Transverse phase space tomography in an accelerator test facility using image compression and machine learning

A. Wolski[*]
*University of Liverpool, Liverpool, United Kingdom,
and the Cockcroft Institute, Daresbury, United Kingdom*

M. A. Johnson, M. King, B. L. Militsyn, and P. H. Williams
STFC/ASTeC, Daresbury Laboratory, Daresbury, United Kingdom

(Received 5 September 2022; accepted 14 November 2022; published 15 December 2022)

We describe a novel technique, based on image compression and machine learning, for transverse phase space tomography in two degrees of freedom in an accelerator beamline. The technique has been used in the CLARA accelerator test facility at Daresbury Laboratory: results from the machine learning method are compared with those from a conventional tomography algorithm (algebraic reconstruction) and applied to the same data. The use of machine learning allows reconstruction of the 4D phase space distribution of the beam to be carried out much more rapidly than using conventional tomography algorithms and also enables the use of image compression to reduce significantly the size of the data sets involved in the analysis. Results from the machine learning technique are at least as good as those from the algebraic reconstruction tomography in characterizing the beam behavior, in terms of the variation of the beam size in response to variation of the quadrupole strengths.

DOI: 10.1103/PhysRevAccelBeams.25.122803

I. INTRODUCTION

Phase space tomography provides a valuable technique for understanding the properties of a beam in a particle accelerator and has been applied in a range of different machines, for example [1–8]. However, conventional tomography techniques present some challenges, including the presence of artifacts in the reconstruction (which can be especially prominent when the number of projections is limited), and the computational time and resources required to construct the phase space distribution with good resolution. Tomography in two transverse degrees of freedom allows characterization of betatron coupling, but the sizes of the data structures required for the analysis increase rapidly with the dimensionality of the system. Storage of a 4D phase space distribution in an array with dimension N along each axis requires a data structure of N^4 numerical values, and the memory resources needed while processing the input data to construct the phase space can be much larger. The demands on computing power increase rapidly with increasing dimensionality of the phase space, and this may limit the use of high-dimensional phase space tomography (with good resolution) in applications where it could make a valuable contribution to machine operation, for example in short-pulse, short-wavelength free electron lasers [9] or injectors for machines using novel acceleration technologies such as plasma cells or dielectric wakefield structures [10,11].

Recent work [11] has shown (in simulation) how phase space tomography can be performed in $2\frac{1}{2}$ degrees of freedom to provide transverse phase space properties as a function of longitudinal position along a bunch. Steps have been taken toward full 6D phase space tomography, but the methods that have been employed (which include the use of machine learning) have not so far allowed the full reconstruction of the 6D phase space [12]. Where betatron coupling or synchrobetatron coupling is present, tracking a beam from a given point in the accelerator to determine its properties as function of position in the beamline requires the full phase space in the coupled degrees of freedom to be described, and in complex machines where multiple correlations can be present, full 6D phase space reconstruction would provide all the necessary information. Techniques allowing reduction of the processing time and data storage requirements for high-dimensional phase space tomography offer the prospect of enabling routine complete and detailed characterization of the charge distribution within bunches in an accelerator, including all cross-plane correlations, with significant benefits for advanced accelerator facilities.

In principle, image compression techniques can be used to reduce the size of the data structures needed to store and

[*] a.wolski@liverpool.ac.uk

process tomography data while maintaining the potential for reconstructing the phase space with a given resolution. Reduction in the size of the data sets can also be accompanied by a reduction in the time taken to process those data sets. However, it is not clear how existing tomography algorithms can be adapted so that they can be applied directly to compressed data. Machine learning techniques offer an alternative to conventional tomography methods and have the potential to allow a direct tomographic analysis of data in a compressed form. Machine learning is already extensively used for image analysis and tomography, particularly in medical contexts [13]. There is also increasing interest in the use of machine learning for a range of applications in accelerator design and operation, including design optimization [14–16], modeling [17], collection and analysis of diagnostic data [18–21], and operational optimization [22,23]. Recent work [12,24,25] has shown the use of neural networks for constructing two-dimensional projections of a six-dimensional phase space, including the use of adaptive feedback in the architecture of the neural network (allowing for the analysis of cases where the beam distribution in phase space varies with time). In [25], principal component analysis is used to reduce the size of the input data by decomposing each data set into a number of modes and selecting those that make dominant contributions to the data. Although data compression generally leads to some loss of information, it is possible by optimizing the number of components to achieve high-resolution phase space reconstructions while still benefitting from a significant reduction in the size of the input data sets.

In the current paper, we report the results of experimental studies aimed at demonstrating the use of machine learning for phase space tomography, working with beam images and phase space distributions stored in compressed form. We describe the principles of the technique, compare the results with those using a conventional tomography algorithm on the same data sets, and discuss the potential advantages of the use of machine learning for this application.

The experimental work that we present has been carried out on CLARA, the Compact Linear Accelerator for Research and Applications at Daresbury Laboratory [26–28]. Relevant features of the facility are outlined in Sec. II, in which we also describe the experimental technique (Sec. II A), and present the results of an analysis of the experimental data using a conventional tomography algorithm, algebraic reconstruction (Sec. II B). In Sec. III, we describe and present results from the tomography analysis based on machine learning. Some conclusions from the work are discussed in Sec. IV.

II. CHARACTERIZATION OF TRANSVERSE PHASE SPACE IN CLARA USING A CONVENTIONAL TOMOGRAPHY TECHNIQUE

Previous studies of phase space tomography in two transverse degrees of freedom using CLARA were reported in [29]. At the time of the previous tomography studies, carried out in 2019, the facility (CLARA Front End), included an electron source and short linac designed to provide bunches at a repetition rate of 10 Hz with charge up to 250 pC, momentum up to 50 MeV/c, and transverse emittance below 1 μm. Because of technical limitations during the tomography data collection, measurements in 2019 were made with a beam momentum of 30 MeV/c and bunch charge up to 50 pC. Since then, CLARA has undergone further development, with a number of changes to components and layout; however, the recent measurements reported here were made with parameters comparable to those used in the original study, specifically with beam momentum 35 MeV/c and bunch charge up to 100 pC. Further development of CLARA is planned, both to extend the energy reach and to test new rf gun technology, in particular, a low-emittance high repetition-rate source (high repetition-rate gun, HRRG). Detailed characterization of the HRRG performance will include studies of the transverse phase space. Work to develop novel phase space tomography techniques, in particular, making use of image compression and machine learning has been motivated by the need to facilitate beam characterization in CLARA generally and HRRG performance in particular. The results reported here are from recent measurements on CLARA in its current form, with the existing 10-Hz rf electron gun.

A. Experimental technique: Design parameters and calibrated model

The tomography technique described in [29] was applied to CLARA, following upgrade work performed since the previous tomography studies. Some changes were made to the details of the experimental procedure to take account of changes in the beam optics and machine layout; however, the overall procedure remained the same in its essential points. A beam momentum of 35 MeV/c was used. Measurements were made using a section of beamline between the exit of the linac (the “reconstruction point”) and a fluorescent screen on which the transverse beam profile could be observed (the “observation point”). The beamline between the reconstruction point and the observation point contains five quadrupoles.

To prepare for the measurements, a machine model [30] was used to determine gradients for the five quadrupoles in the measurement section that would allow control of the betatron phase advances between the reconstruction and observation points, while keeping approximately constant the beta functions at the observation point (see the schematic layout of CLARA in Fig. 1). A sequence of 32 sets of quadrupole gradients was determined, providing a good range of variation in horizontal and vertical betatron phase advance over the sequence. Maintaining constant and approximately equal beta functions at the observation point helps to provide good conditions for beam profile

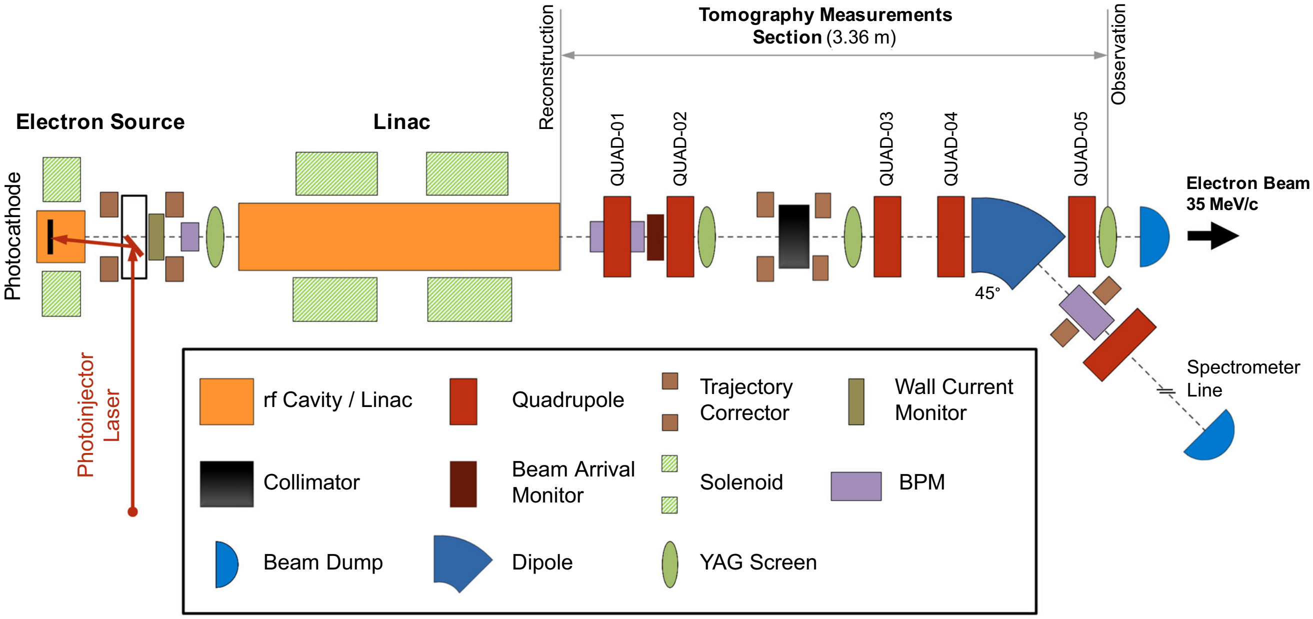


FIG. 1. Layout of CLARA, showing the electron source, magnetic elements, linac, and diagnostics. Distances between elements are shown approximately to scale. The optical functions and phase space distribution at the exit of the linac (the reconstruction point) were calculated from the measured beam profiles on the third YAG screen after the linac (the observation point) for different strengths of the five quadrupoles, QUAD-01 to QUAD-05.

measurements: if the beam image has a large aspect ratio or gets too large or too small, it can be difficult to determine accurately the beam sizes. Data collection consisted of recording the beam profile for each of the 32 steps in the sequence. The order of steps in the sequence was chosen to minimize the changes in the strength of the magnets from each step to the next, and in particular to avoid as far as possible changes in polarity: this helps to reduce the frequency with which the magnets need to be degaussed (the quadrupoles were degaussed at the start of each scan and midway through the scan). At each step, ten screen images were recorded plus an extra image with the photoinjector laser turned off to allow for the subtraction of the background resulting from the dark current. A complete quadrupole scan was carried out first with a bunch charge of 10 pC and then with bunch charge of 100 pC. Although space-charge effects in the injector are significant at 100 pC, in the measurements section at beam momentum 35 MeV/c space charge has little impact.

The analysis presented here is carried out in normalized phase space: this helps to improve the accuracy of the phase space reconstruction [31]. Since the section of beamline in CLARA where the measurements were performed consists only of drift spaces and normal quadrupoles, coupling can be neglected in constructing the normalizing transformation; however, it should be emphasized that the data analysis nevertheless still allows for full characterization of any coupling in the beam. Normalized horizontal phase space coordinates (x_N, p_{xN}) at a particular location along the beamline are related to the physical coordinates (x, p_x) by

$$\begin{pmatrix} x_N \\ p_{xN} \end{pmatrix} = \begin{pmatrix} \frac{1}{\sqrt{\beta_x}} & 0 \\ \frac{\alpha_x}{\sqrt{\beta_x}} & \sqrt{\beta_x} \end{pmatrix} \begin{pmatrix} x \\ p_x \end{pmatrix}, \tag{1}$$

where α_x and β_x are the usual Courant–Snyder optics functions at the specified beamline location. If the phase space distribution is matched to the optics functions, then the distribution in normalized coordinates $\rho_N(x_N, p_{xN})$ will be invariant under rotations in phase space. Furthermore, the transport matrices in normalized phase space are simply rotation matrices (through angles corresponding to the phase advance), so a matched phase space distribution will be invariant under linear transport along the beamline.

In practice, the phase space distribution is not known in advance: the goal of the measurement is to determine the distribution. Phase space normalization cannot, therefore, be carried out using optics functions known to be exactly matched to the phase space distribution. Instead, a machine model is used to generate an expected distribution, and the optics functions describing this distribution are used to normalize the phase space. If the real beam distribution is reasonably close to that expected from the machine model, then in the normalized phase space, the real beam distribution will have at least approximate rotational symmetry. Phase space tomography (in normalized phase space) can be used to determine the actual distribution, which can be transformed back to the physical coordinates using the inverse of the normalizing transformation given in Eq. (1).

TABLE I. Parameter values in the design and calibrated CLARA machine models. Optics functions are given at the tomography reconstruction point (the exit of the linac).

Parameter	Design model	Calibrated model
β_x	15.0 m	5.5 m
α_x	−3.4	0.0
β_y	15.0 m	5.5 m
α_y	−3.4	1.5
Quadrupole magnetic length	100.7 mm	127.0 mm

For the measurements in CLARA, a design model of the machine was used to predict the phase space beam distribution at the reconstruction point (the exit of the linac: see Fig. 1). The values of the optics functions are shown in Table I. A preliminary analysis of the experimental data was carried out using the parameter values from the design model. The results indicated substantial differences between the design values and the real values, largely arising from differences between the operational settings actually used for the injector and linac, and the settings assumed in the machine model when preparing for the experiments. Furthermore, closer investigations found that the magnetic lengths of the quadrupoles in the beamline following the linac (the section used for the tomography studies) were somewhat larger than had been thought, resulting in changes in the transfer matrices between the reconstruction point and the observation point for the quadrupole gradients (calculated using the design model) used during the quadrupole scan.

Differences between the design parameters and the calibrated model are evident in Fig. 2, which shows the beta functions at the observation point and the phase advances from reconstruction to the observation point, at each step in the quadrupole scan using the design quadrupole gradients. Note that the steps were not followed in the order in Fig. 2, which shows the steps in order of increasing horizontal phase advance, followed by increasing vertical phase advance: as mentioned above, the actual order of the steps during the measurements was designed to minimize the changes in quadrupole strengths between successive steps, to reduce the need for degaussing. The quadrupole gradients used in the scan were determined using the design model (top plots in Fig. 2); the same gradients, when used in the calibrated model with the revised quadrupole lengths and optics functions, lead to the observation point beta functions and phase advances shown in the bottom plots in Fig. 2. Following the initial analysis of the quadrupole scan data using the design parameters, the analysis was repeated using the parameters for the calibrated model (and the transfer matrices calculated using the design quadrupole gradients). The optics for the design model are shown in Fig. 2 only to illustrate the intended conditions for the tomography data collection and for comparison with those for the calibrated model. In the remainder of this work, we refer only to the calibrated model.

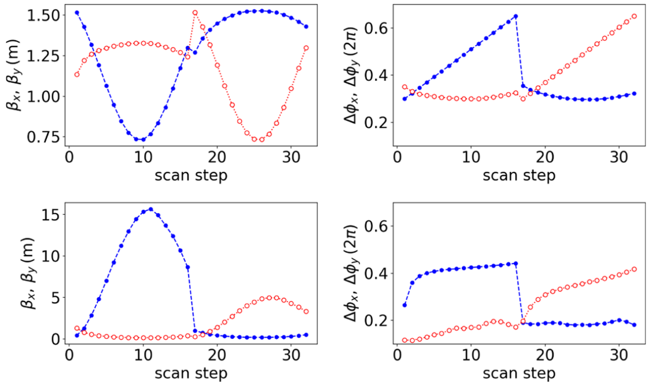


FIG. 2. Optics functions in the CLARA tomography measurements, in the design model (top) and the calibrated model (bottom). Left-hand plots show the beta functions at the observation point at each step in the quadrupole scan; right-hand plots show the phase advances from the reconstruction point to the observation point at each step in the quadrupole scan. Blue solid points (with dashed lines) show the values in the horizontal plane; red open points (with dotted lines) show the values in the vertical plane.

B. Quadrupole scan analysis using the algebraic reconstruction tomography technique

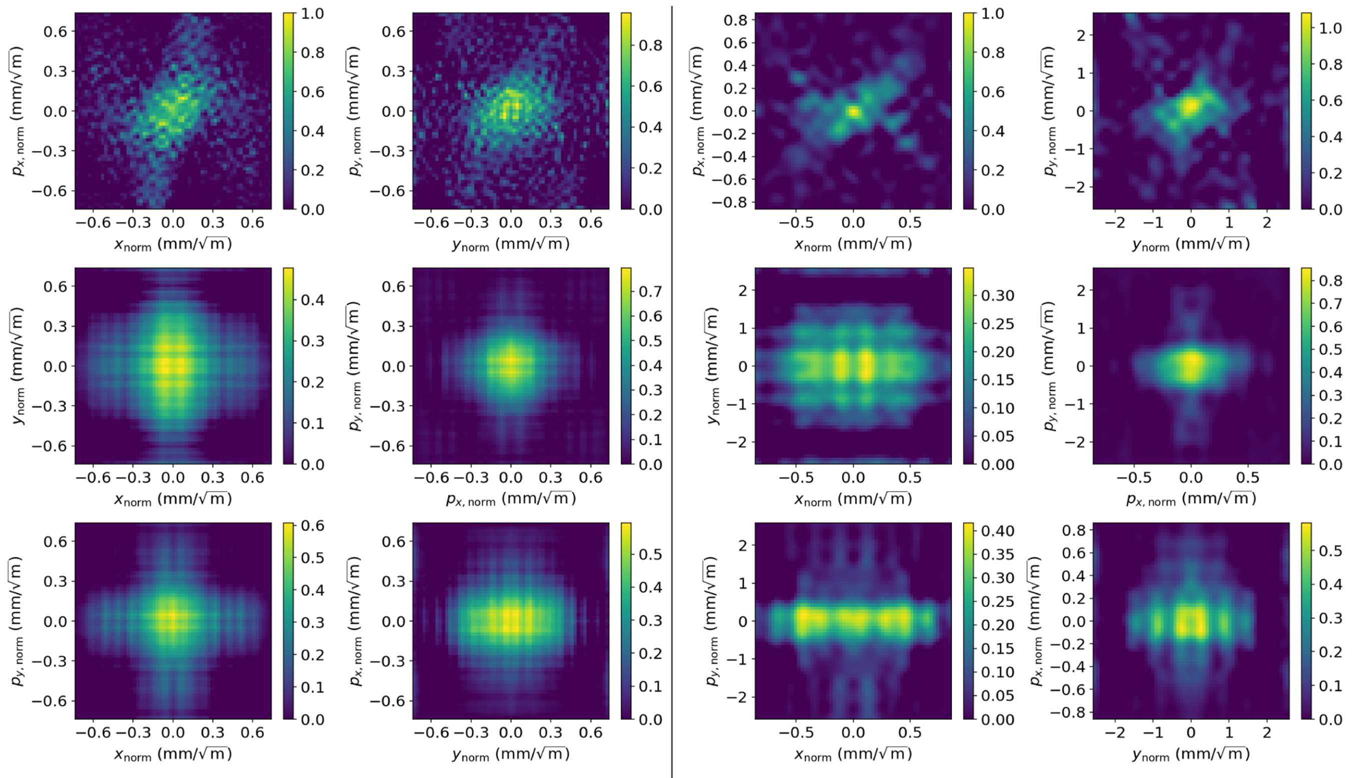
Screen images collected during the quadrupole scans were used in an algebraic reconstruction tomography (ART) code, to determine the 4D transverse phase space charge distribution. The same tomography code was used for the recent data as was used in the studies on CLARA FE: the earlier work included validation of the code, using simulated data [29]. In principle, since the only changes in machine settings made during the course of a quadrupole scan are to the quadrupole gradients, the phase space distribution at the reconstruction point (in the current studies, at the exit of the linac, and upstream of the quadrupoles) should vary a little during the scan.

Beam images collected during a quadrupole scan are prepared for the tomography analysis by first subtracting a background image (to remove any artifacts from dark current) and then cropping and scaling the images. To crop the images, we remove the area outside a certain range of pixels from the point of peak intensity in the image. The same cropping range is used on each step in the quadrupole scan so that the cropped images all have the same dimensions in pixels. The crop limits are chosen so that the beam occupies as much of the cropped images as possible, without clipping the beam in any of the images. To scale the images, we demagnify each image along each axis by the square root of the beta function corresponding to that axis (while maintaining the same number of pixels in each image). In effect, scaling means that given an initial calibration factor in mm/pixel, the calibration factor after scaling will be in mm/$\sqrt{\mathrm{m}}$/pixel. The beta functions used for scaling are found in the optics in the calibrated model (propagating the values from the reconstruction point to the observation point, using the transfer matrix calculated from the corresponding quadrupole strengths). Scaling essentially transforms the images to normalized phase space: this means that if the phase space distribution at the reconstruction point was correctly matched to the optics in the calibrated model, then the scaled beam size (in pixels) would remain constant over the course of the quadrupole scan. Finally, the resolution of the normalized images is reduced (or increased, if necessary) to 39 × 39 pixels.

For the tomography analysis (using ART), we reconstruct the 4D phase space with a resolution equal (in pixels) to the image resolution, i.e., 39 pixels on each axis. The phase space resolution is not in principle constrained by the technique, but is a practical choice, decided by a balance between the desired level of detail in the reconstructed phase space distribution and the computation time and resources needed for the analysis (which can increase rapidly with increasing phase space resolution). The results of the tomography can be validated by transporting, for each step in the quadrupole scan, the 4D phase space distribution from the reconstruction point to the observation point using the transfer matrix calculated from the known quadrupole strengths and drift lengths; and then comparing the projection onto coordinate space with the corresponding observed beam image.

Projections of the reconstructed 4D phase space distribution are shown in Fig. 3 for 10- and 100-pC bunch charges. Note that the scales on the axes for each image are given in normalized phase space (units of mm/$\sqrt{\mathrm{m}}$). Validation images for 10- and 100-pC bunch charges are shown in Fig. 4 for three steps in the quadrupole scan. The screen images are generally reproduced from the coordinate space projection of the reconstructed phase space distribution with good accuracy, supporting the validity of the reconstructed 4D phase space distribution. The screen images with 100-pC bunch charge show significantly more structure than those with 10-pC bunch charge, though the additional structure is not immediately apparent from the projections of the 4D phase space distribution at the exit of the linac. The richer beam structure observed with 100-pC bunch charge is believed to be associated with the properties of the photoinjector laser.

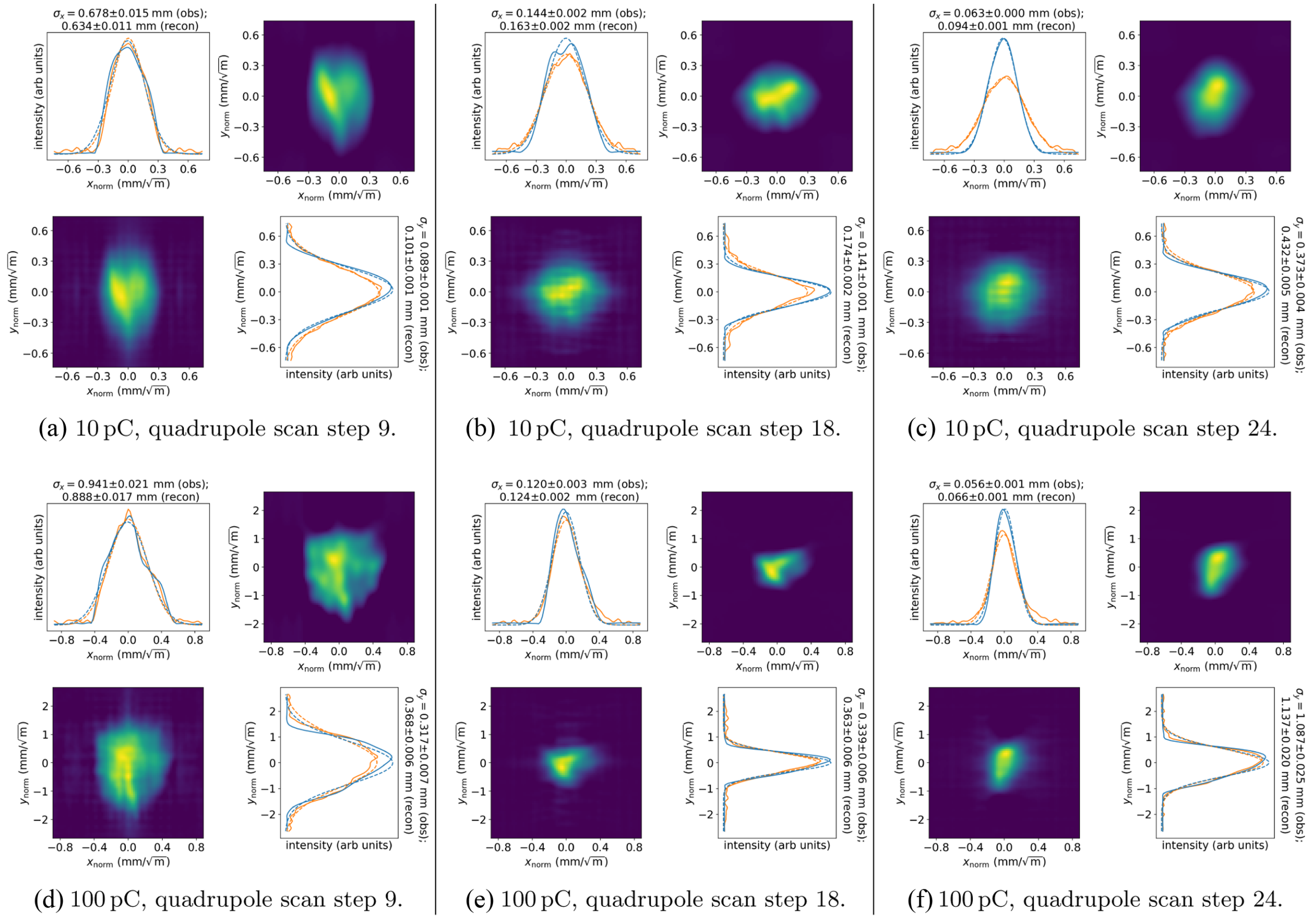
Variations in the beam size at the observation point over the course of a quadrupole scan are shown in Fig. 5. The plots (upper plot for 10-pC bunch charge and lower plot for 100 pC) compare the beam sizes calculated in four different ways: (i) The solid lines (labeled "linear optics") show the beam sizes (calculated at each point in the quadrupole scan) found by calculating the covariance matrix



(a) 10 pC

(b) 100 pC

FIG. 3. Projections of the 4D phase space distribution of the beam in CLARA at the exit of the linac, for (a) 10-pC bunch charge and (b) 100-pC bunch charge, found from algebraic reconstruction tomography.



(a) 10 pC, quadrupole scan step 9.

(b) 10 pC, quadrupole scan step 18.

(c) 10 pC, quadrupole scan step 24.

(d) 100 pC, quadrupole scan step 9.

(e) 100 pC, quadrupole scan step 18.

(f) 100 pC, quadrupole scan step 24.

FIG. 4. Validation images from algebraic reconstruction tomography, from three steps in the quadrupole scan at two different bunch charges: 10 pC in cases (a), (b), (c); and 100 pC in cases (d), (e), (f). Within each set of four plots, the top right and bottom left images show (respectively) the observed and reconstructed beam image at the observation point; continuous lines in the top-left and bottom-right plots show the density projected onto (respectively) the horizontal and vertical axes, broken lines show Gaussian fits (used to determine the beam sizes, with values shown alongside the relevant plots). Blue lines correspond to the observed image, and orange lines correspond to the reconstructed image. Note the different scales on the coordinate axes for 10-pC and 100-pC bunch charges.

describing the reconstructed 4D phase space distribution at the reconstruction point and then transporting the covariance matrix to the observation point. The shaded bands indicate the uncertainties in the beam sizes arising from the uncertainties in the elements of the covariance matrix. (ii) Crosses (labeled "observed beam size" in Fig. 5) show the rms beam sizes obtained from Gaussian fits to projections of the observed beam images onto the horizontal and vertical axes. The error bars indicate the standard deviations of the rms beam sizes over the ten images collected at each step (which dominate over uncertainties associated with the Gaussian fits). (iii) The circular markers (labeled "calibrated model") show the beam sizes at each point in the quadrupole scan expected from the lattice functions in the calibrated model, with emittances found from the reconstructed 4D phase space. The error bars show the uncertainty arising from the uncertainty on the emittance (increased by a factor of 10, to make the error bars more clearly visible). (iv) Points (dots, labeled "tomography") show the rms beam sizes obtained from Gaussian fits to projections (onto the horizontal and vertical axes) of the reconstructed 4D phase space transported from the reconstruction point to the observation point. The error bars in this case indicate the uncertainties in the fit. Although there is a qualitative agreement between the beam sizes in the calibrated model (using the optics functions shown in Table I) and the observed beam sizes, there is a better agreement with the observed beam sizes in the case of linear transport of the covariance matrix calculated from the reconstructed phase space distribution and in the case of linear transport of the phase space distribution.

For completeness, and for comparison of the results from tomographic analysis using ART and analysis using machine learning, the emittances and optics functions at the reconstruction point are given in Table II. The values shown are calculated from the covariance matrices describing the reconstructed 4D phase space distributions, for 10- and

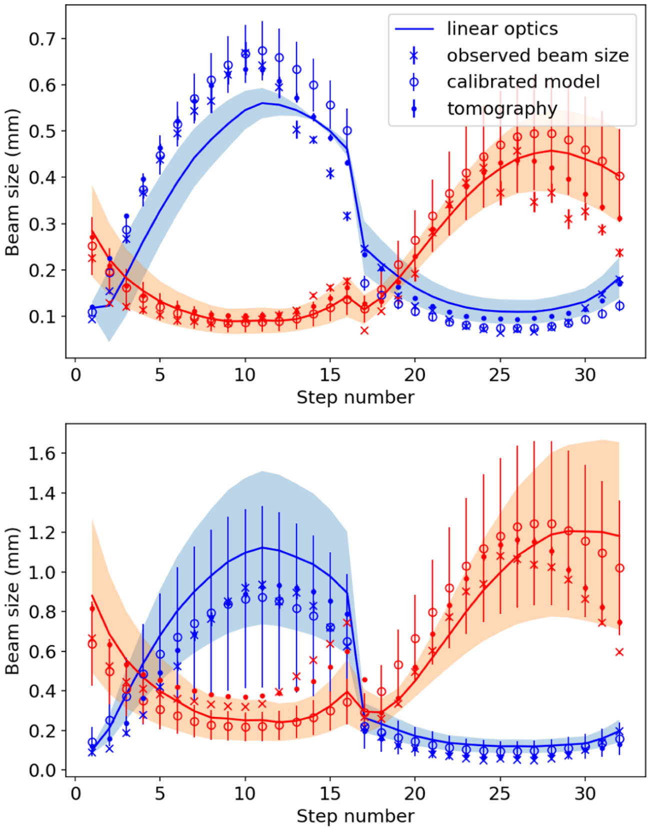


FIG. 5. Variation in horizontal (blue points and lines) and vertical (red points and lines) at the observation point, for 10 pC bunch charge (top) and 100 pC bunch charge (bottom). Error bars on the observed beam sizes (marked as crosses) show the standard deviation of Gaussian fits to the ten beam images collected at the observation point for each step in the quadrupole scan. Error bars on the beam sizes from the tomographic reconstruction (solid points) show the uncertainty in a Gaussian fit to the phase space density projected onto the horizontal or vertical axis. Open circles show the beam sizes calculated by propagating the lattice functions for the calibrated model (Table I) from the reconstruction point to the observation point, and combining with the emittances calculated by a fit to the 4D phase space from ART tomography (Table II). The line shows the beam sizes obtained by propagating the covariance matrix fitted to the 4D phase space distribution reconstructed by ART (Table II), with a shaded range showing the uncertainty arising from the uncertainties on the elements of the covariance matrix.

100-pC bunch charges. Note that the values given are for the normal mode emittances $\gamma\epsilon_{\mathrm{I}}$, $\gamma\epsilon_{\mathrm{II}}$ and optics functions, B^{I}, B^{II} [32]. In terms of these quantities, the covariance matrix is expressed:

$$\Sigma = \epsilon_{\mathrm{I}} B^{\mathrm{I}} + \epsilon_{\mathrm{II}} B^{\mathrm{II}}, \tag{2}$$

where the elements of the covariance matrix are the second-order moments of the beam distribution over all combinations of phase space variables:

$$\Sigma_{ij} = \langle x_i x_j \rangle, \tag{3}$$

with $x_i = x, p_x, y, p_y$, for $i = 1, 2, 3, 4$, respectively. The symmetric matrices B^k can be written in terms of 2×2 submatrices σ^k_{uu} (with $u = x$ or y):

$$B^k = \begin{pmatrix} \sigma^k_{xx} & \sigma^k_{xy} \\ (\sigma^k_{xy})^{\mathrm{T}} & \sigma^k_{yy} \end{pmatrix}. \tag{4}$$

In the absence of coupling,

$$\sigma^{\mathrm{I}}_{xx} = \begin{pmatrix} \beta_x & -\alpha_x \\ -\alpha_x & -\gamma_x \end{pmatrix}, \qquad \sigma^{\mathrm{II}}_{yy} = \begin{pmatrix} \beta_y & -\alpha_y \\ -\alpha_y & -\gamma_y \end{pmatrix}, \tag{5}$$

and

$$\sigma^{\mathrm{I}}_{yy} = \sigma^{\mathrm{II}}_{xx} = \sigma^{\mathrm{I}}_{xy} = \sigma^{\mathrm{II}}_{xy} = 0. \tag{6}$$

III. PHASE SPACE TOMOGRAPHY USING MACHINE LEARNING

Although the results shown in Sec. II suggest that the algebraic reconstruction technique can be of value in constructing the 4D transverse phase space distribution of the beam in a machine such as CLARA, the method can have some limitations. First, the structures visible in the beam images at the observation point (especially at the higher bunch charge) are not clearly evident in any of the projections shown of the 4D phase space distribution at the reconstruction point. The reasons for this are not well understood: it may simply be a result of the relatively poor resolution with which the 4D phase space distribution is determined, or it may be that the orientation of the distribution in phase space is such as to obscure the structure for the chosen 2D projections—note that the structures seen at the observation point are only really evident for particular steps in the quadrupole scan, i.e., for some specific range of betatron phase advances.

A second limitation of the algebraic reconstruction technique is that it can take some time to process the data to obtain the phase space distribution. The demands in terms of processing time and computational resources increase rapidly with increasing resolution of the reconstruction and with increasing dimensionality of the phase space. For the results presented here, a phase space resolution of 39 pixels in each dimension of the 4D phase space is used: this limits the detail visible in the phase space but allows the reconstruction to be completed reasonably rapidly (within a few minutes) using a standard PC. Where a high resolution is required, or a rapid reconstruction would be of value (for example, for several iterations of machine tuning), then more powerful computing resources may be needed if algebraic reconstruction, or a similar tomography technique, is to be used. There is also interest in extending tomography from four to five or six dimensions [12,16]: this can be of particular value in

TABLE II. Emittances and lattice functions describing the 4D phase space distributions obtained by phase space tomography in CLARA. The values given refer to the normal modes [32]. In the absence of coupling, the elements of the matrices $\sigma^{\rm I}_{xx}$ and $\sigma^{\rm II}_{yy}$ are (respectively) β_x and β_y (top left elements) and $-\alpha_x$ and $-\alpha_y$ (top right elements); and all other matrices are zero.

	10 pC		100 pC	
	Algebraic reconstruction	Machine learning	Algebraic reconstruction	Machine learning
$\gamma\varepsilon_{\rm I}$	$1.99 \pm 0.04\ \mu$m	$1.98 \pm 0.03\ \mu$m	$3.35 \pm 0.44\ \mu$m	$3.38 \pm 0.16\ \mu$m
$\gamma\varepsilon_{\rm II}$	$3.39 \pm 0.19\ \mu$m	$2.08 \pm 0.06\ \mu$m	$21.4 \pm 1.7\ \mu$m	$18.0 \pm 0.3\ \mu$m
$\sigma^{\rm I}_{xx}$	$\begin{pmatrix} 8.63\ {\rm m} & 1.06 \\ 1.06 & 0.245/{\rm m} \end{pmatrix}$	$\begin{pmatrix} 4.60\ {\rm m} & 0.260 \\ 0.260 & 0.200/{\rm m} \end{pmatrix}$	$\begin{pmatrix} 14.4\ {\rm m} & 0.782 \\ 0.782 & 0.112/{\rm m} \end{pmatrix}$	$\begin{pmatrix} 8.46\ {\rm m} & 0.389 \\ 0.389 & 0.136/{\rm m} \end{pmatrix}$
$\sigma^{\rm II}_{yy}$	$\begin{pmatrix} 7.11\ {\rm m} & 2.14 \\ 2.14 & 0.786/{\rm m} \end{pmatrix}$	$\begin{pmatrix} 3.44\ {\rm m} & 1.06 \\ 1.06 & 0.572/{\rm m} \end{pmatrix}$	$\begin{pmatrix} 11.7\ {\rm m} & 3.67 \\ 3.67 & 1.24/{\rm m} \end{pmatrix}$	$\begin{pmatrix} 4.50\ {\rm m} & 1.62 \\ 1.62 & 0.804/{\rm m} \end{pmatrix}$
$\sigma^{\rm I}_{xy}$	$\begin{pmatrix} -0.355\ {\rm m} & -0.0675 \\ -0.0760 & -0.0221/{\rm m} \end{pmatrix}$	$\begin{pmatrix} -0.933\ {\rm m} & -0.0708 \\ -0.181 & -0.0897/{\rm m} \end{pmatrix}$	$\begin{pmatrix} -1.02\ {\rm m} & -0.306 \\ -0.100 & -0.0317/{\rm m} \end{pmatrix}$	$\begin{pmatrix} -0.370\ {\rm m} & -0.148 \\ -0.0306 & -0.0185/{\rm m} \end{pmatrix}$
$\sigma^{\rm I}_{yy}$	$\begin{pmatrix} 0.0238\ {\rm m} & 0.00667 \\ 0.00667 & 0.00218/{\rm m} \end{pmatrix}$	$\begin{pmatrix} 0.278\ {\rm m} & 0.0739 \\ 0.0739 & 0.0407/{\rm m} \end{pmatrix}$	$\begin{pmatrix} 0.101\ {\rm m} & 0.0314 \\ 0.0314 & 0.00978/{\rm m} \end{pmatrix}$	$\begin{pmatrix} 0.0177\ {\rm m} & 0.00781 \\ 0.00781 & 0.00375/{\rm m} \end{pmatrix}$
$\sigma^{\rm II}_{xx}$	$\begin{pmatrix} 0.0196\ {\rm m} & 0.00189 \\ 0.00189 & 0.000555/{\rm m} \end{pmatrix}$	$\begin{pmatrix} 0.334\ {\rm m} & 0.0242 \\ 0.0242 & 0.0194/{\rm m} \end{pmatrix}$	$\begin{pmatrix} 0.0376\ {\rm m} & -0.00173 \\ -0.00173 & 0.000135/{\rm m} \end{pmatrix}$	$\begin{pmatrix} 0.0181\ {\rm m} & -0.000737 \\ -0.000737 & 0.000330/{\rm m} \end{pmatrix}$
$\sigma^{\rm II}_{xy}$	$\begin{pmatrix} 0.245\ {\rm m} & 0.0342 \\ 0.0624 & 0.0197/{\rm m} \end{pmatrix}$	$\begin{pmatrix} 0.859\ {\rm m} & 0.0914 \\ 0.209 & 0.105/{\rm m} \end{pmatrix}$	$\begin{pmatrix} 0.148\ {\rm m} & -0.00876 \\ 0.0179 & 0.00863/{\rm m} \end{pmatrix}$	$\begin{pmatrix} 0.261\ {\rm m} & 0.0685 \\ 0.00419 & 0.0100/{\rm m} \end{pmatrix}$

short-wavelength free-electron lasers, for example, where understanding the transverse beam profile and energy spread as a function of longitudinal position in the bunch can be of significant importance.

Approaches based on machine learning may offer ways to address some of the issues associated with conventional tomography techniques for the reconstruction of the beam phase space in four (or more) dimensions. The method presented here, which we apply to the two transverse degrees of freedom, uses a pretrained neural network, to which the beam images at the observation point are provided, in compressed form, as input; the output from the neural network consists of the 4D phase space distribution, again in compressed form. In principle, using a neural network in this way allows a rapid (almost immediate) reconstruction of the 4D phase space distribution once the beam images are provided. The computing resources needed for carrying out the reconstruction can also be much more modest than those needed for algebraic reconstruction tomography. If images in uncompressed form are used, the input and output data sets can still be of significant size, but the use of machine learning enables image compression techniques to be applied, reducing the size of input and output data sets. In principle, a neural network can be trained on images and phase space distributions represented in some chosen compressed form, for example, as discrete cosine transforms (DCTs) [33–35]. Image compression would be difficult to apply in the case of conventional tomography methods, which usually rely on a relationship between the sinogram and the object to be reconstructed that is intrinsically expressed in regular coordinate space. Neural networks offer much greater flexibility and do not require a specific representation of the input or output data.

In using a neural network to perform tomographic reconstruction, an issue does arise with the need to train the network. Training must necessarily be based on simulated data, which would ideally include features characteristic of the beam; but at least in cases where the beam shows some detailed structure, the relevant features may not be known at the time of generating the training data. In the current study, we simply take the approach of generating random phase spaces consisting of a number of superposed 4D Gaussian distributions, with the component distributions in each generated phase space varying randomly in position, shape, and intensity. Given the shape of the phase space distribution in CLARA suggested by tomography using ART, the phase space distributions constructed in this way may not provide ideal training data; however, it is interesting to consider the ability of a neural network to reconstruct phase space distributions presenting features significantly different from those present in the training data. If the techniques described here are to be of value in a reasonably wide range of situations, then they should be able to reproduce phase space distributions with features significantly different from those in the training data: this issue is discussed further in Sec. III B.

A. Implementation of machine learning method

Before presenting the results of tomography using machine learning, we discuss some further details of how the technique was implemented.

For the preparation of training data, phase space distributions were generated as mentioned above, by superposing 4D Gaussian distributions with random variations in position, shape, and intensity. The distributions are constructed in normalized phase space; the sinograms are then obtained by transforming the distribution using phase space rotations (corresponding to the steps in a quadrupole scan) and then projecting the distribution onto the (normalized) x–y plane at each step in the quadrupole scan. Note that we used phase advances corresponding to those in the calibrated model, as shown in Fig. 2 (bottom right). For consistency, it is important that the phase advances should match those resulting from the quadrupole strengths applied in the quadrupole scan, given the lattice functions used for normalizing the phase space. It should be emphasized, however, that the chosen lattice functions do not need to match those describing the actual beam distribution (which, in general, is not known in advance).

Having obtained the sinograms for the simulated 4D phase space distributions, we compress both the phase space distributions and the sinograms using discrete cosine transforms (DCTs). There are several types of DCT: we use a type II DCT, which is the default in many standard scientific computing packages. In the case of a 2D $M \times N$ array, a type II DCT is defined by

$$y_{jk} = \sum_{m=0}^{M-1}\sum_{n=0}^{N-1} x_{mn} \cos\left(\pi j \frac{2m+1}{2M}\right) \cos\left(\pi k \frac{2n+1}{2N}\right), \quad (7)$$

where the values x_{mn} is the component of the initial array, and y_{jk} (for $j = 0 \ldots M-1$, $k = 0 \ldots N-1$) is the component of the transformed array. Compression is achieved by truncating the transformed array at some point, either defined in terms of the magnitudes of the components (which should all be below some specified threshold beyond the truncation point) or simply in terms of a fixed limit on the size of the transformed array. The inverse of the type II DCT of an $M \times N$ array is given by

$$x_{mn} = \frac{1}{MN} \times \sum_{j=0}^{M-1}\sum_{k=0}^{N-1} \alpha_{jk} y_{jk} \cos\left(\pi j \frac{2m+1}{2M}\right) \cos\left(\pi k \frac{2n+1}{2N}\right), \quad (8)$$

where

$$\alpha_{jk} = \begin{cases} 1 & \text{if } j = k = 0, \\ 2 & \text{if } j = 0, k \neq 0, \text{or } j \neq 0, k = 0, \\ 4 & \text{if } j \neq 0, k \neq 0. \end{cases} \quad (9)$$

The expressions in (7) and (8) can be extended to higher-dimensional arrays by including an additional summation for each additional index and making the appropriate modification to the numerical factors in (8). Truncating the transformed array corresponds to reducing the upper limits on the summations in the inverse transformation (8); in this case, the array x_{mn} is reconstructed with approximated values for its elements, but the number of elements in the array remains the same. In the case of an image, the effect of truncating the DCT is to lose some of the fine detail. Figure 6 illustrates image compression using DCTs truncated to different sizes, using (as an example) a beam image collected during the quadrupole scan with 100 pC bunch charge. The original image has a resolution (in pixels) $M \times N = 161 \times 161$. Truncating the DCT to 21×21 results in some loss of clarity, but the main features and some details can still be clearly seen. Truncation to 16×16 results in more significant loss of detail. The "optimum" truncation will depend on the experimental context and should take into account factors such as the machine and beam properties and the quality of diagnostic data. For the case of CLARA, the number of steps in the quadrupole scan will be one of the main limitations on the

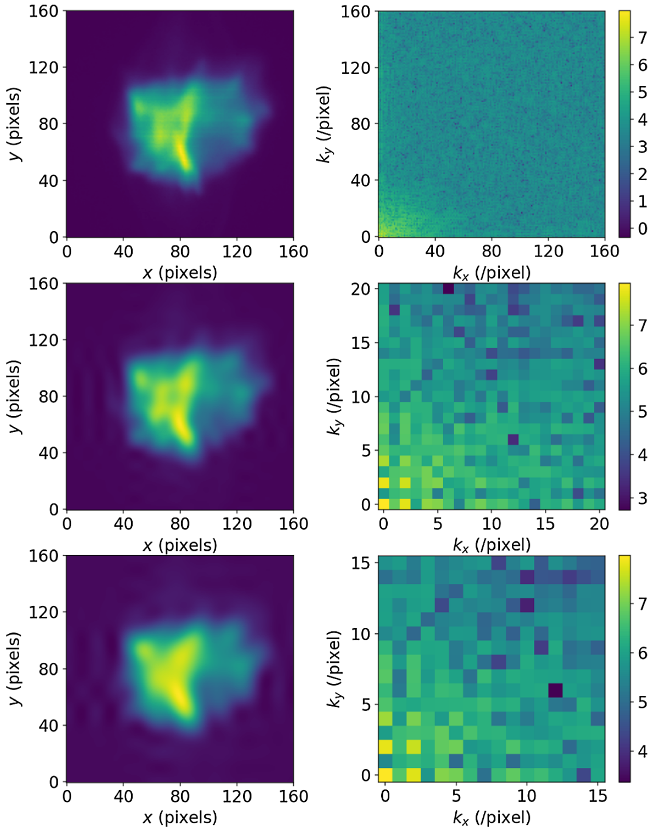


FIG. 6. Image compression using truncation of the DCT. In each pair, the left-hand image shows the beam image in coordinate space reconstructed from the DCT shown in the right-hand image. Images are reconstructed at the resolution of the original image by padding the truncated DCT with zeros as necessary. Top: full resolution, 161×161 pixels. Middle: DCT truncated to 21×21. Bottom: DCT truncated to 16×16. Beam images are from a quadrupole scan with 100-pC bunch charge. The color scale shows the logarithm (to base 10) of the absolute value of the DCT component.

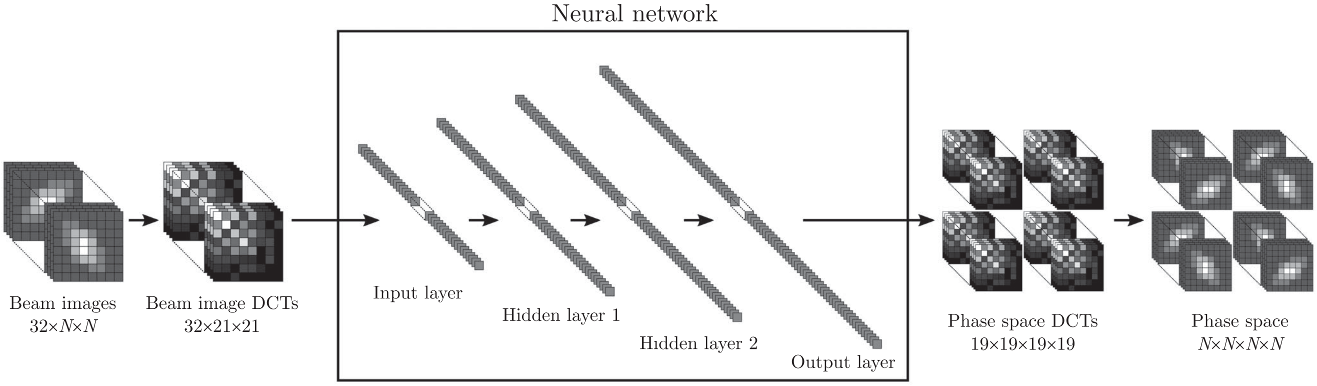


FIG. 7. Schematic showing the steps in the phase space tomography using image compression and machine learning. The beam images collected during a quadrupole scan are compressed by applying a discrete cosine transform (DCT) to each image. The transformed (and compressed) images are provided as input to a neural network, consisting of an input layer, two hidden (dense) layers, and an output layer. The output of the neural network is a DCT of the 4D phase space distribution of the beam: applying an inverse DCT allows reconstruction of the phase space density at any desired resolution.

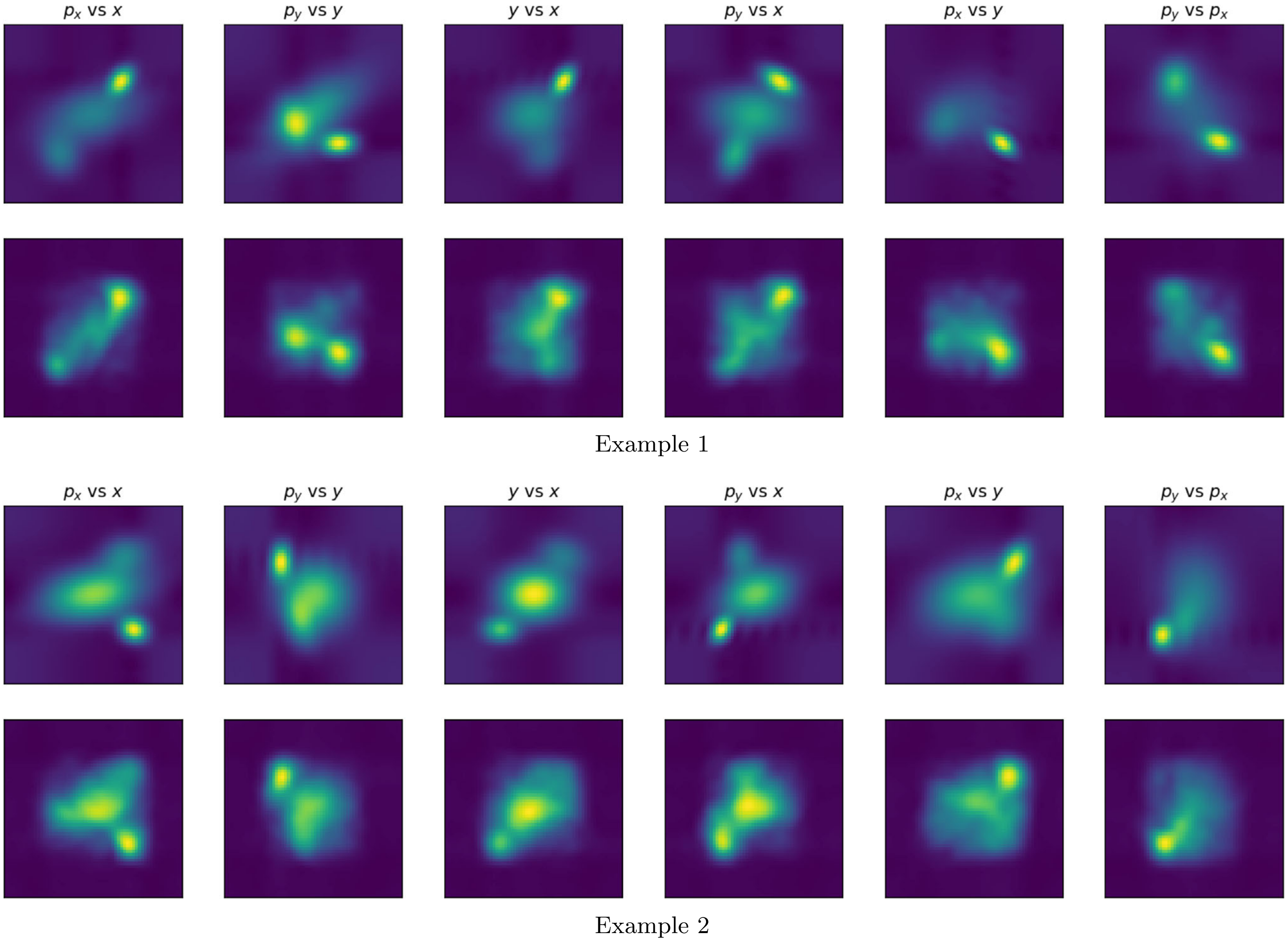


FIG. 8. Examples of reconstruction of charge density in 4D phase space using a neural network. Two different cases are shown, selected from the set of training data but reserved from (i.e., not used during) the training process. In each example, plots in the top row show different projections of the original phase space distribution from which the sinograms (projections onto x-$y$coordinate space following phase space rotations corresponding to steps in a quadrupole scan) are constructed. Plots in the bottom row show the corresponding projections from the 4D phase space reconstructed from the sinograms using the neural network.

level of detail that can be achieved in reconstructing the phase space distribution. Furthermore, in the experimental results that we present here, the range of phase advances from the reconstruction point to the observation point was restricted because of the differences between the design model and the actual machine conditions: this is evident, for example, from the ART reconstruction of the phase space distribution. Although a reconstruction retaining, in principle, smaller-scale details of the beam distribution could be achieved by increasing the number of DCT modes, under current conditions on CLARA, there would be a little practical benefit in doing so.

The training data for the neural network consist of some number of pairs of the DCTs of the sinograms (input) and corresponding phase space distributions (output). The neural network itself is implemented in Keras [36]. We use a rather straightforward architecture. Apart from the input and output layers, there are two hidden layers, defined as dense layers in Keras. To limit overtraining, each dense layer is followed by a dropout layer. We use a resolution of 19 points on each axis for the DCT of the 4D phase space (i.e., 19^4 voxels in total), and a resolution of 21×21 for the DCT of each 2D projection in the set of “images” forming the sinogram. In practice, these resolutions capture sufficient numbers of DCT modes to allow the representation of the screen images and the 4D phase space with good resolution. Note that the size of the data for the 4D phase space using machine learning (19^4) is substantially smaller than the size used for the ART tomography reported in Sec. II (39^4). To return to the issue mentioned above, regarding the loss in detail of the distribution resulting from truncation of the DCTs, we have found that for the data collected in CLARA, increasing the numbers of DCT modes, either in the input sinograms or the reconstructed phase space, does not improve the quality of the results as judged by a comparison between the projections of the phase space at the observation point and the original beam images (as shown, for example, in Fig. 13). In constructing the sinogram, we use phase space rotations corresponding to the phase advances in the calibrated model (see Fig. 2, bottom-right plot), i.e., with 32 steps in the quadrupole scan. With these parameters, the neural network has an input layer with 32×21^2 nodes and an output layer with 19^4 nodes. We use 1500 and 3000 nodes for the first and second hidden (dense) layers, respectively, with a dropout layer specified to set 20% of inputs (selected randomly) to zero for each dense layer during training. The tomography process using image compression and machine learning is illustrated schematically in Fig. 7.

The architecture of the neural network is rather simple, and more sophisticated structures could of course be used. For example, convolutional neural networks (CNNs) are used in [12,24,25]. CNNs are often extremely effective for image analysis tasks; however, in the present case, we have found that better results are generally obtained using dense (not convolutional) layers. The reason for this requires further investigation; however, it may be that CNNs offer limited benefit in the present case because the discrete cosine transform leads to an “encoded” relationship between pixels in local areas of the sinogram. It may be possible, using appropriate design principles, to obtain the benefits from a CNN despite the additional complexity (from point of view of a CNN) associated with the encoding of image information in the DCT. We hope to investigate a wider range of neural network architectures in the future, but at present, the relatively simple network structure based on dense layers is more than sufficient for the current operational needs of CLARA.

A total of 3000 sets of 4D phase space distributions and sinograms were generated as training data; 100 sets were reserved as validation sets for testing the performance of the trained network and were not used in the training process itself. The training was carried out using the Adam optimization algorithm [37]. Training takes several minutes on a standard laptop PC. The training time is comparable to the time taken to process a single data set using ART; however, training only needs to be performed once, to produce a neural network that can (in principle) be applied to any data set collected in a quadrupole scan using given quadrupole strengths. The ART analysis would need to be performed separately for each data set.

B. Testing the neural network using simulated data

Two examples illustrating results from the trained network are shown in Fig. 8. The examples are selected at random from the validation data sets. Each row of images in the figure shows a different projection of a 4D phase space: in each example, the top row shows the projections from the original phase space, and the bottom row shows the

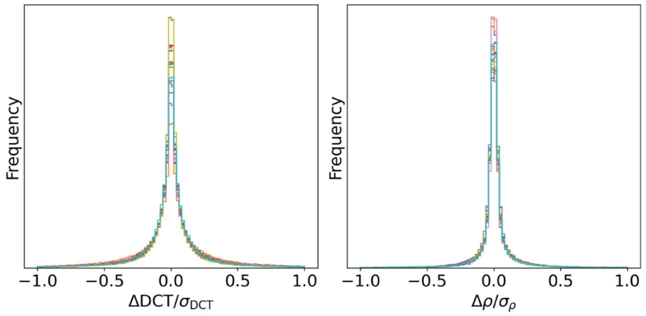


FIG. 9. Residuals of the fit to the phase space density using the trained neural network. Left: histograms showing the frequency of different values of $\Delta\mathrm{DCT}/\sigma_{\mathrm{DCT}}$, where $\Delta\mathrm{DCT}$ is the difference between the known DCT values (in one case from the test data) and the values found by the neural network from the corresponding sinograms, and σ_{DCT} is the standard deviation of the DCT values. Right: histograms showing the same residual analysis, but using the phase space densities, rather than the DCT of the densities. About 20 cases are superposed in each plot: there is little variation between the cases in the distributions of residuals.

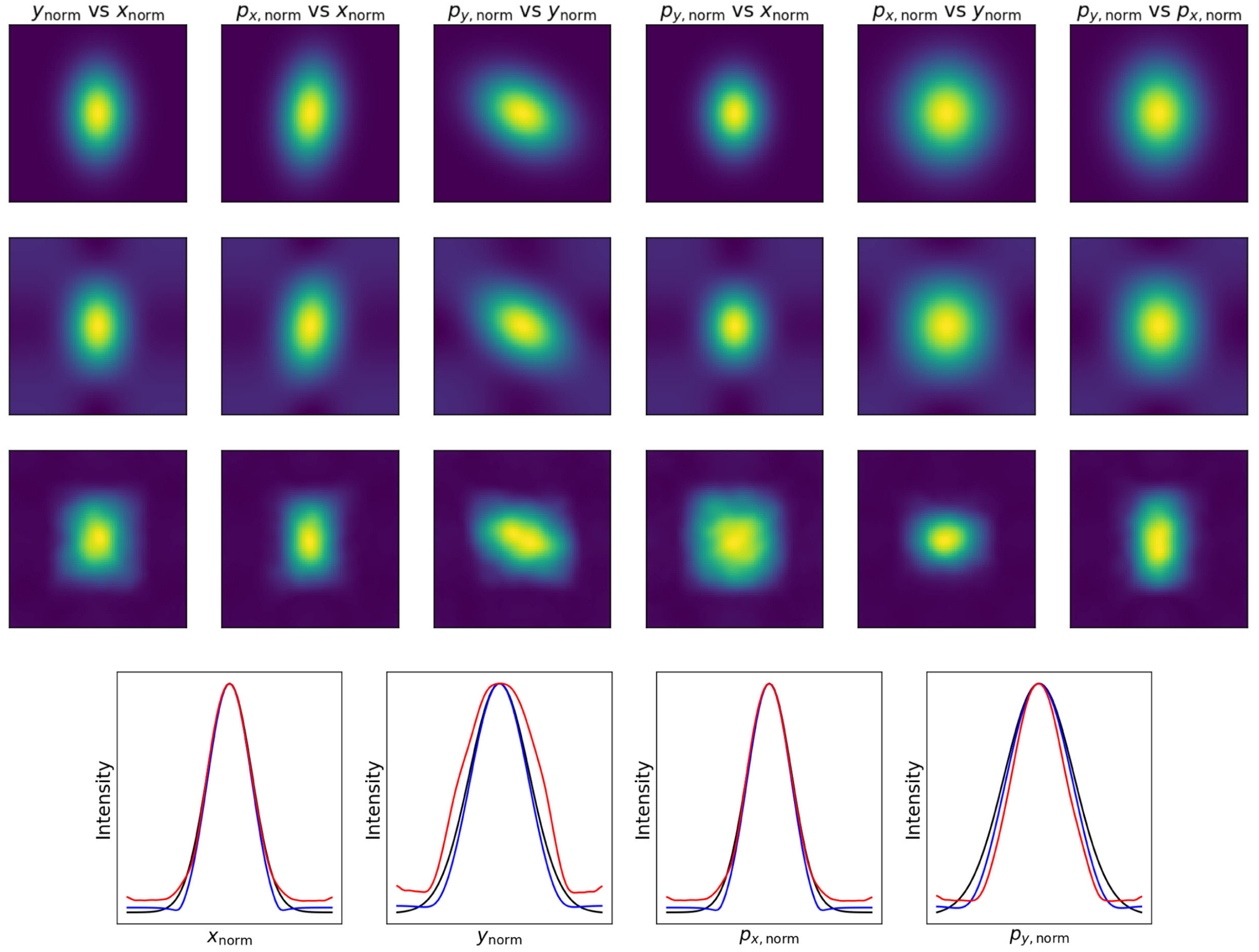


FIG. 10. Comparison of phase space projections between an original simulated phase space based on a simple Gaussian distribution (top row); the distribution reconstructed after compression and decompression using a discrete cosine transform (second row); and reconstruction, using a trained neural network, from images obtained from simulation of a quadrupole scan (third row). The bottom row shows 1D projections of the distribution onto the phase space axes, for the original distribution (black line), the distribution after DCT compression and decompression (blue line), and the distribution reconstructed using the neural network (red line).

projections from the phase space reconstructed by the neural network when provided with the (DCTs of the) corresponding sinograms. While there are clearly some differences between the original and the reconstructed phase spaces, the reconstruction is sufficiently similar to the original to provide a useful practical indication of the beam distribution in phase space.

To characterize further the reliability of the machine learning reconstruction of the phase space, we calculate the residuals between the original phase space density in the test data and the phase space density found from the sinograms using the trained neural network. The residuals are shown in Fig. 9, as histograms of $\Delta\mathrm{DCT}/\sigma_{\mathrm{DCT}}$ and $\Delta\rho/\sigma_\rho$. Here, ΔDCT is the difference between a particular DCT coefficient predicted by the neural network, and the corresponding DCT coefficient in the phase space distribution used to generate the sinogram data provided as input to the network. σ_{DCT} is the standard deviation of the DCT coefficients. $\Delta\rho$ is the difference in the phase space density (at a particular element of 4D phase space) between the original distribution and the distribution found by the neural network, after performing an inverse DCT of the network output; and σ_ρ is the standard deviation of the phase space density. Figure 9 shows histograms of these quantities for 20 cases from the validation data sets. Typically, between 75% and 80% of phase space density values from the neural network are within 0.1 σ_ρ of the true phase space density.

A potential issue in the application of the neural network for experimental data is the extent to which the training data represent the phase space beam distribution in the actual machine. Ideally, the phase space distribution in CLARA will be a 4D Gaussian; however, it was known from screen images collected during other work on the machine that the distribution contained significant non-Gaussian structure and

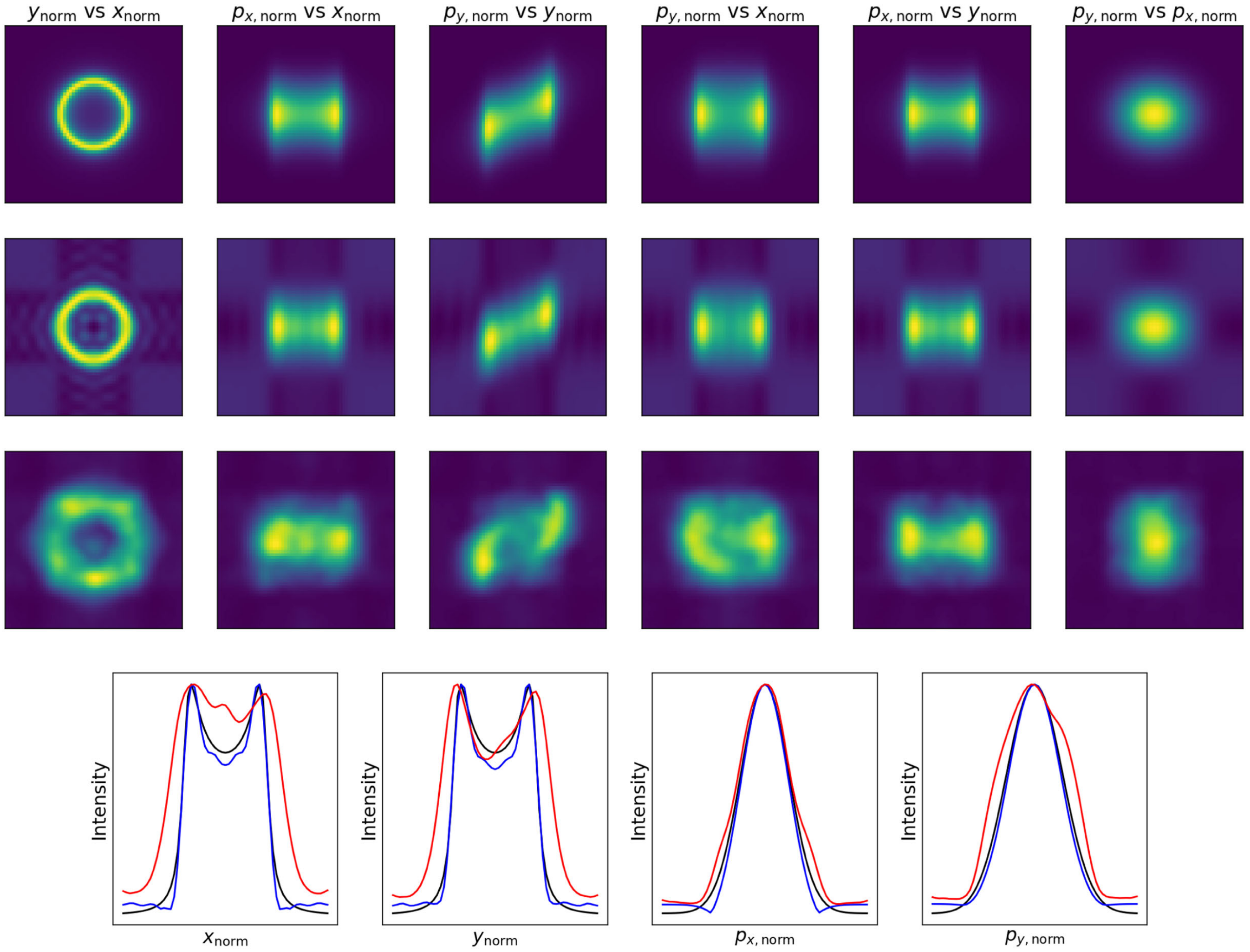


FIG. 11. Comparison of phase space projections between an original simulated phase space based on a circular "ring" in coordinate space (top row); the distribution reconstructed after compression and decompression using a discrete cosine transform (second row); and reconstruction, using a trained neural network, from images obtained from simulation of a quadrupole scan (third row). The bottom row shows 1D projections of the distribution onto the phase space axes, for the original distribution (black line), the compressed and decompressed distribution (blue line), and the distribution reconstructed using the neural network (red line).

that the features visible in the images depended strongly on machine conditions (including, for example, injector and linac settings, and bunch charge). One of the main motivations for the tomography studies presented here was the need to develop a much better understanding of the phase space distribution and its dependence on machine settings.

To test the ability of the neural network to reproduce a phase space distribution with features significantly different from the training data, a number of different distributions were constructed, and the corresponding sinograms were computed in simulation. In constructing the sinograms, the calibrated machine model (as described in Sec. II A) was used. Two examples, illustrating the results, are shown in Figs. 10 and 11. The example in Fig. 10 is based on a simple Gaussian distribution and is intended to show that the neural network does not introduce artificial structure not present in the real phase space distribution and can reproduce with reasonable accuracy the size and shape of the distribution (corresponding to the emittances and lattice functions). Values for the covariance matrix elements found from the original distribution and from the distribution reconstructed using the neural network with simulated quadrupole scan results are compared in Table III. The results suggest that for a simple Gaussian distribution, the reconstruction of the 4D phase space distribution is reliable. Although there is some evidence of non-Gaussian structure in the reconstructed phase space, it should be remembered that the experimental settings for the tomography procedure are not ideal (because of limitations on the number of steps in the quadrupole scan and differences in the optics between the design model and calibrated model) and will impose some limitations on the accuracy with which the phase space distribution can be reconstructed, irrespective of the technique used.

TABLE III. Covariance matrix elements in the simple Gaussian test distribution shown in Fig. 10. The values are obtained from 1D Gaussian fits to projections from the 4D phase space onto the appropriate axes and are given in arbitrary units. Values from the original simulated distribution are compared with the reconstruction from the neural network with simulated quadrupole scan results (first and third rows in Fig. 10, respectively).

Matrix element	Simulated distribution	Neural network reconstruction
$\langle x^2\rangle$	0.619	0.637
$\langle xp_x\rangle$	−0.075	−0.006
$\langle xy\rangle$	$<10^{-4}$	−0.012
$\langle xp_y\rangle$	$<10^{-4}$	0.052
$\langle p_x^2\rangle$	1.625	1.626
$\langle p_xy\rangle$	$<10^{-4}$	$<10^{-4}$
$\langle p_xp_y\rangle$	$<10^{-4}$	$<10^{-4}$
$\langle y^2\rangle$	1.336	1.651
$\langle yp_y\rangle$	0.239	0.197
$\langle p_y^2\rangle$	1.040	0.920

The second example from tests of the neural network with simulated data, shown in Fig. 11, is based on a distribution that is far from ideal (for the real machine) and highly artificial: the distribution consists of a ring in coordinate space, with some correlation between transverse momentum and coordinate in the vertical direction. Although the projection of the distribution in coordinate space shows significant blurring, it is still possible to identify the key features of the distribution. The important conclusion from these tests is that the neural network can still produce meaningful and useful results in terms of the reconstructed 4D phase space distribution of the beam, even for cases where the distribution is significantly different from the distributions used in training the neural network.

As a further remark on the tests (in simulation) of the neural network, it is worth noting that although there is some loss in detail of the phase space distribution from the use of image compression using the discrete cosine transform, this does not appear to be a limitation on the accuracy of reconstruction of the phase space distribution. More extensive simulation studies suggest that one of the main limitations on the accuracy of the reconstruction is the quality of the experimental data: this is evident also in the work reported in previous tomography studies on CLARA using ART [29] (where the machine conditions were initially better understood), and in the ART analysis of the recent experimental data, shown in Sec. II B.

C. Experimental results from tomography using machine learning

The trained neural network was applied to the analysis of the quadrupole scan data collected on CLARA, described

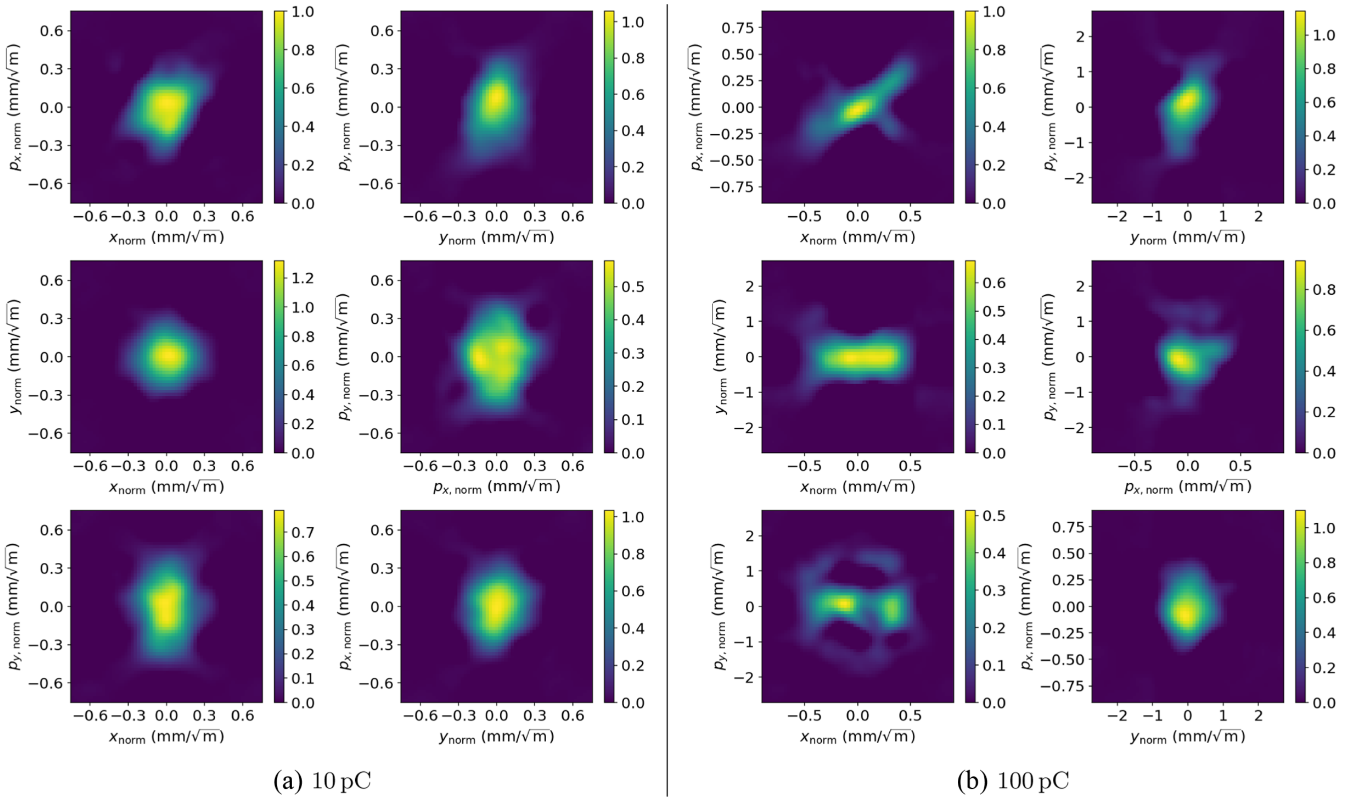
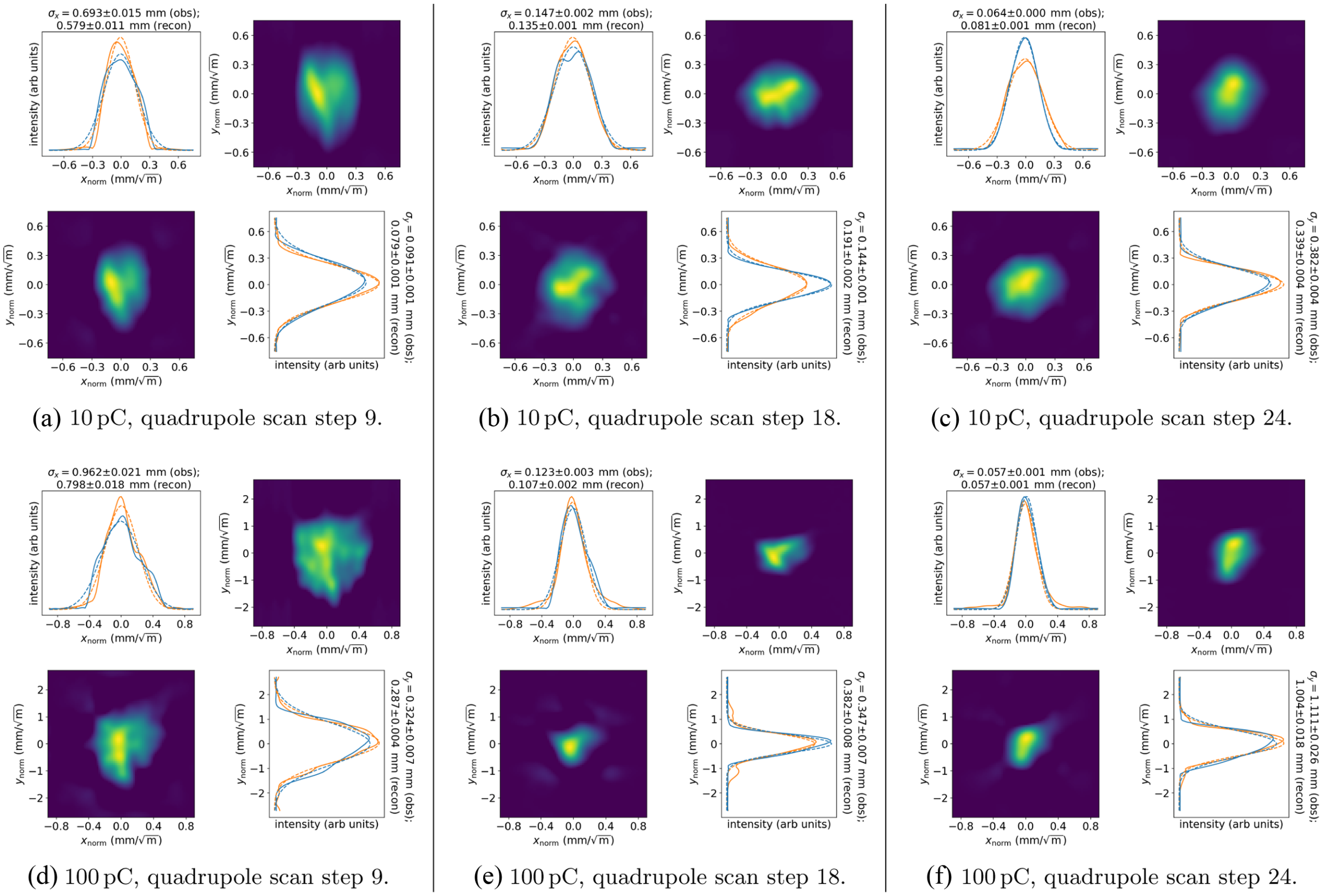


FIG. 12. Projections of the 4D phase space distribution of the beam in CLARA at the exit of the linac, for (a) 10-pC bunch charge and (b) 100-pC bunch charge, found from machine learning.

in Sec. II. The screen images from each step of the quadrupole scan were prepared, in the same way as for the ART analysis, by cropping and then scaling to transform to normalized phase space. The images were then compressed by constructing the DCTs, which were truncated to 21 modes on each axis. The DCTs were provided as input to the trained neural network, which provided the DCT of the 4D phase space distribution, with a resolution of 19 modes along each axis. Projections from the reconstructed 4D phase space distribution for 10-pC and 100-pC bunch charges are shown in Fig. 12.

In Sec. II, we validated the ART reconstruction of the 4D phase space distribution by comparing the projection of the distribution onto x–y coordinate space at the observation point with the observed beam images at different steps of the quadrupole scan. We can make similar comparisons to validate the 4D phase space distribution reconstructed using the neural network: some examples (for the same steps as shown in Fig. 4) are shown in Fig. 13. Once again, we see generally good agreement between the projection of the 4D phase space distribution and the observed images, in both the 10-pC and the 100-pC cases. Comparing with projections from the phase space reconstructed using ART in Fig. 4, the machine learning projections do not all have the same clarity, in terms of the finer details in some of the images. It should be remembered, however, that ART tomography uses beam images with a resolution of 39 × 39 pixels to reconstruct the 4D phase space distribution with a resolution of 39 pixels on each axis. The machine learning technique uses beam images and 4D phase space in a compressed form: the beam images are represented by 21 DCT modes on each axis, and the phase space is represented by 19 DCT modes on each axis. Although this is sufficient to capture a significant amount of detail, the truncation of the DCTs means that the compression is not lossless. Given the compression ratio, the machine learning method retains a reasonable level of detail in the phase space distribution.

Comparisons between the observed and reconstructed beam sizes are shown in Fig. 14: the results here can be



(a) 10 pC, quadrupole scan step 9.
(b) 10 pC, quadrupole scan step 18.
(c) 10 pC, quadrupole scan step 24.
(d) 100 pC, quadrupole scan step 9.
(e) 100 pC, quadrupole scan step 18.
(f) 100 pC, quadrupole scan step 24.

FIG. 13. Validation images for 10-pC bunch charge, found from machine learning, from three steps in the quadrupole scan. Within each set of four plots, the top right and bottom left images show (respectively) the observed and reconstructed beam image at the observation point; continuous lines in the top left and bottom right plots show the density projected onto (respectively) the horizontal and vertical axes, broken lines show Gaussian fits (used to determine the beam sizes, with values shown alongside the relevant plots). Blue lines correspond to the observed image, and orange lines correspond to the reconstructed image.

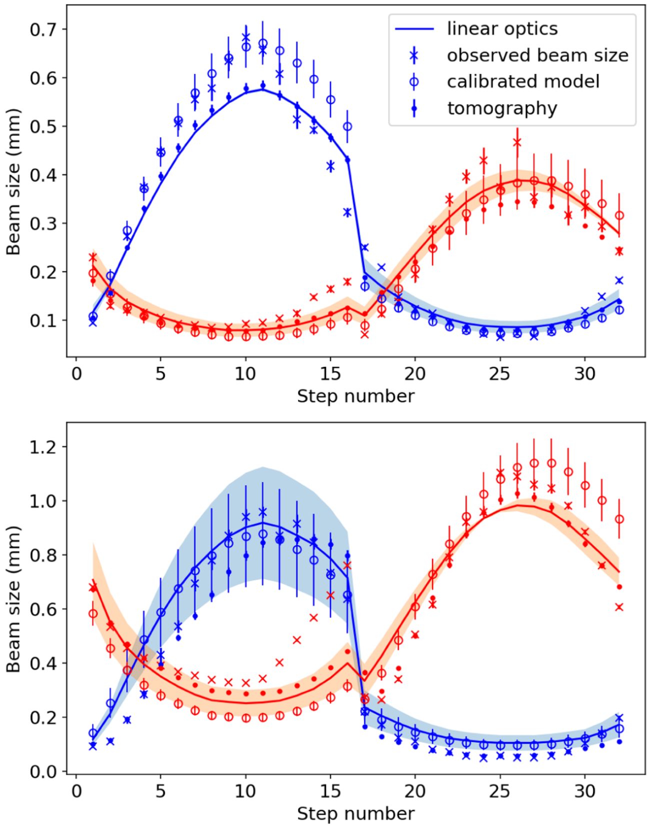


FIG. 14. Variation in horizontal (blue points and lines) and vertical (red points and lines) at the observation point, for 10-pC bunch charge (top) and 10-pC bunch charge (bottom). Error bars on the observed beam sizes (marked as crosses) show the standard deviation of Gaussian fits to the ten beam images collected at the observation point for each step in the quadrupole scan. Error bars on the beam sizes from the tomographic reconstruction (solid points) show the uncertainty in a Gaussian fit to the phase space density projected onto the horizontal or vertical axis. Open circles show the beam sizes calculated by propagating the lattice functions for the calibrated model (Table I) from the reconstruction point to the observation point and combining them with the emittances calculated by a fit to the 4D phase space from machine learning (Table II). The line shows the beam sizes obtained by propagating the covariance matrix fitted to the 4D phase space distribution reconstructed by machine learning (Table II).

compared with those in Fig. 5, which shows the beam sizes reconstructed using ART. While there are some differences in detail in the quality of the match between the beam sizes expected from the reconstructed phase space and the beam sizes observed during the quadrupole scan, both the ART and the machine learning techniques show similar performance in describing the beam behavior.

IV. CONCLUSIONS AND POSSIBLE FURTHER DEVELOPMENTS

The machine learning technique we have described in this paper uses relatively simple methods for reconstructing the 4D phase space. Nevertheless, this approach appears capable of producing useful results, as shown by the comparison between projections onto *x*–*y* coordinate space at the observation point for different quadrupole strengths and beam images collected over the course of a quadrupole scan. Values obtained for parameters describing the distribution (emittances and lattice functions) are consistent with those obtained using a conventional tomography technique. Data collection and analysis were planned using a design model of the machine; despite significant differences between the design model and the actual machine conditions during the collection of experimental data, results from both the ART and the machine learning techniques provide useful information on the beam properties in CLARA.

The use of image compression (in the present case, using discrete cosine transforms) allows a reduction of the size of the data sets that need to be processed, in particular, for representing the 4D phase space. Machine learning allows direct tomographic analysis of compressed beam images and phase space representations, without the additional complications or difficulties that would be encountered in attempting to apply conventional tomography techniques to compressed images.

Inspection of projections of the 4D phase space onto various planes (in particular, comparison of Fig. 3 with Fig. 12) suggests that the machine learning technique is capable of producing a representation of the 4D phase space distribution that appears clearer than that obtained by the conventional tomography algorithm. On the other hand, the beam images obtained by projecting the 4D phase space distribution onto coordinate space at the observation point have slightly higher fidelity in the case of the conventional tomography technique. Nevertheless, the consistency in the results from the two methods suggests that the machine learning approach could have some practical value, especially since the neural network provides results from experimental data almost immediately, whereas the conventional technique requires a potentially lengthy computation time. There are a number of ways in which the machine learning approach could be further developed. With an improved understanding of the operational conditions of CLARA, some optimization would be possible in terms of the quadrupole strengths (and the number of steps) used in the quadrupole scan. More sophisticated neural network architectures or the use of more sophisticated machine learning tools generally, could lead to a better reconstruction of the 4D phase space distribution from a given set of sinograms. There may be some benefits in further increasing the number of sets of training data. An indication of the quality to be expected in the reconstruction can be obtained using simulated data, for example, by calculating the residuals as shown in Fig. 9. Although the phase space distributions in the training data we used for the neural network had very different features from the phase space distribution in the real machine, the trained network was still capable of reconstructing a phase space distribution that provided a good description of beam behavior. It is possible, however, that using training data more closely resembling the real beam (once some initial characterization of the beam has been obtained) could lead to better results.

Discrete cosine transforms may not be the optimal way to represent images and phase space distributions in compressed form for the application described here. A DCT essentially represents a multidimensional array as a set of orthogonal modes, with each mode described by a cosine function. This provides a convenient general purpose approach, but alternative basis functions may allow a more accurate representation of beam images and phase space distributions with fewer modes. It may be possible, for example, to take advantage of properties generally expected of the beam (such as approximate symmetries) to construct a more appropriate basis. The scope for further development is rather wide, and while the results shown here are encouraging and demonstrate the value of machine learning for tomographic reconstruction in principle, more extensive studies would be required to understand the full potential of the technique.

ACKNOWLEDGMENTS

We would like to thank our colleagues in STFC/ASTeC at Daresbury Laboratory for their help and support with various aspects of the simulation and experimental studies of CLARA. In particular, we would like to thank Amy Pollard for useful discussions and advice on machine learning. This work was supported by the Science and Technology Facilities Council, United Kingdom, through a grant to the Cockcroft Institute.

[1] C. B. McKee, P. G. O'Shea, and J. M. J. Madey, Phase space tomography of relativistic electron beams, Nucl. Instrum. Methods Phys. Res., Sect. A **358**, 264 (1995).

[2] V. Yakimenko, M. Babzien, I. Ben-Zvi, R. Malone, and X.-J. Wang, Electron beam phase-space measurement using a high-precision tomography technique, Phys. Rev. ST Accel. Beams **6**, 122801 (2003).

[3] D. Stratakis, R. A. Kishek, H. Li, S. Bernal, M. Walter, B. Quinn, M. Reiser, and P. G. O'Shea, Tomography as a diagnostic tool for phase space mapping of intense particle beams, Phys. Rev. ST Accel. Beams **9**, 112801 (2006).

[4] D. Stratakis, K. Tian, R. A. Kishek, I. Haber, M. Reiser, and P. G. O'Shea, Tomographic phase-space mapping of intense particle beams using solenoids, Phys. Plasmas **14**, 120703 (2007).

[5] Dao Xiang, Ying-Chao Du, Li-Xin Yan, Ren-Kai Li, Wen-Hui Huang, Chuan-Xiang Tang, and Yu-Zheng Lin, Transverse phase space tomography using a solenoid applied to a thermal emittance measurement, Phys. Rev. ST Accel. Beams **12**, 022801 (2009).

[6] Michael Röhrs, Christopher Gerth, Holger Schlarb, Bernhard Schmidt, and Peter Schmüser, Time-resolved electron beam phase space tomography at a soft x-ray free-electron laser, Phys. Rev. ST Accel. Beams **12**, 050704 (2009).

[7] Q. Z. Xing, L. Du, X. L. Guan, C. X. Tang, M. W. Wang, X. W. Wang, and S. X. Zheng, Transverse profile tomography of a high current proton beam with a multi-wire scanner, Phys. Rev. Accel. Beams **21**, 072801 (2018).

[8] Fuhao Ji, Jorge Giner Navarro, Pietro Musumeci, Daniel B. Durham, Andrew M. Minor, and Daniele Filippetto, Knife-edge based measurement of the 4D transverse phase space of electron beams with picometer-scale emittance, Phys. Rev. Accel. Beams **22**, 082801 (2019).

[9] D. Alesinia, G. Di Pirro, L. Ficcadenti, A. Mostacci, L. Palumbo, J. Rosenzweig, and C. Vaccarezza, RF deflector design and measurements for the longitudinal and transverse phase space characterization at SPARC, Nucl. Instrum. Methods Phys. Res., Sect. A **568**, 488 (2006).

[10] D. Marx, R. W. Assmann, P. Craievich, K. Floettmann, A. Grudiev, and B. Marchetti, Simulation studies for characterizing ultrashort bunches using novel polarizable X-band transverse deflection structures, Sci. Rep. **9**, 19912 (2019).

[11] S. Jaster-Merz, R. W. Assmann, R. Brinkmann, F. Burkart, and T. Vinatier, 5D tomography of electron bunches at ARES, in *Proceedings of the 13th International Particle Accelerator Conference, IPAC-2022, Bangkok, Thailand* (JACoW, Geneva, Switzerland, 2022), pp. 279–283.

[12] A. Scheinker, D. Filippetto, and F. Cropp, 6D phase space diagnostics based on adaptively tuned physics-informed generative convolutional neural networks, in *Proceedings of the 13th International Particle Accelerator Conference, IPAC-2022, Bangkok, Thailand* (JACoW, Geneva, Switzerland, 2022), pp. 776–779.

[13] G. Wang, J. C. Ye, and B. De Man, Deep learning for tomographic image reconstruction, Nat. Mach. Intell. **2**, 737 (2020).

[14] Y. Li, W. Cheng, L. H. Yu, and R. Rainer, Genetic algorithm enhanced by machine learning in dynamic aperture optimization, Phys. Rev. Accel. Beams **21**, 054601 (2018).

[15] J. Wan, P. Chu, Y. Jiao, and Y. Li, Improvement of machine learning enhanced genetic algorithm for nonlinear beam dynamics optimization, Nucl. Instrum. Methods Phys. Res., Sect. A **946**, 162683 (2019).

[16] A. Edelen, N. Neveu, M. Frey, Y. Huber, C. Maye, and A. Adelmann, Machine learning for orders of magnitude speedup in multiobjective optimization of particle accelerator systems, Phys. Rev. Accel. Beams **23**, 044601 (2020).

[17] C. Emma, A. Edelen, M. J. Hogan, B. O'Shea, G. White, and V. Yakimenko, Machine learning-based longitudinal phase space prediction of particle accelerators, Phys. Rev. Accel. Beams **21**, 112802 (2018).

[18] G. Azzopardi, G. Valentino, A. Muscat, and B. Salvachua, Automatic spike detection in beam loss signals for LHC

collimator alignment, Nucl. Instrum. Methods Phys. Res., Sect. A **934**, 10 (2019).

[19] X. Xu, Y. Zhou, and Y. Leng, Machine learning based image processing technology application in bunch longitudinal phase information extraction, Phys. Rev. Accel. Beams **23**, 032805 (2020).

[20] C. Tennant, A. Carpenter, T. Powers, A. S. Solopova, and L. Vidyaratne, Superconducting radio-frequency cavity fault classification using machine learning at Jefferson Laboratory, Phys. Rev. Accel. Beams **23**, 114601 (2020).

[21] Z. Omarov and S. Haciömeroğlu, Machine learning assisted non-destructive beam profile monitoring, Nucl. Instrum. Methods Phys. Res., Sect. A **1026**, 166132 (2022).

[22] L. Emery, H. Shang, Y. Sun, and X. Huang, Application of a machine learning based algorithm to online optimization of the nonlinear beam dynamics of the Argonne Advanced Photon Source, Phys. Rev. Accel. Beams **24**, 082802 (2021).

[23] P. Arpaia, G. Azzopardi, F. Blanc *et al.*, Machine learning for beam dynamics studies at the CERN Large Hadron Collider, Nucl. Instrum. Methods Phys. Res., Sect. A **985**, 164652 (2021).

[24] A. Scheinker, Adaptive machine learning for time-varying systems: Low dimensional latent space tuning, J. Instrum. **16**, P10008 (2021).

[25] A. Scheinker, F. Cropp, S. Paiagua, and D. Filipetto, An adaptive approach to machine learning for compact particle accelerators, Sci. Rep. **11**, 19187 (2021).

[26] J. A. Clarke *et al.*, CLARA conceptual design report, J. Instrum. **9**, T05001 (2014).

[27] D. Angal-Kalinin *et al.*, Status of CLARA Front End commissioning and first user experiments, in *Proceedings of the 10th International Particle Accelerator Conference, IPAC-2019, Melbourne, Australia* (JACoW, Geneva, Switzerland, 2019), pp. 1851–1854.

[28] D. Angal-Kalinin *et al.*, Design, specifications, and first beam measurements of the compact linear accelerator for research and applications front end, Phys. Rev. Accel. Beams **23**, 044801 (2020).

[29] A. Wolski, D. C. Christie, B. L. Militsyn, D. J. Scott, and H. Kockelbergh, Transverse phase space characterization in an accelerator test facility, Phys. Rev. Accel. Beams **23**, 032804 (2020).

[30] D. J. Scott, A. D. Brynes, and M. P. King, High level software development framework and activities on VELA/CLARA, in *Proceedings of the 10th International Particle Accelerator Conference, IPAC-2019, Melbourne, Australia* (JACoW, Geneva, Switzerland, 2019), pp. 1855–1858.

[31] K. M. Hock, M. G. Ibison, D. J. Holder, A. Wolski, and B. D. Muratori, Beam tomography in transverse normalised phase space, Nucl. Instrum. Methods Phys. Res., Sect. A **642**, 36 (2011).

[32] A. Wolski, Alternative approach to general coupled linear optics, Phys. Rev. ST Accel. Beams **9**, 024001 (2006).

[33] N. Ahmed, T. Natarajan, and K. R. Rao, Discrete cosine transform, IEEE Trans. Comput. **C-23**, 90 (1974).

[34] Wen-Hsiung Chen and W. Pratt, Scene adaptive coder, IEEE Trans. Commun. **32**, 225 (1984).

[35] K. R. Rao and P. Yip, *Discrete Cosine Transform: Algorithms, Advantages, Applications* (Academic Press, San Diego, CA, 1990), ISBN: 978-0-08-092534-9.

[36] Keras documentation, https://keras.io (retrieved August 11, 2022).

[37] D. P. Kingma and J. Ba, Adam: A method for stochastic optimization, arXiv:1412.6980v9.