TECHNICAL NOTE • OPEN ACCESS

Dark solitons in Bose–Einstein condensates: a dataset for many-body physics research

View the article online for updates and enhancements.

MACHINE
LEARNING
Science and Technology

CrossMark

Dark solitons in Bose–Einstein condensates: a dataset for many-body physics research

Amilson R Fritsch[1] , Shangjie Guo[1] , Sophia M Koh[2] , I B Spielman[1,3] and Justyna P Zwolak[3,*]

[1] Joint Quantum Institute, National Institute of Standards and Technology, and University of Maryland, Gaithersburg, MD 20899, United States of America
[2] Department of Physics and Astronomy, Amherst College, Amherst, MA 01002, United States of America
[3] National Institute of Standards and Technology, Gaithersburg, MD 20899, United States of America
* Author to whom any correspondence should be addressed.

E-mail: jpzwolak@nist.gov

Abstract

We establish a dataset of over 1.6×10^4 experimental images of Bose–Einstein condensates containing solitonic excitations to enable machine learning (ML) for many-body physics research. About 33% of this dataset has manually assigned and carefully curated labels. The remainder is automatically labeled using SolDet—an implementation of a physics-informed ML data analysis framework—consisting of a convolutional-neural-network-based classifier and object detector as well as a statistically motivated physics-informed classifier and a quality metric. This technical note constitutes the definitive reference of the dataset, providing an opportunity for the data science community to develop more sophisticated analysis tools, to further understand nonlinear many-body physics, and even advance cold atom experiments.

1. Introduction

Advances in machine learning (ML), and especially in the area of deep learning, are data driven. Yet, in many fields of science it is a common practice for researchers to either make data available only upon 'reasonable request,' or to not share it at all. As a result, the development of specialized ML techniques as well as trained models is often stymied by the lack of high-quality relevant datasets. Cold atom experiments produce vast amounts of data—images of atom clouds—making cold atoms an ideal system where data availability can both advance ML research and provide new applications relevant to experiment [1–4]. Here we provide a two component dataset consisting of absorption images of dark solitons in atomic Bose–Einstein condensates (BECs). The first component of this dataset is a curated revision of the *Dark solitons in BECs dataset v.1.0* [5, 6] containing over 6×10^3 images, which we carefully amend to assure high quality labels. The second component contains approximately 1×10^4 additional preprocessed and automatically labeled images.

BECs are widely investigated systems that exhibit quantum phenomena on a macroscopic scale. For example, they can be manipulated to contain solitonic excitations including conventional solitons, vortices, and many more. Broadly speaking, solitonic excitations are solitary waves that retain their size and shape and often propagate at constant speed. They are present in many systems, at scales ranging from microscopic [7, 8], to terrestrial [9–15] and even astronomical [16].

Unlike naturally occurring physical systems, the parameters governing BECs are under strict experimental control. In atomic BECs, solitons can be classified as dark or bright, generally occurring for systems with repulsive or attractive interactions, respectively. Here, we focus on repulsively interacting BECs that therefore support dark solitons. In one-dimensional (1D) BECs, only a single type of dark soliton exists, corresponding to the kink soliton in three-dimensional (3D) systems. Even in highly elongated 3D BECs, kink solitons are stable only when strict conditions on the propagation velocity and the trap geometry are satisfied. Otherwise they decay, ultimately producing solitonic vortices, stable solitonic excitations in 3D BECs.

Kink solitons manifest as a plane-like reduction in the BEC's 3D density, while solitonic vortices add to this reduction a vortex-line of zero density about which the BEC flows (with either sense of vorticity). Visually, kink solitons and solitonic vortices both appear as depletions in the BEC's atomic density, with solitonic vortices having additional structure that is not present for kink solitons. Given this, one might expect such images of BECs to be straightforward to classify. However, due to physical limitations of the imaging system capturing only a single-sided view of the 3D BEC, they are not. For example, the plane of reduced density can be canted with respect to the BEC, both for kink solitons and solitonic vortices (leading to our label 'canted excitation'). The solitonic vortex [17, 18] can be predominantly confined to the top or the bottom of the BEC (leading to what we label as 'top partial' and 'bottom partial' excitations, respectively); in addition, vorticity leaves its imprint on density as well (giving our labels 'clockwise vortex' and 'counterclockwise vortex'). Finally, when viewed from the side, the density-zero of a solitonic vortex becomes indistinguishable from a shallow kink soliton [18] (leading to our joint label of 'longitudinal soliton').

The expanded dataset, *Dark solitons in BECs dataset v.2.0*, contains experimental images of dark solitons created using the procedure described in reference [19]. The constituent images reflect a diverse range of parameters, for example containing from zero to upwards of four excitations of numerous types that themselves have a wide range of velocities. Due to the similarities between kink solitons and other excitations present in our images, annotators can potentially mislabel data, attributing different labels to the same type of excitation.

The full dataset contains over 1.6×10^4 experimental images, including 5378 images manually classified into three carefully curated classes: 'no excitations' (class-0), 'lone excitation' (class-1), and 'other excitations' (class-2). The lone excitation class is additionally tagged with the excitation position, the physics-informed excitation (PIE) class, and quality score. The remaining images are not manually classified; rather they are automatically labeled using the SolDet package [20]. This dataset is available via the National Institute of Standards and Technology Public Data Repository [5] to provide an opportunity for the data science community to develop more sophisticated analysis tools for soliton research and to further understand nonlinear many-body physics.

2. Dataset curating: materials and methods

In 2021, we released the 'Dark solitons in BECs dataset' [5] consisting of approximately 6.3×10^3 preprocessed absorption images taken from multiple experiments performed in a single lab over a span of two months with human assigned labels. Based upon the number of solitonic excitations observed in a given image of a BEC, the data was organized into three classes: no excitations (class-0, accounting for 19.8% of the dataset), lone excitation (class-1, accounting for 55.4%), and other excitations (class-2, accounting for 24.8%). While the initial agreement rate between annotators was relatively high at 87%, the remaining 13% of the dataset had to be 'further analyzed and discussed until an agreement' between annotators was reached[4], as stated in reference [6]. Such discussions of labels might introduce an undesirable bias in the labels, especially when the data is challenging to interpret. This bias is in turn imprinted into any ML model trained using that data, thereby putting the model's reliability into question.

While manually re-examining the dataset in the context of reference [4], we confirmed the presence of inconsistencies in the human assigned labels. We found three types of labeling errors:
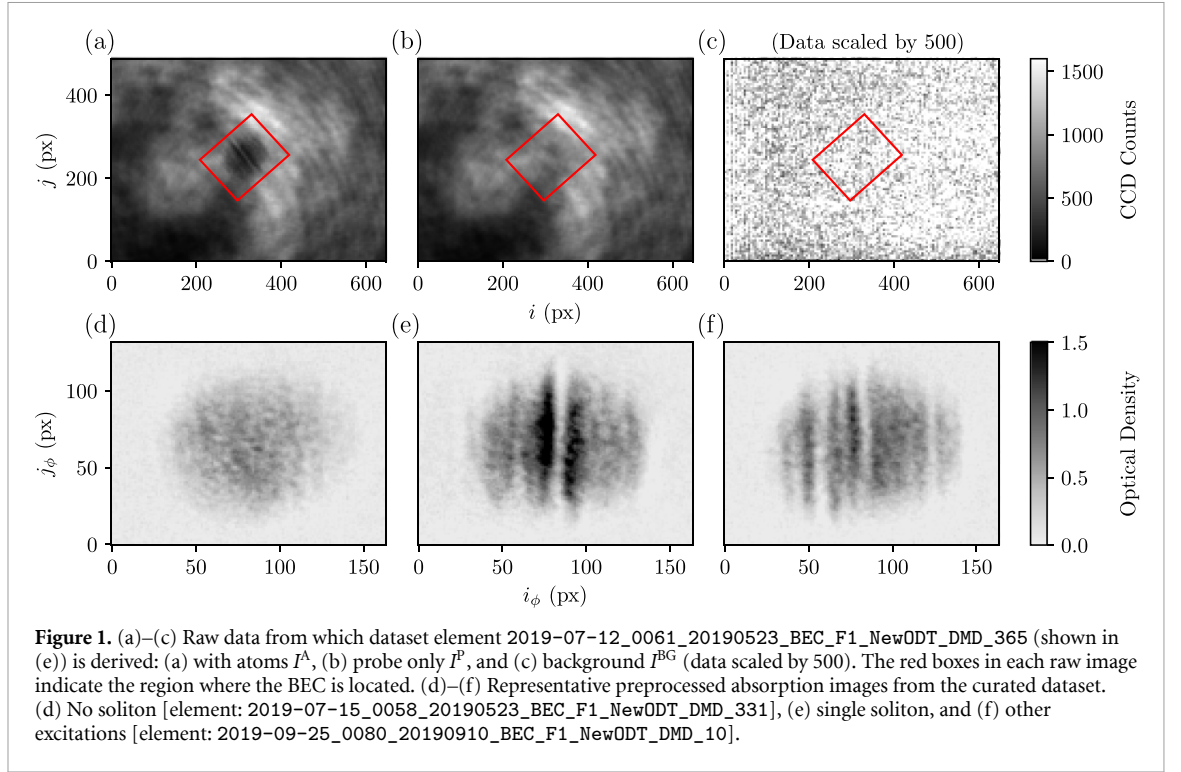
(e1) Images in class-2 containing only a single excitation.
(e2) Images in class-1 containing more than one solitonic excitation.
(e3) Images in class-0 containing a distinct excitation.

Moreover, in reference [6], the excitation's location for the lone excitation class was determined using fits centered on the deepest density depletion. Our reexamination showed that some solitonic excitations were far from this point (error type (e4)). As a result, we decided to use a combination of ML and statistical analyses to identify potentially incorrectly labeled data as well as to curate the dataset.

Building on the deep ensembles approach, originally proposed as a means to estimate models' predictive uncertainty [21], we implement an iterative five-fold cross-validations[5] with strict agreement constraints to curate the original dataset. As described in detail below, we employ two classifications schemes for the

---

[4] The agreement was substantially higher for the easier to interpret class-0 (95.7%) than for other two cases (88.7% for class-1 and 76.3% for class-2), indicating a likely decrease in label reliability with increased data complexity.

[5] The k-fold cross-validation is a resampling method that involves dividing the full dataset into k partitions and then performing a series of training and testing runs, with each run using $k-1$ partitions to train a given module and the remaining one partition to test it. The process is repeated k times to fully cover the dataset.

Figure 1. (a)–(c) Raw data from which dataset element 2019-07-12_0061_20190523_BEC_F1_NewODT_DMD_365 (shown in (e)) is derived: (a) with atoms I^A, (b) probe only I^P, and (c) background I^{BG} (data scaled by 500). The red boxes in each raw image indicate the region where the BEC is located. (d)–(f) Representative preprocessed absorption images from the curated dataset. (d) No soliton [element: 2019-07-15_0058_20190523_BEC_F1_NewODT_DMD_331], (e) single soliton, and (f) other excitations [element: 2019-09-25_0080_20190910_BEC_F1_NewODT_DMD_10].

cross-validation. First, we verify the human assigned labels using a set of convolutional neural network (CNN) classifiers trained using the original dataset. In addition, to ensure diversity of the trained models [22], we use a set of object detectors (ODs) [23, 24] trained to localize all solitonic excitations within each BEC image and compare the number of detected excitations with both the original and CNN labels. After each iteration, images with insufficient agreement are further analyzed and, if necessary, removed from the dataset and set aside. Following the data curation process, we add the location of all excitations obtained from the ODs as an additional label.

The curated dataset is used to train an implementation of SolDet [20], a general-purpose framework for feature identification in cold atom experiments [4]. We then use the PIE classifier SolDet module to add fine-grained and physically-motivated labels (e.g. longitudinal solitons, solitonic vortices, and 'partial' solitons) for the lone excitation class. Furthermore, we employ SolDet to automatically label about 1×10^4 additional experimental images (class-9) as well as all images set aside during the curation process (class-8) [5].

2.1. Data preprocessing

In the raw data, shown in figures 1(a)–(c), the BEC occupies only a small region of the image (inside the red box) and the long axis of the BEC is rotated with respect to the camera. The horizontal and vertical axes in figures 1(a)–(c) are labeled in terms of camera pixels i and j.

The angle between the camera and the BEC depends on the experimental setup and is obtained from fits to a representative subset of absorption images (in the case of our data the rotation angle is about $\phi = 40$ degrees). The BEC's angle, position, and size—all necessary for proper cropping—are determined by fitting every image to a column-integrated 3D Thomas-Fermi distribution

$$n_{i,j}^{\mathrm{TF}} = n_0 \max\left\{ \left[1 - \left(\frac{i_\phi}{R_i} \right)^2 - \left(\frac{j_\phi}{R_j} \right)^2 \right], 0 \right\}^{3/2} + \delta n, \tag{1}$$

describing the density distribution of 3D BECs integrated along the imaging axis [25]. There are seven parameters in this fit: the rotation angle ϕ; the BEC center coordinates $[i_0, j_0]$ in the original image frame; the peak 2D density n_0; the Thomas–Fermi radii $[R_i, R_j]$; and an offset δn (from small changes in probe intensity between images). We define the rotated coordinates as $i_\phi = (i - i_0)\cos(\phi) + (j - j_0)\sin(\phi)$ and $j_\phi = (j - j_0)\cos(\phi) - (i - i_0)\sin(\phi)$. The initial estimates for $[i_0, j_0]$ are obtained by summing the 2D density image along the vertical and horizontal directions to obtain two 1D projections and selecting the average position of the five largest values. The largest value of the image is taken as the estimate for n_0 and $[R_i, R_j] = [66, 55]$ as the estimate for the radii. The estimate for the offset δn is zero. The 164×132 pixel

Table 1. Labeling statistics. The first two rows describe the labeling distribution in *Dark solitons in BECs dataset v.1.0* and include the number of images with full agreement in each class. The third row counts the number of images in each class that pass the first round of CNN and OD tests. The last row shows the number of images remaining in each class in the curated dataset.

Labeling phase	Class-0	Class-1	Class-2	Class-8	Class-9
Original dataset	1237	3468	1552	—	—
Original dataset: 3-agree	1184	3077	1184	—	—
CNN and OD test	1234	3449	1388	186	—
Curated dataset	1130	3212	1036	879	10 221

extent of the cropping region is determined from the average radii $[R_i, R_j] = [66(5), 58(3)]$ obtained from fits to the original 6.2×10^3 images[6]. In the resulting images, dark solitons appear as vertically aligned density depletions and are often easily visually identified, see figures 1(e) and (f).

To ease the manual analysis and labeling process and facilitate training of the ML models, the absorption images are first rotated to align the BEC with the image frame and then cropped to discard the large fraction of the image that does not contain relevant information. Finally, an elliptical mask (determined based on the $[R_i, R_j]$ radii) is applied to the image to eliminate the noise outside the BEC [6].

The same preprocessing techniques are applied to the over 1×10^4 previously unlabeled absorption images now included in the expanded 'Dark solitons in BECs dataset 2.0'.

2.2. Data labeling process: dark solitons in BECs dataset v.1.0

As discussed in reference [6], the *Dark solitons in BECs dataset v.1.0* [5] consists of images labeled by three independent annotators. These labels organized the data into three disjoint classes: class-0 indicating no excitations, class-1 indicating lone excitation, and class-2 indicating other excitations (e.g. different types of excitations, multiple excitations, ambiguous data). The expectation was for class-0 to consist of images that unambiguously contain no solitonic excitations; for the class-1 to contain images with exactly one solitonic excitation; and for class-2 to contain all other images.

The initial labeling process was carried out in batches. At each stage, a subset of anywhere between 508 and 1209 images were independently labeled by each annotator and the resulting labels were compared. The labels with full agreement were accepted. When only two out of three annotators agreed (moderate disagreement), the images were reinspected and further discussed until an agreement was reached. Finally, images labeled differently by each annotator (strong disagreement), were added to class-2. The top two rows in table 1 show the distribution of images between classes in *Dark solitons in BECs dataset v.1.0*, as well as the number of images with full agreement between annotators in each class.

2.3. Data labeling process: dark solitons in BECs dataset v.2.0

During the first phase of data curating, we performed a pair of five-fold cross validation tests using the original dataset. The first cross validation used CNN models trained on all three classes and the second used ODs trained with only class-0 and class-1. After cross validation we tagged each excitation located by the OD with a quality estimate [4]. The quality estimator yields the likelihood that a fit to the 1D profile of a given excitation has parameters in the range expected for a solitonic excitation. The likelihood was established based on a statistical analysis of fits to features previously identified as solitonic excitations in comparison with all other density depletions.

The cross-validations results in each image in class-0 and class-1 being assigned two predicted labels that are used to identify ambiguous data. To enable direct comparison of the two models, the OD predicted label is class-0 if no excitations are identified, class-1 if one excitation is found, and class-2 in all other cases. An image is flagged as potentially mislabeled if and only if

(a)   the CNN prediction disagrees with the assigned class and
(b)   the CNN and OD predictions are the same.

We found 343 potentially mislabeled images: 14 in class-0, 40 in class-1, and 289 in class-2. We note that our intent is to use cross-validation and deep ensembles to assist data curating, not to change the ground truth. Thus, we do not overwrite the original class labels during the curating process. Rather, all flagged images are further analyzed and all potential excitations are tested with the quality metric. At this stage, all

---

[6] We use a notation value(uncertainty) to express uncertainties, for example 1.5(6) cm would be interpreted as (1.5 ± 0.6) cm. All uncertainties herein reflect the uncorrelated combination of single-sigma statistical and systematic uncertainties.

Table 2. Complete dictionary of the labels appearing in *Dark solitons in BECs dataset v.2.0*. For each element we provide the key, definition together with all possible instances of a given label, and data type. The `excitation_position`, `excitation_PIE`, and `excitation_quality` labels are assigned only to data where `label_v3 = 1`. The SolDet labeles (`soldet_CNN`, `soldet_OD`, `soldet_PIE`, and `soldet_QE`) are assigned to all of the class-8 and class-9 data as well as to the 10% of the curated manually labeled data used for testing during SolDet training.

Dictionary key	Definition	Data type
`file_name`	Information about which data file a given set of labels refers to	String
`label_v1`	the original, human assigned label as in *Dark solitons in BECs dataset v.1.0*	Integer
	0: no excitations	
	1: lone excitation	
	2: other excitations	
	9: unlabeled	
`3-agree`	Indicates whether all annotators agreed on the originally assigned label	Integer
	1: true	
	0: false	
	−1: unlabeled data	
`label_v2`	Intermediate label resulting from the first curation phase	Integer
	0,1,2,9: same as for `label_v1`	
	8: data determined to be potentially mislabeled	
`label_v3`	Curated label	Integer
	0,1,2,8,9: same as for `label_v2`	
`excitation_position`	True position of the excitation for the lone excitation class	List
`excitation_PIE`	Physically-motivated label provided by the PIE classifier	List
	A: longitudinal soliton (0)	
	B: top partial (1)	
	C: bottom partial soliton (2)	
	D: clockwise solitonic vortex (3)	
	E: counterclockwise solitonic vortex (4)	
	F: canted (5)	
`excitation_quality`	Quality metric for excitations in the lone excitation class	List
`soldet_CNN`	CNN classifier label from SolDet	Integer
`soldet_OD`	List of positions returned by SolDet	List
`soldet_PIE`	List of classes returned by PIE classifier for all excitation localized in `soldet_OD`	List
`soldet_QE`	List of quality estimates for all excitation localized in `soldet_OD`	List

images in the extended *Dark solitons in BECs dataset v.2.0* have `label_v1` (either the original label or class-9 for new, therefore unlabeled, data), and in addition they are assigned a new intermediate label `label_v2` (see table 2). `label_v2` is set equal to `label_v1` except for data determined to be truly mislabeled where `label_v2` is assigned to a new class-8 effectively removing it from the curated dataset. The resulting distributions between classes is shown in table 1.

In the next stage of data curating, we further refine labels for the data that are not in class-8 or class-9 in `label_v2` using five distinct deep ensembles of size ten trained through a repeated five-fold cross-validation. Prior research suggest that ensembles of ten models are sufficient to reliably assess the predictive uncertainty [21]. Building on that, we use ten five-fold-cross-validated OD models[7] trained using `label_v2` class-0 and class-1 (4683 images in total).

Each image is tagged with ten OD predicted labels, each consisting of the number of excitations detected and their positions. Given the random initialization of the training sessions, we treat the deep ensemble as giving a uniformly-weighted set of predicted labels, with each model prediction considered equally reliable [21]. The OD class predictions are used to define a measure of class-based disagreement

$$D_{\text{class}} = \frac{\#(\text{OD class prediction} \neq \texttt{label\_v2})}{M}, \tag{2}$$

where $\#(\cdot)$ denotes the number of instances for which condition $(\cdot)$ occurs and M is the size of the deep ensemble. A score of 0 indicates full agreement within the deep ensemble with the ground truth label while score of 1 indicates that all models predict an incorrect class. In addition, each image in class-1 is assigned a preliminary excitation position (FIT_{pos}): the minimum of the background subtracted 1D density profile [6, 19].

---

[7] There were 50 models in total—ten per cross-validation—each trained on 80% of the data. To ensure diversity of models, the dataset was shuffled between the ten training sessions.

The subsequent data analysis continues our aim of identifying mislabeled images and is carried out separately for class-0, 1, and 2. The result of this analysis is stored in `label_v3`.

Class-0: *Deep ensemble disagreement.* For each image, we calculate the class-based disagreement score D_{class}. Since the initial agreement rate between annotators was around 87%, we opt to require at least 90% agreement between the ODs. Thus, images with $D_{\text{class-0}} \leqslant 0.1$ are retained in the curated dataset with `label_v3` set equal to `label_v2` (1130 images). We note that the remaining 104 images had either (a) $D_{\text{class}} \in (0.1, 0.5]$ (75 images); or (b) half or more models predicting class-1 (29 images). All of these are assigned to class-8.

Class-1: *Step 1: Deep ensemble disagreement.* Like for class-0, we first compare the number of OD-identified excitations with `label_v2` and find 3222 images for which $D_{\text{class}} \leqslant 0.1$. Of the remaining 227 there are 178 for which $D_{\text{class}} \in (0.1, 0.5]$, 28 images for which five or more models predict class-0, and 21 images in which five or more models assign class-2. All these images are assigned to class-8.

Step 2: Position range check. While the ODs agree on the number of solitonic excitations, the positions in pixel units found by the ODs might differ. All images for which $\max(\mathbf{OD}_{\text{pos}}) - \min(\mathbf{OD}_{\text{pos}}) > 3$ are assigned to class-8 (10 images)[8]. Here $\mathbf{OD}_{\text{pos}}$ is the vector of OD position predictions and $\max(\mathbf{v})$ and $\min(\mathbf{v})$ denote the maximum and minimum element of the of the vector $\mathbf{v}$, respectively.

Step 3: Position alignment check. As an additional consistency check, we compare the average OD position for images against the preliminary FIT_{pos} position. Images for which $|\overline{\mathbf{OD}}_{\text{pos}} - \text{FIT}_{\text{pos}}| > 3$ (15 images) were manually reviewed. We found that in each case the excitation position was better located by the OD, so the excitation position label was updated to that returned by the OD.

Class-2: *Step 1: Deep ensemble disagreement.* Data in class-2, by design, includes ambiguous images. Thus, the goal of curating this class is to ensure that it does not contain images that would with high confidence be classified as belonging to either class-0 or class-1 by the deep ensemble. Since all models were trained only on the class-0 and class-1 data, we use an ensemble consisting of all 50 models. We find that out of the 1388 images in class-2, 30 had 90% of models predict class-0 (and were therefore assigned to class-8) and 336 had 90% of models predict class-1 (and were further analyzed in step 2). The remaining 1022 images are retained in the curated dataset with `label_v3` set equal to `label_v2`.

Step 2: Position range check. To confirm the OD predicted class-1, we compare the range of OD positions for those models predicting one excitation and find that $\max(\mathbf{OD}_{\text{pos}}) - \min(\mathbf{OD}_{\text{pos}}) < 3$ for 322 of images, suggesting that these are very likely class-1 data mislabeled as class-2. These images are also assigned to class-8.

The deep-ensembles-based data curating process resulted in assigning to class-8 a total of 693 images from the original dataset. The resulting dataset contains 1130 images in class-0, 3212 in class-1, 1036 in class-2, and 879 in class-8 (data labeled as class-9, associated with unlabeled data, are unchanged). The final classifications are contained in `label_v3`.
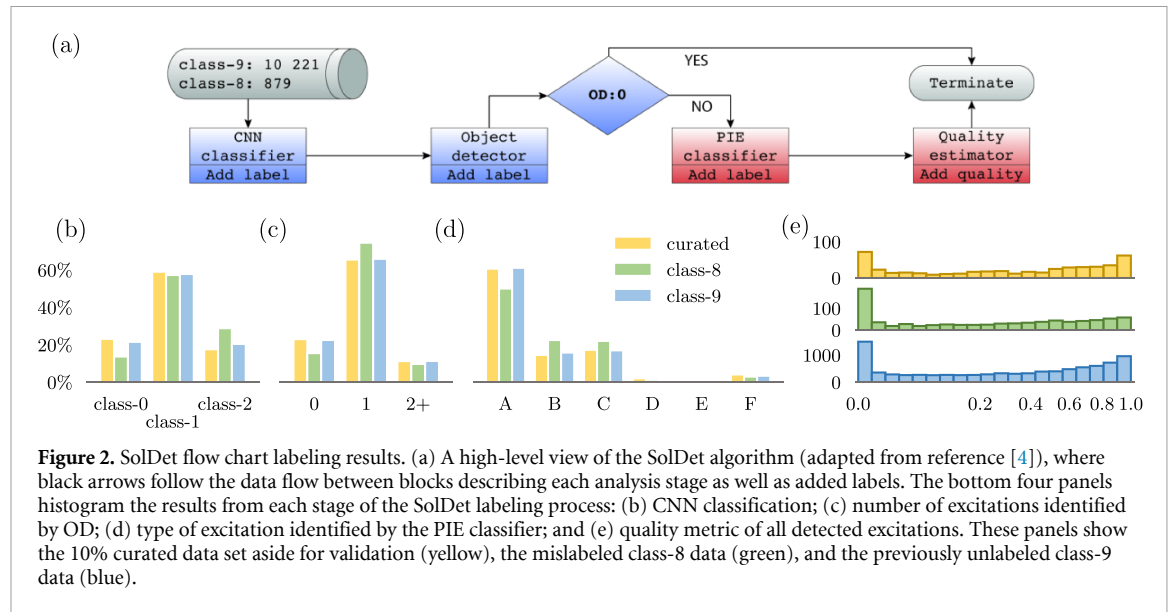
2.4. Label refinement: dark solitons in BECs dataset v.2.0

To further refine labels for images in class-1, we use the PIE classifier and quality estimator from the SolDet package [20]. PIE classifier partitions class-1 into physically-motivated sub-classes stored in `excitation_PIE`. The PIE classifier operates by splitting each image into top and bottom halves and determining the associated 1D profiles to which the quality estimator is separately applied. In addition to returning an overall quality estimate, stored in `excitation_quality`, this algorithm also returns parameters such as the excitation position, width, and so forth.

Then, a sequence of thresholds driven by different top-bottom combinations of these parameters is applied to determine the label. The values defining all thresholds were arrived at by exploring the data accepted and rejected by the cut to minimize the false positive identification of longitudinal solitons, as described in reference [4].

Within the 3212 images in class-1, the PIE classifier categorized 2229 images as proper longitudinal solitons (class-A). Out of the remaining images, 378 were classified as top 'partial' solitons (class-B) and 418 as bottom 'partial' solitons (class-C); 28 were categorized as clockwise vortex (class-D) and 38 as counterclockwise vortex (class-E); 121 were categorized as canted excitations (class-F).

---

[8] This threshold corresponds to the average width of solitons in our images which we observed to be about four pixels.

Figure 2. SolDet flow chart labeling results. (a) A high-level view of the SolDet algorithm (adapted from reference [4]), where black arrows follow the data flow between blocks describing each analysis stage as well as added labels. The bottom four panels histogram the results from each stage of the SolDet labeling process: (b) CNN classification; (c) number of excitations identified by OD; (d) type of excitation identified by the PIE classifier; and (e) quality metric of all detected excitations. These panels show the 10% curated data set aside for validation (yellow), the mislabeled class-8 data (green), and the previously unlabeled class-9 data (blue).

3. Automated expansion of the dataset v.1.0

In this section, we describe how we leverage the full SolDet package [20] to automatically analyze and label previously unlabeled images (a subset of about 1×10^4 identified in `label_v3` as class-9) that were not part of *Dark solitons in BECs dataset v.1.0*. These data include images representing class-0, 1, and 2; many class-2 candidates possess multiple excitations as well as ambiguous and confusing structures that may hinder human labeling. Since these data were not previously considered, they make an ideal test case for the SolDet package [4]. Roughly 90% of the class-0, 1, and 2 data from the curated dataset were used to train SolDet, leaving the remaining 10% of these classes for validation. In addition, we also apply SolDet to all of class-8 allowing us to cross-check the mislabeled assignment.
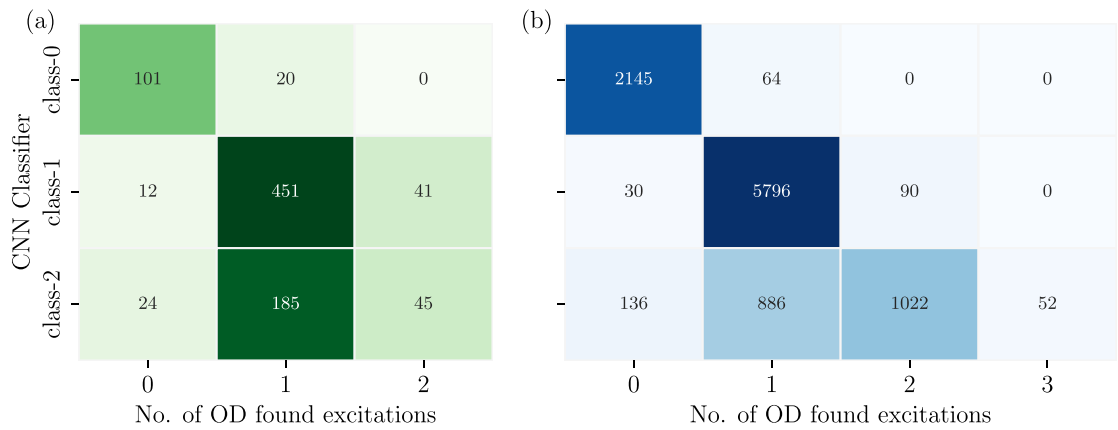
For this application, the SolDet package is configured to give the labeling flow depicted in figure 2(a), involving a sequential application of CNN and OD modules followed by the PIE classifier. The CNN module categorizes images as class-0, 1, or 2 while the OD module tags each image with positions of all detected excitations. If OD assigns class-0, the image is labeled accordingly and the process terminates for that image. Otherwise, the PIE classifier is executed for each excitation detected by the OD. Finally, all excitations located by OD are additionally tagged with a quality estimate [4].

The output of each module is included as a separate label: `soldet_CNN` for the CNN classifier; `soldet_OD` for the vector of positions; `soldet_PIE` for the vector of classes returned by PIE classifier; and `soldet_QE` for the vector of quality estimates. This enables the end user to choose a desired level of agreement between the labeling modules or the longitudinal solitons' quality necessary for a particular application. Thus, unlike images in the curated dataset, these previously unlabeled images are not assigned a single ground truth class.
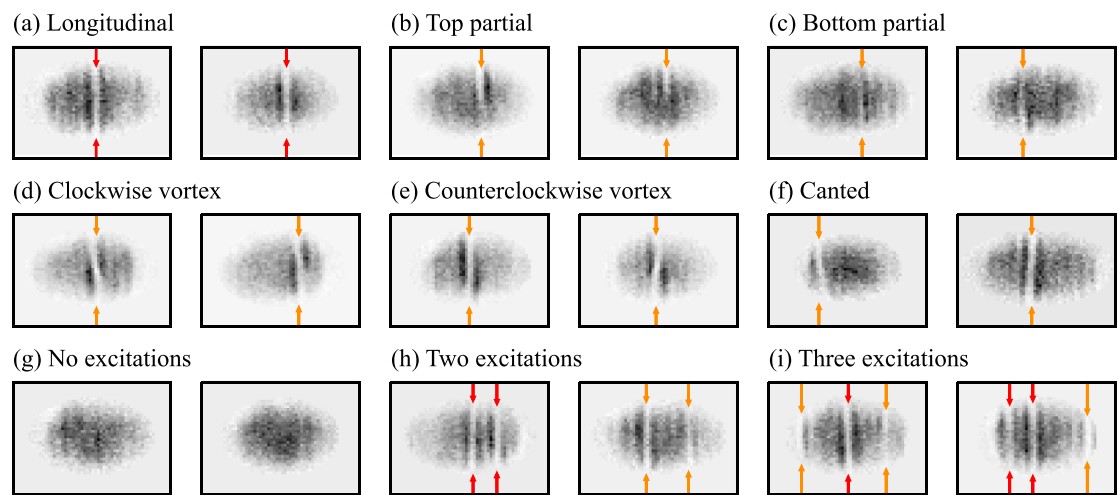
Figures 2(b)–(d) compares the label assignment for three different subsets of our dataset: a subset of the curated data (10%) automatically generated by SolDet for validation (shown in yellow), class-8 (mislabeled data; shown in green), and class-9 (unlabeled data; shown in blue). Panel (b) shows that CNN classes for each case follow a very similar distribution, but where data in class-8 has a reduced likelihood of being classified in class-0. While the OD results depicted in panel (c) are very similar to the data in (b), there are slight differences. For example in the class-8 data, the OD finds somewhat more class-1 excitations than the the CNN did (by about 17.3%). This likely results from the annotators disagreeing on the labeling of partial solitons and solitonic vortices, all of which are identified as a single excitation by the OD. The PIE classifier in (d) finds a modest deficit of class-A (longitudinal soliton) in the mislabeled data, confirming that such excitations are the most straightforward to annotate. Lastly, figure 2(e) shows that the quality metric for all excitations identified by the OD are consistent across the three subsets of our data.

Figure 3 compares the outcome of the CNN and OD modules for (a) class-8 data and (b) class-9 data. In both cases when the CNN classifier assigns class-0 or class-1, the OD is unlikely to find any or other number than one excitation, respectively. However, when the CNN assigns class-2, the OD assignment is strongly biased to class-1 for the class-8 (mislabeled) data, with 73% assessed to contain only one excitation. This bias likely results from the process of labeling and curation in which many class-1 candidates were moved into

Figure 3. Matrices comparing the outputs of the CNN module and the OD module for (a) the mislabeled class-8 data as well as (b) the previously unlabeled class-9 data.

Figure 4. Representative class-9 data from the *Dark solitons in BECs dataset v.2.0*. (a) Longitudinal soliton; (b) top partial; (c) bottom partial; (d) clockwise solitonic vortex; (e) counterclockwise solitonic vortex; (f) canted excitation; (g) no excitation; (h) two excitations; and (i) three or more excitations. Red arrows indicate longitudinal solitons while orange ones are for other solitonic excitations.

class-8 to avoid false positives. The OD assignment seems almost random for the previously unlabeled class-9 data assigned class-2 by the CNN, with 42% assessed to contain only one excitation and 49% assessed to contain two excitations.

For the mislabeled (class-8) data the performance is significantly degraded in the converse case: about 30% of the OD class-0 and 1 data and almost 50% of OD class-2 data is assigned one of the alternative CNN classes. For the unlabeled (class-9) data, the disagreement between OD and CNN assigned classes is much lower, at 7%, 14%, and 8% for OD class-0, 1, and 2+, respectively.

Figure 4 depicts 18 images, with a variety of classifications and positions, all from the automatically labeled (class-9) portion of the expanded dataset. By design, SolDet effectively categorizes class-1-A images, giving the examples in (a). Panels (b)–(f) show illustrative examples from the remaining classes. Because the primary function of the PIE classifier is to reject images that are not class-1-A (avoid false positive longitudinal solitons) the B–F labels are of lower quality. The bottom row displays data from class-0 (no excitations; panel (g)) and class-2 (other excitations; panels (h)–(i)). Together, these show that SolDet is very effective in delineating between class-0,1 and 2.

In each panel the arrows identify the location of the excitation from the OD, showing it is effective across-the-board in locating excitations. The arrows are colored according to the PIE classifier result: Red arrows mark the location of longitudinal solitons (class-*-A) and the orange arrows mark all other classes. Even in cases with many excitations (h), (i), SolDet correctly identifies high quality longitudinal solitons.

4. Conclusion and outlook

We find that SolDet reliably labels new experimental data (class-9) as well as data categorized as potentially mislabeled (class-8). An inspection of the images shows that the the assigned classes are visually similar to the manually labeled data. Furthermore, the labels automatically assigned to class-9 have a statistical distribution that is very similar to the training dataset. By contrast, the labels assigned to class-8—a selectively filtered subset of the original dataset—are significantly different.

As we mentioned, SolDet is configured to identify and correctly locate longitudinal soliton within BEC (class-*-A). The reliability of the additional PIE classes could be improved by, e.g. further refining cuts defining the physically-motivated categories or slicing the image into more than two pieces.

The enlarged *Dark solitons in BECs dataset v.2.0* dataset includes quantitative estimates of all longitudinal solitons quality as well as new fine-grained solitonic excitation categories of all detected excitations. It is a freely available to the whole ML and physics community the opportunity to develop novel ML techniques to cold atom systems and to further explore the intersection of ML and quantum physics.

Data availability statement

The data that support the findings of this study are openly available at the following URL/DOI: https://doi.org/10.18434/mds2-2363.

Acknowledgments

Appendix

Experimental setup

In our experiments, we implement well established techniques to achieve the Bose–Einstein condensation of neutral atoms [26]. Our ^{87}Rb BECs contain $N \approx 2.4 \times 10^5$ atoms in an elongated crossed optical dipole trap with harmonic frequencies $[\omega_x, \omega_y, \omega_z] = 2\pi \times [9.1(1), 94.5(6), 153(1)]$ Hz.

To create solitons, we manipulate the BEC local density and phase by applying an external potential generated by a far detunned laser light patterned by a digital micromirror device (DMD). The DMD (Texas Instruments DLP LightCrafter Module—DLP3000) [9] has an array of 608×684 mirrors, $\approx 7.6\,\mu$m on a side, that can be independently flipped to create arbitrary patterns. The pattern generated by the DMD is imaged onto the atoms using an imaging system that demagnifies the light 12 times and the intensity of the laser light is controlled by an acousto-optic modulator.

The protocol to create solitons, which starts after the BEC is formed, is summarized as follows: with the DMD previously programmed to reflect a narrow 3 px wide stripe, we increase the laser power creating a dimple potential that depletes the BEC local density by about 70%. We then change the DMD pattern to illuminate half of the BEC extension and pulse the light for a variable time to imprint the phase. Since the accumulated phase is proportional to the duration in which the light is pulsed, the pulse duration is varied to create solitons at different speeds. To avoid creating additional density modulations after imprinting the phase, the DMD is reconfigured back to the narrow stripe, the dimple is reapplied and its magnitude is ramped to zero. More details about this protocol can be found in [19].

After solitons are created we let them oscillate in the trap for a variable evolution time that allow us to obtain variation in the soliton properties, such as the oscillation amplitude, initial position, propagation velocity, and lifetime. Since our elongated trap geometry does not produce truly 1D BECs, kink solitons that are initially created can eventually decay into solitonic vortices during the time they oscillate in the trap. After

---

[9] Certain commercial equipment, instruments, or materials are identified in this paper in order to specify the experimental procedure adequately. Such identification is not intended to imply recommendation or endorsement by the National Institute of Standards and Technology, nor is it intended to imply that the materials or equipment identified are necessarily the best available for the purpose.

the evolution time, the trap is turned off and the cloud expands for 12 ms before it is imaged using standard absorption imaging [26].

In the standard absorption imaging technique, the fraction of a resonant probe light absorbed by the atomic cloud is used to extract information about the atoms. For the standard absorption imaging, we acquire three images that are combined to obtain the optical density. In the first image $I_{i,j}^{A}$, the atomic cloud is illuminated with the probe light and its shadow is recorded in a CCD camera, see figure 1(a). To compute the fraction of the light that is absorbed by the atoms, a second image $I_{i,j}^{P}$ of the probe beam, without atoms, is acquired, figure 1(b). The third image $I_{i,j}^{BG}$ is then acquired without the probe light to get the background light in the experiment, figure 1(c). All images are acquired with the same duration and the probe beam has the same intensity for the first two images. The three images are then combined to to obtain the 2D optical density

$$\sigma_0 \, n_{i,j} \approx -\ln\left[\frac{I_{i,j}^{A} - I_{i,j}^{BG}}{I_{i,j}^{P} - I_{i,j}^{BG}}\right], \tag{3}$$

where $\sigma_0 = 3\lambda^2/(2\pi)$ is the resonant cross-section and λ is wavelength of the probe laser.

All images (figures 1(a)–(c)) are acquired by a 648×488 pixel camera (Point Grey FL3) with $5.6\,\mu$m square pixels, labeled by i and j. The imaging system, with an optical resolution of $\approx 2.8\,\mu$m, has $\approx 6\times$ magnification, generating images with effective pixel size of $0.93\,\mu$m.

ORCID iDs

Amilson R Fritsch ⓘ https://orcid.org/0000-0003-1419-8591
Shangjie Guo ⓘ https://orcid.org/0000-0002-6910-636X
Sophia M Koh ⓘ https://orcid.org/0000-0001-5481-0285
I B Spielman ⓘ https://orcid.org/0000-0003-1421-8652
Justyna P Zwolak ⓘ https://orcid.org/0000-0002-2286-3208

References

[1] Ness G, Vainbaum A, Shkedrov C, Florshaim Y and Sagi Y 2020 Single-exposure absorption imaging of ultracold atoms using deep learning *Phys. Rev. Appl.* **14** 014011
[2] Seo S, Lee J H, Lee S-B, Park S-E, Seo M H, Park J, Kwon T Y and Hong H-G 2021 Maximized atom number for a grating magneto-optical trap via machine-learning assisted parameter optimization *Opt. Express* **29** 35623–39
[3] Leykam D, Rondon I and Angelakis D G 2021 Dark soliton detection using persistent homology (arXiv:2107.14594)
[4] Guo S, Koh S M, Fritsch A R, Spielman I B and Zwolak J P 2022 Combining machine learning with physics: a framework for tracking and sorting multiple dark solitons *Phys. Rev. Res.* **4** 023163
[5] Database: data.nist.gov 2020 Dark solitons in BECs dataset (available at: https://doi.org/10.18434/mds2-2363)
[6] Guo S, Fritsch A R, Greenberg C, Spielman I B and Zwolak J P 2021 Machine-learning enhanced dark soliton detection in Bose-Einstein condensates *Mach. Learn.: Sci. Technol.* **2** 035020
[7] Burger S, Bongs K, Dettmer S, Ertmer W, Sengstock K, Sanpera A, Shlyapnikov G V and Lewenstein M 1999 Dark solitons in Bose–Einstein condensates *Phys. Rev. Lett.* **83** 5198–201
[8] Denschlag J *et al* 2000 Generating solitons by phase engineering of a Bose–Einstein condensate *Science* **287** 97–101
[9] Russel J S 1837 Report of the committee on waves *Report of the 7th Meeting of the British Association for the Advancement of Science, Liverpool* (London: John Murray) pp 417–68
[10] Hasegawa A and Tappert F 1973 Transmission of stationary nonlinear optical pulses in dispersive dielectric fibers. II. Normal dispersion *Appl. Phys. Lett.* **23** 171–2
[11] Osborne A R and Burch T L 1980 Internal solitons in the Andaman Sea *Science* **208** 451–60
[12] Mollenauer L F, Stolen R H and Gordon J P 1980 Experimental observation of picosecond pulse narrowing and solitons in optical fibers *Phys. Rev. Lett.* **45** 1095–8
[13] Yomosa S 1987 Solitary waves in large blood vessels *J. Phys. Soc. Jpn.* **56** 506–20
[14] Hashizume Y 1985 Nonlinear pressure waves in a fluid-filled elastic tube *J. Phys. Soc. Jpn.* **54** 3305–12
[15] Lakshmanan M 2012 Solitons, tsunamis and oceanographical applications of *Mathematics of Complexity and Dynamical Systems* ed R A Meyers (New York: Springer) pp 1603–17
[16] Stasiewicz K, Shukla P K, Gustafsson G, Buchert S, Lavraud B, Thidé B and Klos Z 2003 Slow magnetosonic solitons detected by the cluster spacecraft *Phys. Rev. Lett.* **90** 085002
[17] Tylutki M, Donadello S, Serafini S, Pitaevskii L P, Dalfovo F, Lamporesi G and Ferrari G 2015 Solitonic vortices in Bose–Einstein condensates *Eur. Phys. J. Spec. Top.* **224** 577–83
[18] Donadello S, Serafini S, Tylutki M, Pitaevskii L P, Dalfovo F, Lamporesi G and Ferrari G 2014 Observation of solitonic vortices in Bose–Einstein condensates *Phys. Rev. Lett.* **113** 065302
[19] Fritsch A R, Mingwu L, G Reid H, Piñeiro A M and Spielman I B 2020 Creating solitons with controllable and near-zero velocity in Bose–Einstein condensates *Phys. Rev. A* **101** 053629
[20] Zwolak J P, Guo S, Koh S M, Fritsch A R and Spielman I B 2022 SOLDET: solitonic feature detection package GitHub (available at: https://doi.org/10.18434/mds2-2363)

[21] Lakshminarayanan B, Pritzel A and Blundell C 2017 Simple and scalable predictive uncertainty estimation using deep ensembles *Advances in Neural Information Processing Systems* vol 30

[22] Ovadia Y, Fertig E, Ren J, Nado Z, Sculley D, Nowozin S, Dillon J, Lakshminarayanan B and Snoek J 2019 Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift *Advances in Neural Information Processing Systems* vol 32

[23] Redmon J, Divvala S, Girshick R and Farhadi A 2016 You only look once: unified, real-time object detection *Proc. of the IEEE Computer Society Conf. on Computer Vision and Pattern Recognition* pp 779–88

[24] Voulodimos A, Doulamis N, Doulamis A and Protopapadakis E 2018 Deep learning for computer vision: a brief review *Comput. Intell. Neurosci.* **2018** 7068349

[25] Castin Y and Dum R 1996 Bose–Einstein condensates in time dependent traps *Phys. Rev. Lett.* **77** 5315–9

[26] Ketterle W, Durfee D S and Stamper-Kurn D M 1999 Making, probing and understanding Bose–Einstein condensates *Proc. of the Int. School of Physics "Enrico Fermi"* ed M Inguscio, S Stringari and C E Wieman (Amsterdam: IOS Press) pp 67–176