2.

Parameter conversions

Here, we document the conversions from control parameters to physical parameters for the quadrupole current to magnetic field gradient as well as the AOM command for the cooling and repump lasers, specifying their power and intensity.

The calibration of quadrupole current I to magnetic field gradient $\mathbf{B}'$ from the MOT's quadrupole coils is $\mathbf{B}' = I \times [5.3(2),10.5(4),5.3(2)] \times 10^{-3}$T m^{-1}A^{-1}. This is determined by measuring the minimum current at which ^{39}K atoms can be magnetically trapped. Our standard MOT operates with I = 15.2 A, yielding a strong-axis gradient of 0.160(6) T m^{-1}.

AOMs vary the intensity of the incident laser light by Bragg scattering the optical field off a traveling acoustic wave; for a single-pass setup (like our main cooling laser), this process is described by a sin^2 law for low drives and transitions to J_1^2 for very strong drives (where J_1 is a Bessel function of the first kind). For the purpose of obtaining a calibration, the Bessel function model satisfactorily fits the data over our whole control range with the definition

$$P = P_0 J_1^2[A(V_{\mathrm{AOM}} + \delta V)], \quad \text{(Equation 9)}$$

with fit parameters P_0, A, and δV. For a double-pass setup like our repumper, the Bessel function is raised to the fourth power. During dataset collection, power measurements were recorded for standard shot control voltages yielding the data shown in Figure 8. We cross-calibrated by measuring the power of one of the six beams as a function of its respective AOM control voltage for both the cooling and repump lasers. These were then fit to Equation 9, yielding the fitting parameters shown in Table 5. With measured beam waist w = 13(3) mm ($1/e^2$ radius), this gives per-beam peak intensities of 15(4) mW cm^{-2} and 3.4(8) mW cm^{-2} for the cooling and repump beams, respectively.

Number and temperature labeling

The number and temperature labels are determined from joint fits to all M absorption images within a given set (labeled by $k = 1, \ldots, M$) of column density $\rho_{\mathbf{i}}^{(k)}$ to a TOF-dependent Gaussian profile whose width increases in time $t_{\mathrm{TOF}}^{(k)}$.

Please cite this article in press as: de Sousa et al., Nondestructive characterization of laser-cooled atoms using machine learning, Newton (2026), https://doi.org/10.1016/j.newton.2026.100518

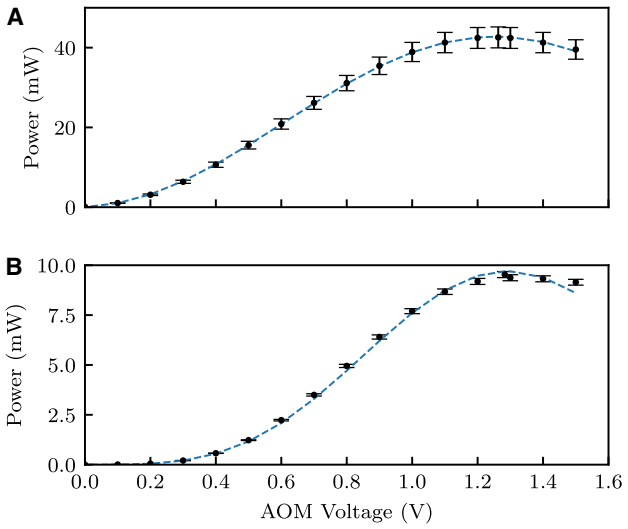


Figure 8. Per-beam power and intensity calibrations

Results for the cooling and repump lasers are shown in (A) and (B), respectively. Markers indicate measurements, and the error bars reflect the variation in power due to day-to-day experimental drift. The dashed curves are fits to Equation 9, with parameters listed in Table 5.

We model the initial $t_{\mathrm{TOF}} = 0$ density distribution as a Gaussian profile and assume that the velocity is thermally distributed (i.e., also Gaussian). The TOF expansion along the $\mathbf{e}_x$ and $\mathbf{e}_z$ directions (observed by our $\mathbf{e}_y$ TOF camera) is observed to be anisotropic, necessitating different temperature parameters T_x and T_z. Altogether, this leads to a model distribution that is Gaussian at all times with widths

$$w_j^2(t_{\mathrm{TOF}}) = w_j^2(0) + \frac{k_{\mathrm{B}} T_j}{m} t_{\mathrm{TOF}}^2, \text{with} \quad j \in \{x, z\}, \tag{Equation 10}$$

where $w_j(0)$ denotes the initial size of the cloud. In what follows, we generalize the notation $t_{\mathrm{TOF}}^{(k)}$ introduced for TOF times to compactly label quantities associated with our specific observation times; i.e., $w_j\left(t_{\mathrm{TOF}}^{(k)}\right) \to w_j^{(k)}$. We use (0) to denote (unobserved) quantities at $t_{\mathrm{TOF}} = 0$.

These definitions lead to the fit function for TOF density distributions imaged on our CCD camera,

$$n_{\mathbf{i}}^{(k)} = N\left[2\pi \prod_{j \in x,z} w_j^{(k)}\right]^{-1} \exp\left[-\frac{1}{2}\sum_{j \in x,z}\left(\frac{r_j - r_{0,j}^{(k)}}{w_j^{(k)}}\right)^2\right], \tag{Equation 11}$$

where pixels sample positions $(x,z) \equiv \mathbf{r} = \Delta x \mathbf{i}$; Δx is the magnified and down-sampled pixel size, and $r_{0,j}^{(k)}$ is the spatial center of the distribution. The spatial center

$$\mathbf{r}_0^{(k)} = \mathbf{r}_0 + \mathbf{v} t_{\mathrm{TOF}}^{(k)} + \delta_{k,1} \mathbf{r}_{\mathrm{step}} \tag{Equation 12}$$

was specified by fit parameters: an initial position $\mathbf{r}_0$, an initial velocity $\mathbf{v}$, and a small offset $\mathbf{r}_{\mathrm{step}}$ ($\delta_{k,1}$ is the Kronecker-δ function). We introduced $\mathbf{r}_{\mathrm{step}}$ to account for a systematic shift in the center position of the cloud in the 1 ms images; this shift was found to be only along $\mathbf{e}_z$.

Table 5. AOM calibration parameters

Laser	P_0 (mW)	A (V^{-1})	δV (mV)
Cooling	126(8)	1.441(6)	24(3)
Repump	85(3)	1.40(1)	24(constrained)

Because the double-pass AOM response is flat near $V_{\mathrm{AOM}} = 0$, the fit fails to converge when δV is free; we therefore constrain it to the value determined from the cooling fits.

We then minimize the objective function,

$$E = \sum_k \sum_{\mathbf{i}} \left[\frac{\rho_{\mathbf{i}}^{(k)} - n_{\mathbf{i}}^{(k)}}{\Sigma_{\mathbf{i}}^{(k)}}\right]^2, \tag{Equation 13}$$

to determine the fit parameters; each contribution to E is weighted by pixel-by-pixel uncertainties $\Sigma_{\mathbf{i}}^{(k)}$ derived from photon shot noise in the individual images used in absorption imaging. We minimized with respect to the parameters $\left(N, T_x, T_z, w_x^{(0)}, w_z^{(0)}, r_{0,z}, v_z, r_{\mathrm{step},z}\right)$ and found that the remaining parameters $(r_{0,x}, v_x, r_{\mathrm{step},x})$ were constant across the dataset, within the uncertainties. This led to our final labels N and $T = (T_x + T_z)/2$.

The F1 measure

Data are labeled as containing an MOT when the overall fluorescence signal exceeds the background. This requires determining a threshold value that results in a classification model that accurately identifies all instances of MOTs. To do this, we use the F1 measure, which quantifies the number of true positives (TPs) relative to the number of false positives (FPs) and false negatives (FNs). In our context, a TP corresponds to counts above the threshold for the pink difference distribution in Figure 4. Similarly, a TN corresponds to counts below the threshold for the no-atom distribution. Since there will be overlap between the two distributions, some false identifications will occur. An FP will occur for counts above the threshold on the no-atoms distribution, and an FN will occur for counts on the difference distribution below the threshold. In terms of these parameters, the F1 measure is defined by

$$F_1 = \frac{2\mathrm{TP}}{2\mathrm{TP} + \mathrm{FP} + \mathrm{FN}}. \tag{Equation 14}$$

This equally weighs a model's ability to produce correct identifications when it provides one (its precision) and its ability to correctly identify all relevant identifications (its recall). Therefore, when the threshold is chosen to maximize F_1, the model performs well on both metrics.

RESOURCE AVAILABILITY

Lead contact

Requests for further information and resources should be directed to and will be fulfilled by the lead contact, I.B. Spielman (ian.spielman@nist.gov).

Materials availability

This study did not generate new materials.

Data and code availability

- The experimental datasets acquired and analyzed during the current study are publicly available at doi:10.5281/zenodo.19340423.

- Relevant code describing the models is directly provided in the supplemental information.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

ACKNOWLEDGMENTS

The authors thank E.B. Norrgard and D. Schug for carefully reading the manuscript. F. Salces-Carcoba created the CAD file used in Figure 1A. This work was partially supported by the National Institute of Standards and Technology and the National Science Foundation through the Physics Frontier Center at the Joint Quantum Institute (PHY-1430094) and the Quantum Leap Challenge Institute for Robust Quantum Simulation (OMA-2120757). G.d.S. acknowledges the São Paulo Research Foundation (2024/20892-8).

AUTHOR CONTRIBUTIONS

Conceptualization, J.P.Z. and I.B.S.; methodology, G.d.S., M.D., D.D., J.P.Z., and I.B.S.; investigation, G.d.S., M.D., D.D., and B.E.; writing – original draft, G.d.S., M.D., D.D., and B.E.; writing – review and editing, G.d.S., M.D., D.D., B.E., J.P.Z., and I.B.S.; funding acquisition, J.P.Z. and I.B.S.; resources, J.P.Z. and I.B.S.; supervision, J.P.Z. and I.B.S.

DECLARATION OF INTERESTS

J.P.Z. is an advisory board member for *Newton*.

DECLARATION OF GENERATIVE AI AND AI-ASSISTED TECHNOLOGIES IN THE WRITING PROCESS

During the preparation of this work, the authors used Grammarly to proofread text and ChatGPT to provide adversarial critiques. After using this tool or service, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.newton.2026.100518.

Received: October 9, 2025
Revised: March 9, 2026
Accepted: April 17, 2026

REFERENCES

1. Carleo, G., Cirac, I., Cranmer, K., Daudet, L., Schuld, M., Tishby, N., Vogt-Maranto, L., and Zdeborová, L. (2019). Machine learning and the physical sciences. Rev. Mod. Phys. *91*, 045002.
2. Degrave, J., Felici, F., Buchli, J., Neunert, M., Tracey, B., Carpanese, F., Ewalds, T., Hafner, R., Abdolmaleki, A., de Las Casas, D., et al. (2022). Magnetic control of tokamak plasmas through deep reinforcement learning. Nature *602*, 414–419.
3. Zwolak, J.P., and Taylor, J.M. (2023). Colloquium: Advances in automation of quantum dot devices control. Rev. Mod. Phys. *95*, 011006.
4. Medeiros, L., Psaltis, D., Lauer, T.R., and Özel, F. (2023). The Image of the M87 Black Hole Reconstructed with PRIMO. Astrophys. J. Lett. *947*, L7.
5. Merchant, A., Batzner, S., Schoenholz, S.S., Aykol, M., Cheon, G., and Cubuk, E.D. (2023). Scaling deep learning for materials discovery. Nature *624*, 80–85.
6. Karagiorgi, G., Kasieczka, G., Kravitz, S., Nachman, B., and Shih, D. (2022). Machine learning in the search for new fundamental physics. Nat. Rev. Phys. *4*, 399–412.
7. Hasan, M.Z., and Kane, C.L. (2010). Colloquium: Topological insulators. Rev. Mod. Phys. *82*, 3045–3067.
8. Carrasquilla, J., and Melko, R.G. (2017). Machine learning phases of matter. Nat. Phys. *13*, 431–434.
9. Rem, B.S., Käming, N., Tarnowski, M., Asteria, L., Fläschner, N., Becker, C., Sengstock, K., and Weitenberg, C. (2019). Identifying quantum phase transitions using artificial neural networks on experimental data. Nat. Phys. *15*, 917–920.
10. Hänsch, T.W., and Schawlow, A.L. (1975). Cooling of gases by laser radiation. Opt. Commun. *13*, 68–69.
11. Raab, E.L., Prentiss, M., Cable, A., Chu, S., and Pritchard, D.E. (1987). Trapping of neutral sodium atoms with radiation pressure. Phys. Rev. Lett. *59*, 2631–2634.
12. Ludlow, A.D., Boyd, M.M., Ye, J., Peik, E., and Schmidt, P.O. (2015). Optical atomic clocks. Rev. Mod. Phys. *87*, 637–701.
13. Kaufman, A.M., and Ni, K.K. (2021). Quantum science with optical tweezer arrays of ultracold atoms and molecules. Nat. Phys. *17*, 1324–1333.
14. Anderson, M.H., Ensher, J.R., Matthews, M.R., Wieman, C.E., and Cornell, E.A. (1995). Observation of Bose-Einstein condensation in a dilute atomic vapor. Science *269*, 198–201.
15. Roslund, J.D., Cingöz, A., Lunden, W.D., Partridge, G.B., Kowligy, A.S., Roller, F., Sheredy, D.B., Skulason, G.E., Song, J.P., Abo-Shaeer, J.R., and Boyd, M.M. (2024). Optical clocks at sea. Nature *628*, 736–740.
16. Tranter, A.D., Slatyer, H.J., Hush, M.R., Leung, A.C., Everett, J.L., Paul, K.V., Vernaz-Gris, P., Lam, P.K., Buchler, B.C., and Campbell, G.T. (2018). Multiparameter optimisation of a magneto-optical trap using deep learning. Nat. Commun. *9*, 4360.
17. Ren, Z., Yan, X., Wen, K., Chen, H., Hajiyev, E., He, C., and Jo, G.B. (2024). Creation of a tweezer array for cold atoms utilizing a generative neural network. APL Quantum *1*, 046111.
18. Wigley, P.B., Everitt, P.J., van den Hengel, A., Bastian, J.W., Sooriyabandara, M.A., McDonald, G.D., Hardman, K.S., Quinlivan, C.D., Manju, P., Kuhn, C.C.N., et al. (2016). Fast machine-learning online optimization of ultra-cold-atom experiments. Sci. Rep. *6*, 25890.
19. Vendeiro, Z., Ramette, J., Rudelis, A., Chong, M., Sinclair, J., Stewart, L., Urvoy, A., and Vuletić, V. (2022). Machine-learning-accelerated Bose-Einstein condensation. Phys. Rev. Res. *4*, 043216.
20. Metcalf, H.J., and van der Straten, P. (1999). Laser Cooling and Trapping (Springer-Verlag).
21. Lett, P.D., Watts, R.N., Westbrook, C.I., Phillips, W.D., Gould, P.L., and Metcalf, H.J. (1988). Observation of atoms laser cooled below the doppler limit. Phys. Rev. Lett. *61*, 169–172.
22. Griffiths, J., Wrathmall, S.A., and Gardiner, S.A. (2025). Single-shot thermometry of simulated Bose–Einstein condensates using artificial intelligence. Preprint at arXiv. https://arxiv.org/abs/2506.16925.
23. Chu, S., Hollberg, L., Bjorkholm, J.E., Cable, A., and Ashkin, A. (1985). Three-dimensional viscous confinement and cooling of atoms by resonance radiation pressure. Phys. Rev. Lett. *55*, 48–51.
24. Pritchard, D.E., Raab, E.L., Bagnato, V., Wieman, C.E., and Watts, R.N. (1986). Light traps using spontaneous forces. Phys. Rev. Lett. *57*, 310–313.
25. de Sousa, G., Doris, M., D'Amato, D., Egleston, B., Zwolak, J., and Spielman, I. (2026). Dataset underlying the manuscript: Nondestructive characterization of laser-cooled atoms using machine learning. Zenodo. https://doi.org/10.5281/zenodo.19340423.
26. Baltrušaitis, T., Ahuja, C., and Morency, L.P. (2019). Multimodal machine learning: A survey and taxonomy. IEEE Trans. Pattern Anal. Mach. Intell. *41*, 423–443.
27. Maas, A.L., Hannun, A.Y., and Ng, A.Y. (2013). Rectifier nonlinearities improve neural network acoustic models. Proc. ICML *28*, 6.
28. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. (2019). PyTorch: An

Imperative Style, High-Performance Deep Learning Library (Curran Associates Inc).

29. Cranmer, M. (2023). Interpretable machine learning for science with PySR and SymbolicRegression.jl. Preprint at arXiv. https://doi.org/10.48550/arXiv.2305.01582.
30. Lou, Y., Caruana, R., Gehrke, J., and Hooker, G. (2013). Accurate intelligible models with pairwise interactions. In Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD '13 (Association for Computing Machinery), pp. 623–631.
31. Schug, D., Kovach, T.J., Wolfe, M.A., Benson, J., Park, S., Dodson, J.P., Corrigan, J., Eriksson, M.A., and Zwolak, J.P. (2025). Automation of quantum dot measurement analysis via explainable machine learning. Mach. Learn.: Sci. Technol. *6*, 015006.
32. Gaudesius, M., Zhang, Y.C., Pohl, T., Kaiser, R., and Labeyrie, G. (2021). Phase diagram of spatiotemporal instabilities in a large magneto-optical trap. Phys. Rev. A *103*, L041101.
33. Reinaudi, G., Lahaye, T., Wang, Z., and Guéry-Odelin, D. (2007). Strong saturation absorption imaging of dense clouds of ultracold atoms. Opt. Lett. *32*, 3143.
34. Seroka, E.M., Curiel, A.V., Trypogeorgos, D., Lundblad, N., and Spielman, I.B. (2019). Repeated measurements with minimally destructive partial-transfer absorption imaging. Opt. Express *27*, 36611–36624.

NEWTON, Volume *2*

Supplemental information

Nondestructive characterization of laser-cooled atoms using machine learning

Guilherme de Sousa, Michael Doris, Dario D'Amato, Brady Egleston, Justyna P. Zwolak, and Ian B. Spielman

Note S1: TRAINING CURVES

Here we present training curves obtained from the 10-fold cross-validation procedure. All models are trained for a total of 4 000 epochs. Loss values are monitored throughout the training process, and the model that produces the lowest validation score has its weights and biases saved. This is known as checkpointing and allows a model to be reinitialized from a snapshot of its state during training. These checkpoints are essentially validation-based stopping conditions. For each model, we retain the loss metrics and checkpoints for all 10 folds.

A summary of the training and validation behavior for the RT-augmented models is shown in Fig. (S1). As the complexity and capacity of the models increase, the overall loss scores decrease, and their behavior differs. The LIN model saturates around 20 epochs into its training cycle, whereas the other three begin to plateau later. After 20 epochs, the LIN model shows no further improvement, as evidenced by the spread in its checkpoints due to noise around this saturation. Although the increased complexity of the MM model improves the loss score, its minimal change and eventual saturation indicate limited capacity to learn additional information. Like the LIN model, it exhibits a similar spread in checkpoints.

In contrast, the MLP and CNN models began to plateau around 300 epochs, with drastically slowed learning, an effect that is more evident at smaller scales. This is further evidenced by the clustering of checkpoints toward the end of their training cycles and the slow decrease seen in the insets.

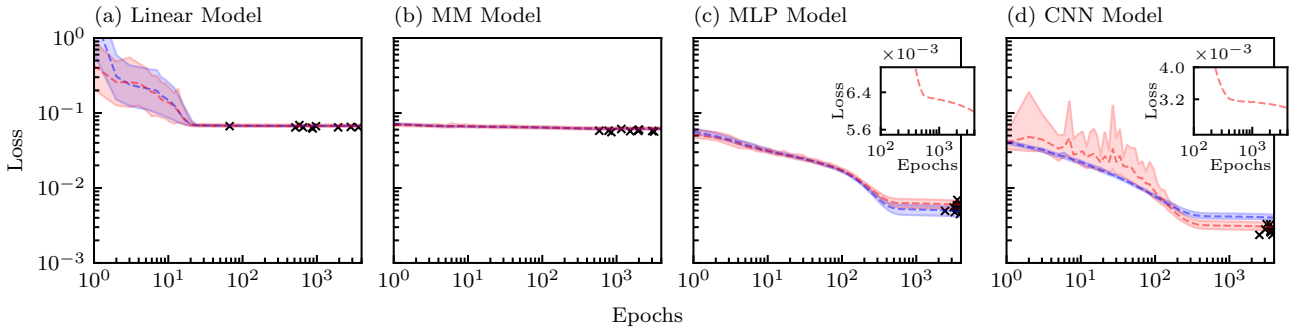


FIG. S1. Loss curves for RT augmented models with 10-fold cross validation. The training (blue) and validation (red) loss curves include the mean (dashed) along with non-symmetric single-sigma error bands. Black markers indicate the final checkpoint for each fold. Insets present the mean validation curve at a smaller scale for the MLP and CNN models after 100 epochs. Data is subsampled to display at most 20 points per decade in the epochs axis.

TABLE S1. Architectures of all models used in this work. All Dense layers are assumed to include bias, and the number inside the parentheses of the Conv2D layer indicates the size of the kernel. The notation LeakyReLU* indicates BatchNorm2D → LeakyReLU → Dropout(0.2), and is used during training.

CON, # parameters = 4			LIN, # parameters = 12			MM, # parameters = 24 580			MLP, # parameters = 7 840 036			CNN, # parameters = 1 397 156		
Layer type	Size	Size	Layer type	Size	Size	Layer type	Size	Size	Layer type	Size	Size	Layer type	Size	Size
Input	-	-	Input	(64, 48)	(64, 48)	Input	(64, 48)	(64, 48)	Input	(64, 48)	(64, 48)	Input	(64, 48)	(64, 48)
Linear	$0 \rightarrow 4$		Flatten	(64, 48)→ 3 072	(64, 48)→ 3 072	Flatten	(64, 48)→ 3 072	(64, 48)→ 3 072	Flatten	(64, 48)→ 3 072	(64, 48)→ 3 072	Conv2D(9)	(1,64,48)→(8,64,48)	(1,64,48)→(8,64,48)
Output	4		Sum	$3\,072 \rightarrow 1$	$3\,072 \rightarrow 1$	Fuse	(3 072,3 072)→6 144		Linear	$3\,072 \rightarrow 1\,024$	$3\,072 \rightarrow 1\,024$	LeakyReLU*	-	-
			Fuse	(1, 1)→2		Linear	$6\,144 \rightarrow 4$		LeakyReLU	-	-	AvgPool2d	(8,64,48)→(8,32,24)	(8,64,48)→(8,32,24)
			Linear	$2 \rightarrow 4$		Output	4		Linear	$1\,024 \rightarrow 512$	$1\,024 \rightarrow 512$	Conv2D(9)	(8,32,24)→(16,32,24)	(8,32,24)→(16,32,24)
			Output	4					LeakyReLU	-	-	LeakyReLU*	-	-
									Linear	$512 \rightarrow 256$	$512 \rightarrow 256$	AvgPool2d	(16,32,24)→(16,16,12)	(16,32,24)→(16,16,12)
									LeakyReLU	-	-	Conv2D(9)	(16,16,12)→(32,16,12)	(16,16,12)→(32,16,12)
									Fuse	$(256, 256) \rightarrow 512$		LeakyReLU*	-	-
									Linear	$512 \rightarrow 256$		AvgPool2d	(32,16,12)→(32,8,6)	(32,16,12)→(32,8,6)
									LeakyReLU	-		Conv2D(9)	(32,8,6)→(64,8,6)	(32,8,6)→(64,8,6)
									Linear	$256 \rightarrow 256$		LeakyReLU*	-	-
									LeakyReLU	-		AvgPool2d	(64,8,6)→(64,4,3)	(64,8,6)→(64,4,3)
									Linear	$256 \rightarrow 128$		Flatten	(64,4,3)→768	(64,4,3)→768
									LeakyReLU	-		Fuse	(768,768)→1 536	
									Linear	$128 \rightarrow 32$		Linear	1 536→512	
									LeakyReLU	-		LeakyReLU	-	
									Linear	$32 \rightarrow 4$		Linear	512→256	
									Output	4		LeakyReLU	-	
												Linear	256→128	
												LeakyReLU	-	
												Linear	128→64	
												LeakyReLU	-	
												Linear	64→4	
												Output	4	

TABLE S2. Complete results for all models, including every training and testing augmentation. Each column contains a model trained with a different augmentation. For a given metric (FVU, ΔT, ΔN, $\mathcal{F}_T$, $\mathcal{F}_N$), each row shows the score of a trained model when tested with a given augmentation (U, T, R, RT). The numbers in parentheses show the most significant digit of the standard deviation.

		In-distribution test set															
		CON	LIN				MM				MLP				CNN			
	Training / Testing	–	U	R	T	RT	U	R	T	RT	U	R	T	RT	U	R	T	RT
FVU	U	0.0	0.440(2)	0.4394(5)	0.4347(4)	0.4364(7)	0.868(1)	0.787(1)	0.493(1)	0.448(2)	0.9703(2)	0.9704(4)	0.951(2)	0.920(6)	0.9719(5)	0.9690(9)	0.964(2)	0.952(2)
	R		0.443(2)	0.4437(5)	0.4354(4)	0.4375(7)	-3.2(2)	0.780(1)	0.439(2)	0.453(1)	-1.6(5)	0.9703(2)	0.47(6)	0.925(4)	0.0(1)	0.9693(8)	0.61(4)	0.952(2)
	T		0.29(2)	0.25(2)	0.435(1)	0.4355(9)	$-3.0(2)\times10^{2}$	-69(3)	0.513(2)	0.474(4)	$-8(6)\times10^{1}$	$-6(4)\times10^{1}$	0.948(2)	0.921(4)	-2.7(5)	-1.6(5)	0.962(2)	0.947(2)
	RT		0.30(2)	0.29(1)	0.436(1)	0.437(1)	$-3.3(2)\times10^{2}$	-66(4)	0.39(2)	0.478(3)	$-1.2(9)\times10^{2}$	$-5(3)\times10^{1}$	0.41(6)	0.922(4)	-2.8(6)	-1.6(6)	0.62(4)	0.948(3)
ΔT [mK]	U	11.011(8)	11.005(5)	10.99(1)	11.196(5)	11.188(8)	7.05(2)	8.32(2)	10.96(2)	10.92(2)	2.70(8)	3.20(6)	4.27(5)	4.66(6)	3.43(3)	3.78(5)	3.9(1)	4.52(8)
	R		10.968(6)	10.95(1)	11.192(5)	11.181(8)	8.75(6)	8.45(2)	11.00(2)	10.96(1)	6.5(2)	3.22(7)	5.56(7)	4.73(5)	6.2(2)	3.79(5)	5.8(1)	4.57(8)
	T		10.79(1)	10.75(1)	11.174(5)	11.150(7)	29.5(6)	16.2(2)	10.73(1)	10.757(8)	18(5)	14(3)	4.46(4)	4.98(5)	8.4(1)	8.4(1)	4.19(9)	4.78(8)
	RT		10.76(1)	10.71(1)	11.168(6)	11.144(7)	30.7(5)	15.9(2)	10.71(2)	10.76(1)	21(7)	14(3)	5.77(7)	5.01(6)	8.6(2)	8.5(1)	5.9(1)	4.80(8)
$\Delta N[\times10^{6}]$	U	43.791(3)	24.31(5)	24.44(4)	23.75(2)	23.77(2)	11.87(7)	16.81(8)	22.17(2)	23.09(3)	3.67(6)	3.8(1)	7.3(2)	10.3(5)	3.4(1)	4.2(1)	5.3(2)	6.9(2)
	R		24.02(5)	24.13(5)	23.57(2)	23.56(2)	62(1)	16.5(1)	23.28(5)	23.16(4)	22(1)	3.75(7)	19.8(6)	10.0(4)	20.0(6)	4.2(1)	17.4(4)	6.9(2)
	T		28.0(2)	28.4(4)	23.94(5)	23.97(5)	274(9)	158(3)	23.11(9)	24.1(1)	$5(1)\times10^{1}$	43(3)	7.7(2)	10.4(4)	48(2)	42(1)	5.8(2)	7.5(2)
	RT		27.7(2)	28.1(2)	23.75(3)	23.82(5)	285(9)	158(3)	25.1(1)	23.76(7)	$5(1)\times10^{1}$	42(3)	19.8(5)	10.4(4)	48(2)	43(1)	17.7(4)	7.5(2)
$\mathcal{F}_T$	U	0.75542(6)	0.7426(2)	0.7427(8)	0.7516(2)	0.7511(5)	0.476(1)	0.559(1)	0.725(1)	0.730(1)	0.56(6)	0.27(2)	0.40(6)	0.46(3)	0.21(2)	0.25(3)	0.30(5)	0.31(2)
	R		0.7391(2)	0.7389(8)	0.7512(2)	0.7504(5)	1.39(4)	0.5639(9)	0.736(1)	0.7330(8)	1.4(1)	0.30(3)	0.65(3)	0.43(3)	0.85(7)	0.25(2)	0.55(4)	0.31(2)
	T		0.765(5)	0.774(4)	0.7505(4)	0.7494(5)	11.3(2)	5.40(8)	0.725(1)	0.731(1)	7(2)	6(2)	0.37(4)	0.39(3)	1.4(1)	1.0(1)	0.29(5)	0.31(2)
	RT		0.758(3)	0.761(4)	0.7498(4)	0.7484(4)	11.9(2)	5.4(1)	0.746(5)	0.7302(9)	7(3)	5(1)	0.66(2)	0.39(3)	1.4(1)	1.0(1)	0.55(3)	0.31(3)
$\mathcal{F}_N$	U	3.80(1)	1.068(6)	1.048(4)	1.137(4)	1.122(5)	0.486(5)	0.573(3)	0.985(4)	1.005(4)	0.169(7)	0.138(7)	0.26(2)	0.37(3)	0.109(5)	0.14(1)	0.20(1)	0.29(1)
	R		1.078(6)	1.058(4)	1.140(4)	1.125(5)	1.02(2)	0.563(5)	1.003(5)	1.017(3)	0.59(2)	0.16(1)	0.60(3)	0.37(3)	0.45(3)	0.148(7)	0.48(3)	0.29(1)
	T		1.102(7)	1.086(6)	1.137(6)	1.124(5)	3.9(1)	2.20(4)	1.006(5)	1.030(6)	0.8(2)	0.66(7)	0.27(1)	0.39(2)	0.60(2)	0.55(2)	0.21(1)	0.30(1)
	RT		1.110(7)	1.096(7)	1.139(5)	1.127(4)	3.9(1)	2.20(4)	1.009(5)	1.036(4)	0.9(1)	0.66(6)	0.57(2)	0.39(2)	0.61(2)	0.55(2)	0.47(2)	0.29(1)