Title

Virtual-diagnostic-based time stamping for ultrafast electron diffraction

Journal

Physical Review Accelerators and Beams, 26(5)

ISSN

1098-4402

Authors

Cropp, F

Moos, L

Scheinker, A

et al.

Publication Date

2023-05-01

DOI

10.1103/physrevaccelbeams.26.052801

Virtual-diagnostic-based time stamping for ultrafast electron diffraction

F. Cropp,[1,2,*] L. Moos,[3] A. Scheinker,[4] A. Gilardi,[2] D. Wang,[2] S. Paiagua,[2] C. Serrano,[2] P. Musumeci,[1] and D. Filippetto[2,†]

[1]*Department of Physics and Astronomy, UCLA, Los Angeles, California 90095, USA*
[2]*Lawrence Berkeley National Laboratory, Berkeley, California 94720, USA*
[3]*Special Circumstances, 113 Cherry Street 94153, Seattle, Washington 98104, USA*
[4]*Los Alamos National Laboratory, Los Alamos, New Mexico 87544, USA*

(Received 25 January 2023; accepted 27 March 2023; published 3 May 2023)

In this work, nondestructive virtual diagnostics are applied to retrieve the electron beam time of arrival and energy in a relativistic ultrafast electron diffraction (UED) beamline using independently measured machine parameters. This technique has the potential to improve the temporal resolution of pump and probe UED scans. Fluctuations in time of arrival have multiple components, including a shot-to-shot jitter and a long-term drift which can be separately addressed by closed loop feedback systems. A linear-regression-based model is used to fit the beam energy and time of arrival and is shown to be able to predict accurate behavior for both long- and short-time scales. More advanced time-series analysis based on machine learning techniques can be applied to improve this prediction further.

DOI: 10.1103/PhysRevAccelBeams.26.052801

I. INTRODUCTION

A recent trend in accelerator and beam physics has been the use of virtual diagnostics to measure indirectly one or more beam parameters using larger sets of upstream, nondestructive measurements of accelerator and machine parameters, which are correlated with the downstream beam properties. Various mathematical tools ranging from linear and nonlinear interpolations to more complex machine-learning-based techniques can be used to create high fidelity predictive models from training data obtained by destructive beam measurements. Then the model can be used to retrieve the beam parameter of interest once the destructive measurements cease. This is critically advantageous when measurements of the given parameter are particularly time consuming or require running in a particular working point on the beamline which is not compatible with the end-user application (e.g., [1–6]).

In particular, in systems where beam fluctuations strongly affect the accelerator performances, a very attractive opportunity exists to take advantage of virtual diagnostics models to improve the reliability in delivering a known set of beam parameters to an application even in presence of active feedback systems. This is because while feedback systems can be used to monitor machine parameters and keep them close to a given working point, these loops are not perfect (i.e., still allow a residual amount of jitter) and a complete compensation requires a beam-based diagnostic. In addition, the monitoring is typically limited to a single variable and the control algorithm does not take into account cross-correlation terms with other machine parameters. A global smart control system taking advantage of powerful and reliable virtual diagnostics models has therefore the potential to outperform such local feedback loops.

As an example, temporal stability is particularly important in pump-probe ultrafast techniques such as Ultrafast Electron Diffraction (UED). In a UED experiment, temporal resolution is defined as

$$\tau = \sqrt{\Delta t_{e^-}^2 + \Delta t_{\text{laser}}^2 + \Delta t_{\text{jitter}}^2 + \Delta t_{VM}^2}, \tag{1}$$

where Δt_{e^-} is the electron bunch length, and Δt_{laser} is the laser pulse length. These quantities can be reduced using a bunching cavity and laser compressor, respectively. Δt_{VM} is the velocity mismatch term, which can be neglected for ultrarelativistic beams and thin samples. That leaves the limiting factor of Δt_{jitter}, the time-of-arrival jitter between the laser pulse and electron bunch. Time stamping has been proposed in the past (e.g., [7,8]) to sort the UED patterns and retrieve the actual temporal trace of an ultrafast process. Nevertheless, depending on the particular implementation, accurate time stamping strongly constrains the machine setup (charge, crystal proximity, THz deflector) which might not be fully compatible with high-quality

*ericcropp@physics.ucla.edu
†dfilippetto@lbl.gov

diffraction patterns. Taking advantage of a virtual diagnostic would greatly increase the range of applicability of time stamping in UED, potentially improving the temporal resolution of the technique.

One of the beam parameters most strongly connected to the time of arrival of the beam at the sample is the beam energy. In linear transport theory, the connection is mathematically represented by the matrix element R_{56} which connects the final time of arrival with the relative energy deviation $\Delta E/E$ from the reference particle. For example, in a drift, higher energy particles arrive sooner. With more complex arrangements of beamline elements, which include buncher cavity and bending dipoles, the relation can become more complex (see [9]), as discussed in Sec. III B. Still, the beam energy is often the dominant contribution to the particle time of arrival at a given plane in the beamline and a kinetic energy virtual diagnostic could be useful to refine the predictions of the relative time-of-arrival fluctuations in a UED setup [10]. We also note here that there are other cases where nondestructive measurements of the beam energy (which otherwise requires bending the beam in a dipole spectrometer) would greatly improve accelerator performances. For example, in multishot measurements of transverse phase spaces, such as a quadrupole or solenoid scan emittance measurements [11,12], energy fluctuations change the focusing strength of the magnets, which would be considered to be constant for such a scan; poor energy stability is catastrophic to such a measurement. Even single-shot emittance measurement techniques, such as [13], require knowing the beam energy.

In this paper, we develop nondestructive virtual diagnostics for the beam time of arrival (TOA) and kinetic energy, which take into account nondestructive machine parameters measured upstream. Time stamping techniques in UED pose unique challenges, requiring single-shot TOA measurements on very low charge beams with very high resolution (<100 fs), which has only been achieved via destructive measurements [14]. Our work shows that the use of advanced mathematical methods can help break the paradigm of measurement accuracy versus beam charge. Indeed, we show that the electron beam parameters can be inferred from the accelerator context, i.e., measurable instantaneous machine parameters, with the same level of precision obtained by performing destructive measurements. Thus, the temporal resolution becomes independent from the beam charge and only dependent on the precision of the measurement of machine parameters.

The experiments were carried out at the LBNL HiRES beamline for UED, where we were able to reconstruct the beam energy and TOA for each shot with an accuracy beating our feedback systems using a simple linear interpolation virtual diagnostics model. By applying machine learning (ML) forecasting techniques, the reliability of the prediction further improves. The application of ML has been shown to solve or mitigate a plethora of accelerator control and diagnostic problems, for example, for navigating efficiently the multidimensional parameter space to find control set points [15,16], for inverting a large parameter space to make a parasitic diagnostic [17,18] or for nondestructive virtual diagnostics [1,2]. ML has also been combined with model-independent adaptive feedback for automatic control of the longitudinal phase space of the electron beam in the LCLS [19]. Further, UED has benefited from ML-based static models and virtual diagnostics [20,21]. The application of ML to forecasting for accelerators is a burgeoning effort that shows unique promise because of the time series structure of measured beam data [22].

In the next section, we discuss the operations at HiRES and the measurement systems for the parameters that are used in establishing the virtual diagnostics. The results of a linear-regression-based virtual diagnostic are shown in two cases where beam TOA and beam energy are used to train the model and benchmark its fidelity. In the last section of the paper, we compare the application of more complex ML models to the linear regression model.

II. SYNCHRONOUS DATA ACQUISITION AND ANALYSIS AT HIRES

A. Data Acquisition at HiRES

The HiRES accelerator includes a continuous-wave-class, normal-conducting electron photogun working at 185.7 MHz [23] (RF1 in Fig. 1) and a subsequent bunching cavity (RF2) operating at the seventh harmonic of the gun, i.e., 1.3 GHz. The present maximum electron beam repetition rate is fixed by the photocathode laser to 250 kHz, while an acousto-optic deflector at the end of the optical amplification chain can select user-defined patterns and/or lower the repetition rate. The nominal beam energy is 750 keV and all measurements in the paper were taken with an approximate beam charge of 15 fC.

Referring to Fig. 1, a dipole magnet (D1) downstream of the gun (RF1) and rf buncher cavity (RF2) is used to select between two beamlines, each providing access to a series of diagnostic tools. In particular, a deflecting cavity along the straight line (RF3) provides accurate pulse length and time-of-arrival information, with a calibration of 23.37 fs/pixel at the downstream imaging screen (VS3). The side beamline (UED beamline), branches off at an angle of 60° with respect to the straight line, resulting in high dispersion at the imaging screen VS2, and enabling high resolution energy measurements. The energy calibration $\Delta E/E$ at the screen is 2.5×10^{-5}/pixel and can be increased or decreased using the quadrupole triplet just upstream VS2 (Q1). In the measurements presented in this paper, the calibration for $\Delta E/E$ was 1.7×10^{-5}/pixel.

Owing to its unique set of beam parameters and its flexibility, the HiRES has been used for both UED applications [24,25] and for developing new technologies for compact and large-scale user facilities [26–28].

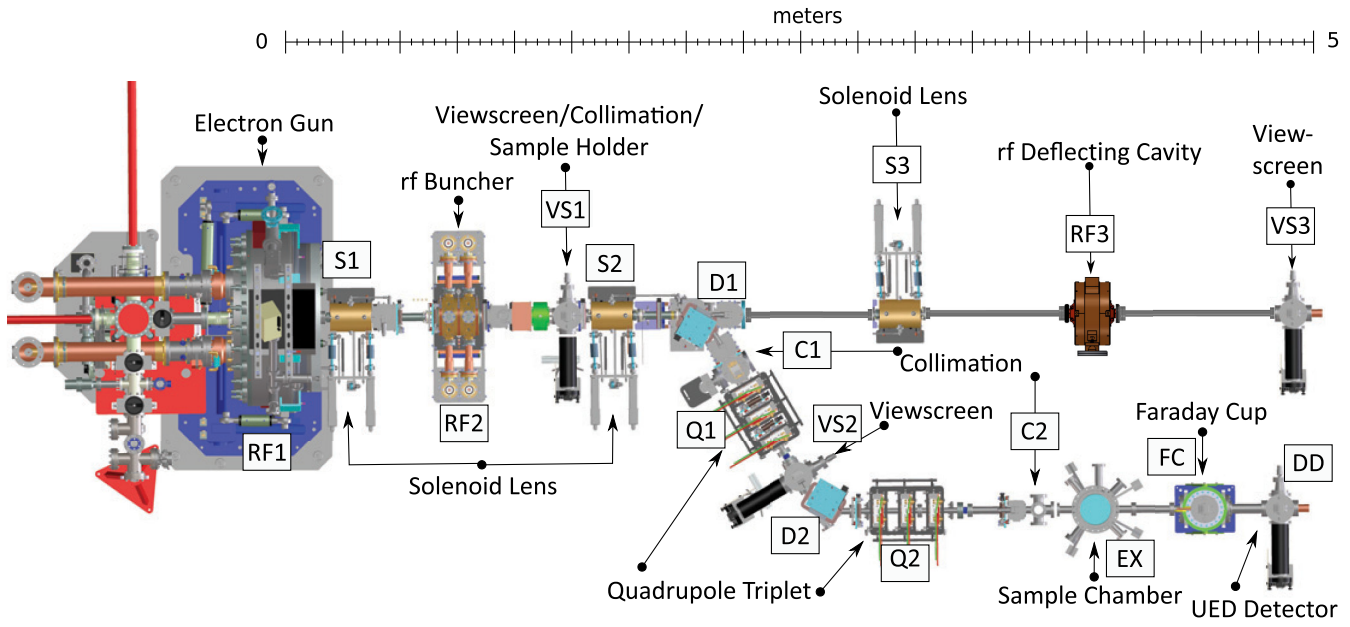

FIG. 1. The HiRES beamline. The UED beamline starts at D1 and goes through the dogleg to DD, while the diagnostic beamline goes straight from D1 to VS3. Adapted with permission from [9].

For example, the low-level-rf control electronics (LLRF), one of the most critical subsystems for ensuring electron beam energy stability, has been developed at LBNL and then deployed at the LCLS-II accelerator at SLAC [29]. The system allows precision control and measurement of amplitude and phase, with minimal cross-talk (more the 100 dB isolation in the upgraded version) and white noise background below 150 dBc/Hz. Such development is a key component for developing high precision feedback loop controls and for high fidelity prediction of beam parameters.

The aim of this work is to develop a novel virtual diagnostic tool for online high-precision time and energy stamping. The tool has been tested for single-shot beam predictions using information collected passively during beam runs and, while in these first tests, the acquisition was limited to 1 Hz, minor modifications to the timing system would allow much faster repetition rates, in the kHz range and beyond.

The development of a virtual diagnostic starts with building a model of the system, correlating measurements of beam parameters with machine parameters. Measuring the beam energy or time of arrival requires intercepting the electron beam with scintillator screens and analyzing the resulting images for the beam centroid after bending through a dipole or a time-dependent kick from a transverse deflecting cavity (TCAV).

The signal-to-noise ratio (SNR) of the training datasets is of particular importance as the model will be trained on the processed variables extracted from the images, and any error in the calculation for the parameter corresponds to an effective loss of information. In order to boost image SNR, it is possible to integrate multiple electron beam pulses (because of the high repetition rate of the system, each only 4 μs apart), so long as the timescale of system changes is longer than the averaging period.

At fixed rf power, the rate of change of the phase or amplitude of an electromagnetic field in a resonant cavity is limited by the cavity bandwidth. Indeed, the latter acts as a filter for external disturbances, so that every noise component outside its bandwidth is strongly attenuated. In the case of our 186 MHz cw-rf gun, with a quality factor Q greater than 10^4, we can estimate the timescale of field fluctuations:

$$\tau_{\text{noise}} = \frac{1}{\Delta f} = \frac{Q}{f} > 50\ \mu\text{s} \tag{2}$$

where f is the resonance frequency of the cavity. Further, there is an intrapulse proportional-integral-derivative (PID)-type feedback system engaged, which should further reduce jitter. The engagement of the intrapulse feedback can be seen after approximately 250 μs in the rf traces in Fig. 2.

For the rf bunching cavity and deflecting cavity, rms fluctuations of the amplitude and phase of the rf in both cavities show only a minimal increase when integrated for 40 μs relative to the case when integrated for 4 μs (the inherent uncertainty in laser shot time of arrival). The slight increase in fluctuations in each cavity is expected to increase the uncertainty in beam time of arrival at the final screen by less than 10 fs each. Therefore, most of the data produced in this work have been collected by averaging ten

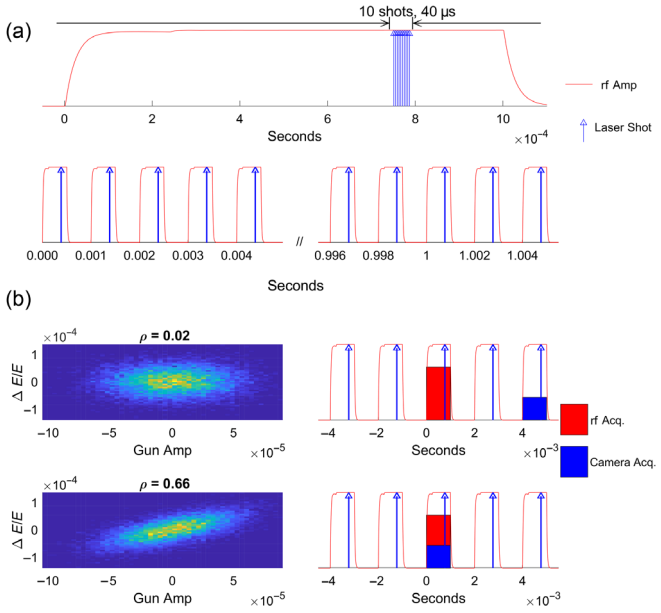

FIG. 2. (a) Synchronization scheme: rf amplifier run at 50% duty cycle with a 2-ms period. Up to ten laser shots arrive at 4 μs intervals for a short period toward the end of each rf pulse, where the rf is generally most stable. (b) Correlation plots of rf gun amplitude and relative energy deviation (measured at the dipole spectrometer) for synchronized and temporally misaligned acquisition schemes.

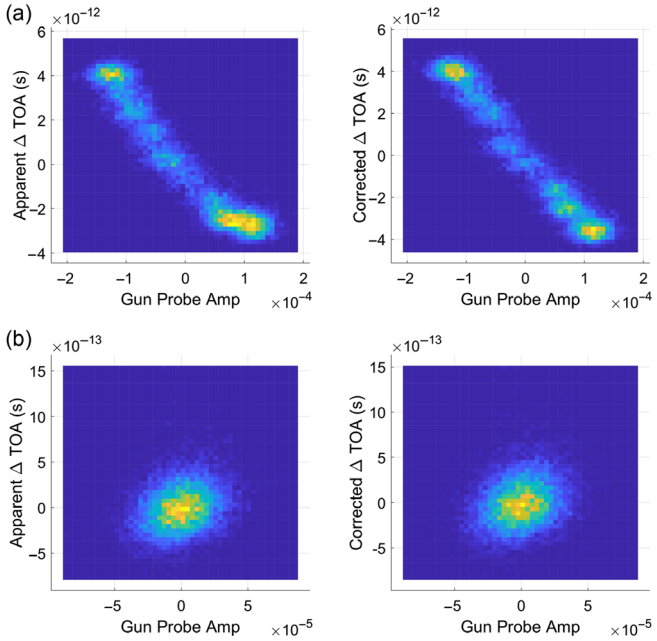

FIG. 3. Synchronous measurements of beam TOA at TCAV: correlations of TOA with electron gun amplitude (without the use of the rf bunching cavity, similar to data in Fig. 2). Left: TCAV rf jitters are not taken into account in postprocessing. Right: fluctuations in TCAV rf amplitude and phase are used to correct beam time-of-arrival measurements. In (a), long-term drifts are uncompensated. In (b), a moving average is subtracted to show only short timescale jitters. Note the much smaller y scale in the bottom plots.

beams per image, in order to increase the SNR in the images (see Fig. 2).

In order to obtain the most accurate model and predictions, the heterogeneous data acquired (a mix of images and waveforms) require deterministic time alignment with a precision equal to or better than τ_{noise}. Figure 2 describes our timing setup. The electron gun is used in pulsed mode for these experiments, with a total duration of the rf pulse of 1 ms and a repetition rate of 500 Hz (corresponding to a duty cycle of 50%). The optical gate sending the burst of ten consecutive laser pulses can be activated at any time along the rf pulse, with a precision of 4 μs. In Fig. 2(b), we show the correlation plots of electron beam relative energy deviation and the amplitude of the field in the electron gun, in the simple case where no other cavity is used. The ρ coefficient on top of each plot corresponds to the value of the correlation function between the two. The data are shown in the case of synchronized and not synchronized acquisition, providing a clear idea of the information lost without precise time alignment.

The beam position on the screen measured during the initial characterization of the accelerator is not only determined by the variable we are interested in predicting. Beam position can change because of a magnet current change or because of laser pointing fluctuations on the cathode. Therefore a complete model should calculate correlations with all the relevant parameters of the accelerator. Photocathode laser beam images are saved with μs-scale alignment precision synchronously with electron beam images, but fortunately, not all data require the same level of time alignment. Variations in parameters such as cavity temperatures, water flows, and magnet currents mostly contribute to machine drifts, and only require synchrony at the subsecond level, which can be achieved via software. At HiRES, continuous data storing of user-requested machine settings is performed automatically by an online database with 10-Hz periodicity, providing the necessary information to include all machine parameters in the model.

Data for the deflecting cavity deserve a special discussion. As mentioned above, data on this cavity should in principle be taken synchronously, as small short-term fluctuations are expected. In Fig. 3, the correlation plots between the electron gun amplitude and the measured beam time of arrival before and after the compensation of TCAV short-term fluctuations are shown. Given the small contributions of these jitters to the measured short-term fluctuations, we acquire TCAV data via the database and therefore account only for long-term drifts of the field in the cavity.

B. Data analysis and prediction

In the following section, virtual diagnostics based on multivariable linear regression are presented. Multiple linear regression is a statistical modeling technique where

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{3}$$

where $\mathbf{Y}$ is a vector of the observed quantities, $\mathbf{X}$ is a matrix of dimension number of observations by number of predictors, $\boldsymbol{\beta}$ is a vector of regression coefficients (dimension: number of predictors), and $\boldsymbol{\epsilon}$ is the error term, a vector of individual errors on each observation. Estimates of regression coefficients, $\hat{\boldsymbol{\beta}}$, are learned by minimizing residuals.

Multiple linear regression is a powerful tool because of its explainable nature. Rather than producing a black-box model, which produces predictions through an opaque process, multiple linear regression produces an interpretable and explainable model. As it will be shown below, the regression coefficients are learned and can be analyzed in order to characterize quantitatively the impact that each predictor has on the overall prediction.

It is also important to observe that linear regression between two variables is agnostic to the time relation between different data points as it is inherently a time-independent method. However, the datasets in this work are all time-series datasets. While linear regression is effective in quantifying the consistent effect that the predictors have on the observation, the method will fail to identify the time-dependent noise processes that perturb the system and affect both the predictors and the observations.

In this case, two adaptations to the usual prescription for linear regression were made: (i) the data were not randomized, in order to preserve the time-series ordering and (ii) the model is trained on the first part of the data, while the last part of the data is reserved for validation. The linear regression results below serve as a practical baseline result that—by itself—shows promise for improving stability, but as shown in the last section of the paper could be further improved upon by taking into account temporal evolution using more complex models.

III. ONLINE PREDICTIONS OF ELECTRON BEAM PARAMETERS

A. Time stamping

The temporal resolution in UED experiments [see Eq. (1)] is often dominated by the relative time-of-arrival fluctuations between excitation laser and probing electron beam. Therefore an online diagnostic capable of precise nondestructive measurement of TOA would have a profound impact on the overall instrument performance. For the data presented in this section, the accelerator setup matched the beam and machine parameters used during UED experiments. As such, the rf bunching cavity (RF2) is set for temporal compression, with nominal field amplitude and zero-degree injection phase (the so-called *zero-crossing* phase). The fields in both the electron gun and the bunching cavity are stabilized in amplitude and in phase by fast, FPGA-based PID-type feedback loops.

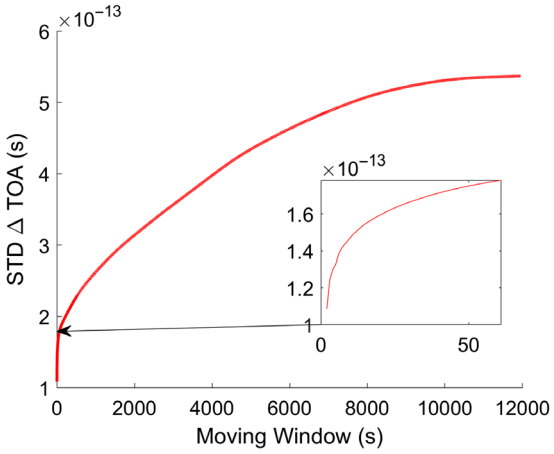

FIG. 4. Standard deviation of transverse beam centroid calibrated to TOA relative to the reference beam as a function of the width of the acquisition time window. Inset shows a short timescale.

In Fig. 4, we show the standard deviation of the time-of-arrival of the beam at the deflecting cavity (measured by converting the centroid variation of the beam on the screen using the pixel-to-time deflector calibration), as a function of the temporal duration for the data acquisition. This quantity continuously increases due to short- and long-term drift. In the inset, zooming in on the 1-min time scale, it is shown how short-term drifts account for less than 200 fs of temporal jitter. On the other hand, with an increase in the temporal width of the acquisition window, the overall stability of the system is observed to degrade at longer timescales. Depending on the duration of the intervals in between re-establishing a new time-zero position in UED pump probe scans [14], the integrated resolution can become as large as 600 fs.

Figure 5 shows the evolution of the beam TOA at the TCAV over about 3 h. The data are divided into two sections (highlighted by the vertical dashed line), with 75% of the points used for developing a model of the system (the training data set), and the last 25% is used to validate it. A further test set is not required to test generalization, as no hyperparameter tuning was required.

We then use the linear regression model described in Sec. II B for prediction. Inputs to the model include rf amplitude and phase of the electron gun, the bunching cavity and the deflecting cavity, the photocathode laser arrival time at the cathode with respect to the rf wave, and an image of its transverse shape, intensity, and position at the cathode.

The red line in Fig. 5 shows the result of the regression. The model is able to learn the correlations in the data and predict to a high degree of accuracy, decreasing the uncertainty in the long-term data (RMSE) by more than a factor of two in the test set. This represents a major improvement, one that reduces the uncertainty from the hours-long timescale in Fig. 4 to the minutes-long stability.

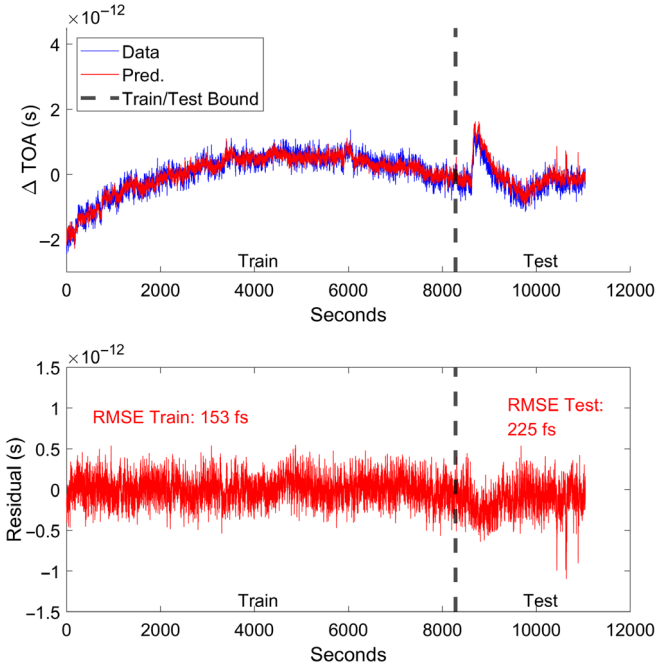

FIG. 5. Time-of-arrival fluctuation measured using the TCAV screen while the PID-type feedback was engaged. Residual drifts not corrected by the feedback are present. The linear regression is also shown. The uncertainty due to long-term drifts is reduced to the 200 fs level similar to the shot-to-shot, short-term jitter shown in Fig. 4.

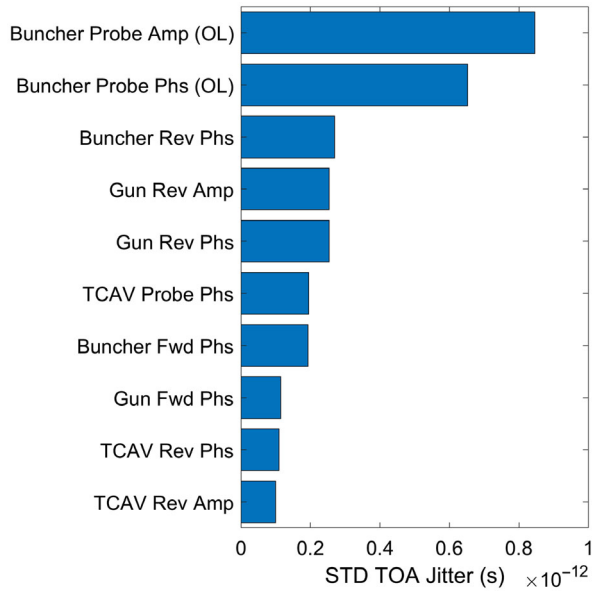

FIG. 6. Top 10 model predictors and the associated TOA movement with 1 STD movement in the validation set. Details of the predictors can be found in Table I in Appendix A.

Notably, the model was able to predict accurately the outcome of a sudden large phase shift in the TCAV in the test dataset (visible at around 8600 seconds). Such a jump is mostly due to a sudden variation in the settings of our diagnostic device, and not to an actual change in beam TOA. By tracking the parameters of the measurement system, i.e., TCAV rf amplitude and phase, real temporal shifts can be isolated from simple beam centroid fluctuations due to variations of the fields in the TCAV. Indeed, once the system correlations have been learned from the training dataset, all the coefficients that would contribute to a beam movement on the screen but not necessarily to a change in TOA can be removed.

Although the improvements made so far are impressive, they do not fully take into account the complexity of a drifting, time-series dataset. The discrepancy arises because of the distinction between the root mean square error (RMSE) of the linear regression prediction from the actual measurement, and the standard deviation (STD), which measures the deviation from the mean value. It is essential to carefully choose the reference point for assessing the virtual diagnostic's best performance. With traditional feedback, minimizing the system's drift is the typical approach until it becomes negligible, and the uncertainty is then determined by the STD. Essentially, when destructive measurements are turned off, the beam is presumed to have the same properties as the last-measured beam. Thus, in a drifting system, the uncertainty in the TOA would increase with time (due to the variation in the mean value). However, by utilizing the linear regression virtual diagnostic, it is possible to compensate for these drifts, and the uncertainty remains relatively constant over time, with little to no degradation.

As described in Sec. II B, the impact of each predictor on the overall time of arrival estimate can be extracted from the model. In Fig. 6, the top predictors' impacts [1 standard deviation of variation converted to TOA prediction using the corresponding β from Eq. (3)] are reported. This can be helpful for several reasons, including (i) to see if conventional feedback systems can be better tuned and (ii) to see if the perceived TOA variation is due to measurement uncertainty (i.e., the TCAV measuring the TOA is jittering) or if the TOA is actually moving. Although the effect of the TCAV is significant, the dominating contribution is that of the buncher, meaning that the TOA is actually moving, despite conventional feedback systems. It also suggests that these conventional feedback systems could be improved for the buncher cavity. A full list of parameters included in the model is shown in Appendix A.

These results show how the combination of a linear-regression-based model and time-aligned data can help enhance the performance of traditional feedback systems.

B. Energy stamping

A similar approach can be used to obtain very accurate predictions of the electron beam energy. To showcase this capability, we make use of separate beamline settings. In particular, the electron beam is transported into the UED line and measured at the VS2 screen (Fig. 1) after

acceleration by the electron gun, while the rf bunching cavity (RF2) is left off for simplicity of interpretation. We acquired 3 h of data for the two different cases of stabilized and unstabilized accelerating fields in the gun (using the active LLRF PID stabilization loop mentioned earlier).

Results of energy stability measurements in the two cases are shown in the histogram of Fig. 7(a). The effect of the fast feedback in stabilizing the energy is evident, with rms relative energy stability going from approximately 10^{-3} to 2×10^{-4} over a 3-h run. Nevertheless, a clear structure is evident in the stabilized case, with a double peaked distribution of unknown cause, suggesting even better performance may be achieved.

The application of our linear regression model to both scenarios results in the histograms of Fig. 7(b). Here we plot the residual error left after comparing the model predictions with the measured ground truth. We can make a few observations: first, the application of the model increases the precision with which we can assert the energy of each electron beam, by a tenfold factor for feedback-off, and by a small factor in the feedback-on case. Second, the residual error distribution is now much closer to a Gaussian, to be expected when only random noise is left, and nothing else can be learned from the system. This is therefore the first indication that our model is close to optimal. Third, the final RMSE is similar in the stabilized and unstabilized cases. This last point is quite interesting, as it shows that virtual diagnostic tools have comparable performances with respect to traditional feedback systems. For applications where time of arrival may vary, but in a well-controlled fashion, such as pump-probe experiments, this suggests that in the future, software-based approaches could outperform hardware-based approaches in the stabilization and control of particle beams. In the future, considering the likely increase in computing power, network speed, and bandwidth, feedback could even be based on virtual diagnostics.

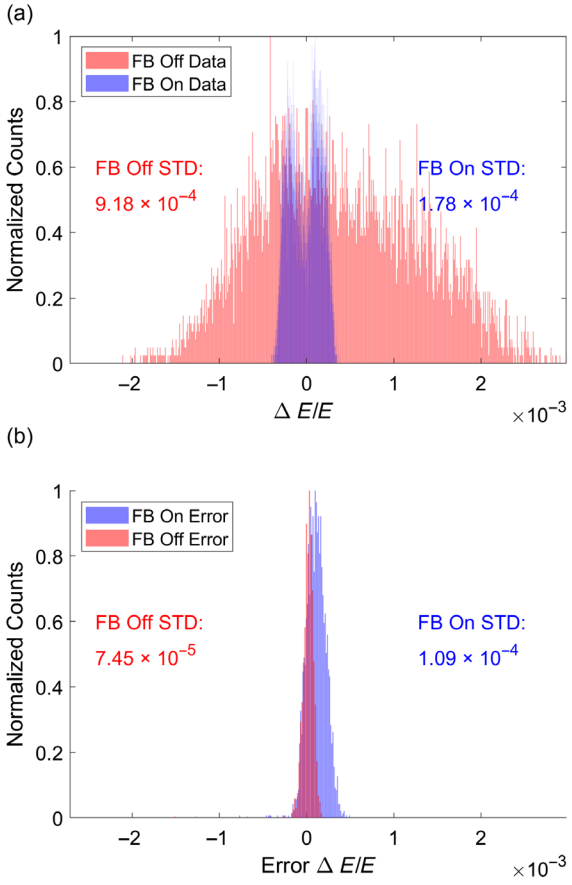

FIG. 7. Linear regression predictions with and without traditional PID-type feedback (FB) engaged. (a) The variation from the mean of the training data is shown, after conversion to relative energy deviation. (b) the validation errors are shown. Note that the FB off case shows greater improvement than the FB on case on both an absolute and a relative scale.

Using a matrix formalism for the linear transport in longitudinal phase space, one can compare the virtual diagnostics results for the energy and time-of-arrival measurements. At the TCAV screen,

$$\Delta t = \left(\frac{R_{56,\mathrm{gb}}}{h \cdot R_{56,\mathrm{gb}} + 1} + R_{56,\mathrm{bs}} \right) \frac{\Delta E}{E}, \tag{4}$$

where $R_{56,\mathrm{gb}}$ and $R_{56,\mathrm{bs}}$ are related to the drift distances from the gun to the buncher and from the buncher to screen, respectively. These R_{56} elements are given in time-energy coordinates for convenience. For HiRES, these can be calculated to be 0.73 and 3.20 ns, respectively. h is the R_{65} term associated with a thin lens description of the buncher cavity. For HiRES, calibration measurements indicate this to be $(-2.02\ \mathrm{ns})^{-1}$. This can be expressed in engineering units to be approximately

$$\Delta t\,[\mathrm{ns}] = 4.3415 \frac{\Delta E}{E}, \tag{5}$$

where Δt is given in nanoseconds. Using an uncertainty from the energy stamping virtual diagnostic of about 7.45×10^{-5} (see the feedback off case in Fig. 7), we find an uncertainty of 227 fs, when considering the limitations discussed below, which is comparable to the 225-fs uncertainty of the time-stamping virtual diagnostic.

C. Limitations of the method

There are many possible limitations that leave room to increase the precision of this method beyond what is achieved in this work. To start, the model used up to now is linear, so any presence of small nonlinearity in the system will not be captured. Also, the model may change with time, and therefore some sort of adaptive tuning could be exploited (see [27,30]). Maybe the most important of all limitations is the accuracy in the measurement of key parameters and the level of noise in the measured ground

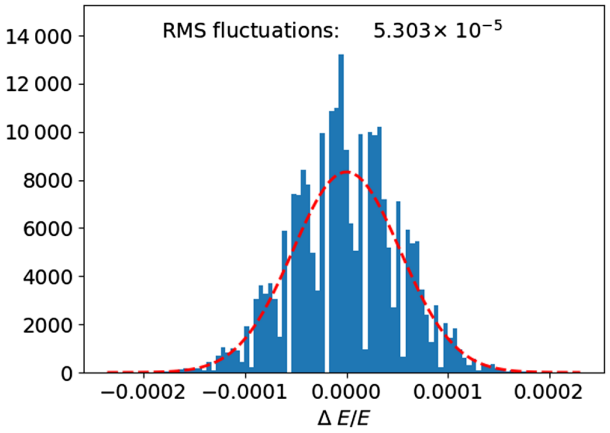

FIG. 8. Measured current fluctuations extrapolated to perceived relative energy fluctuations for a 750-keV electron beam, as in the experiment.

truth used for training the model. In this work, we are aiming at final predictions with accuracy at the 10^{-5} level, which requires more than 100 dB SNR in measurements of radiofrequency signals. Potentially more dangerous is the requirement on the currents energizing the different magnets in the beamline. Measuring 10^{-5} variations on this current requires specialized hardware that is usually not available for each magnet of the accelerator. Therefore, we perform an experimental sensitivity study and verify the dipole D1 as the one with the highest impact on the beam position on the screen. The current fluctuations driving the magnetic dipole were then measured with high precision by a specialized setup. A Danisense DS50ID ultrastable fluxgate current transducer with a 16-bit digitizer was set up to measure the current provided to the dipole from the CAEN A3620 power supply when set to a nominal value corresponding to a 750-keV electron beam. A 66-h long measurement of the current was taken and showed fluctuations on the high 10^{-5} level, corresponding to apparent relative energy functions on the 5×10^{-5} level. See Fig. 8 for more details. The rms fluctuations found during this test, although not contextual with beam measurements, are of the same scale as the residual error we obtained in both cases of Fig. 7, showing that direct synchronous measurement of the current in the dipole magnet could increase the precision of the virtual tool.

IV. ADVANCED PREDICTIONS

While the above methods are effective, the regression approach presented above relies on the assumption that TOA or energy can be extracted from linear correlations with the predictors. In this section, an approach is shown that leverages advancements in forecasting to use a temporal fusion transformer (TFT) architecture to reduce temporal correlation in the residuals and further improve the prediction performance of the model.

A. Autoregressive models and TFTs

Autoregressive models are a class of models used to represent sequential data that include recurrent neural networks (RNNs), long-term and short-term memory networks (LSTMs), and transformers. RNNs estimate the probability $p(y_{t+1}|h_t, y_t)$ with a prior consisting either of weights generated from an initialization strategy or previously computed hidden state h_t [31]. In practice, representing nonmonotonic and complex relationships in sequences requires improving these models by stacking multiple neurons or using bidirectional LSTMs [32–35]. While still widely used, there are also problems with LSTMs, namely catastrophic forgetting, where successive updates of the memory cell with new data cause the network weights to forget historical data [36,37]. Catastrophic forgetting affects the network's ability to be pretrained on large datasets prior to fine-tuning for a specific task [38].

Transformers are a member of the class of autoregressive models that build upon models such as those described above. Transformers use submodules consisting of stacked LSTM neurons as well as novel components such as self-attention in order to train on many datasets or large sequences. While the transformer architecture was originally used in the field of natural language processing, it has subsequently been expanded to a variety of other modalities including time-series forecasting and visual processing [39,40]. Transformers consist of an encoder that maps feature vector $\{x_{1,t}, x_{2,t}....x_{n,t}\}$ to a continuous representation $\{z_{1,t}, z_{2,t}....z_{n,t}\}$, which is then used by a decoder to generate a sequence of m predictions $\{y_{1,t}, y_{2,t}....y_{m,t}\}$.

TFTs, described in [41], are a transformer architecture and the current state of the art for multihorizon forecasting. At each time step t, a context window of length k consisting of past predictors, along with the past values of the ground truth, is sent to the encoder. Known machine parameters at the prediction time steps (or horizons) are also encoded. TFTs contain multiple variable selection networks that reduce network complexity through the modulation of the probability that a given predictor's signal propagates to deeper layers of the network. This reduces the need for data preprocessing, the impact of noisy variables, and prevents overfitting. The TFT decoder includes a multiheaded attention module for long-term temporal pattern recognition. Through multiheaded attention, weights are learned that reflect the degree to which encoded variables attend or correlate with one another [42]. These weights are then passed from the decoder to a dense layer that produces the quantile predictions.

In summary, the TFT offers an explainable model that predicts a distribution of results instead of a point prediction. TFTs extract variable importance using both attention and automated variable selection [41]. Quantile predictions allow confidence of a prediction to be assessed, which is useful for human evaluation and downstream control tasks.

B. Method

In offline forecasting tasks, access to past ground-truth data allows for the context window to be populated with observations of the target as shown in Eq. (6). A prediction at time step t is given by

$$\hat{Y}_t = f(\{\mathbf{X}_{t-k}, \ldots, \mathbf{X}_t\}, \{Y_{t-k}, \ldots, Y_{t-1}\}), \quad (6)$$

where $\mathbf{X}$ at each time step is a vector of predictors, such as machine parameters. Y at any given time step is the ground-truth data, in this case, the beam TOA, where $\hat{Y}$ is the prediction of this quantity. k is the context window, or history, that the model is given.

In the context of the HiRES virtual diagnostic, the lack of ground-truth data during deployment about the beam TOA has to be negotiated once destructive measurements cease. In such an online or a multihorizon task, one approach would be to introduce previous predictions recursively as shown in Eq. (7).

$$\hat{Y}_t = f(\{\mathbf{X}_{t-k}, \ldots, \mathbf{X}_t\}, \{\hat{Y}_{t-k}, \ldots, \hat{Y}_{t-1}\}) \quad (7)$$

In this approach, residual bias introduced in the model's estimates would accumulate with each recursive prediction, and over thousands of timesteps would become significant. Even without a significant increase in error, error would be time correlated rather than being normally distributed. In order to avoid this problem, once destructive measurements cease, previous ground-truth measurements must be replaced with time-independent predictions as shown in Eq. (8).

$$\hat{Y}_t = f(\{\mathbf{X}_{t-k}, \ldots, \mathbf{X}_t\}, \{g(\mathbf{X}_{t-k}), \ldots, g(\mathbf{X}_{t-1})\}) \quad (8)$$

where $g(x)$ is any time-independent model. In the work presented herein, the linear regression model as shown in Sec. III A, as described in Eq. (3), is used to replace ground-truth data, after training with access to the ground-truth data, as shown in Eq. (6). Thus, during training, the TFT has access to a context window of long-term trend information in $\mathbf{X}$ and $\mathbf{Y}$ in order to learn from a more complete view of the system's dynamics. Following training, during online application, we utilize accurate estimates of $\mathbf{Y}$ as provided by the linear-regression-based estimates $g(\mathbf{X}_t)$ and therefore the approach does not suffer from catastrophic degradation of the predictions after destructive measurements are no longer available.

The results in this section make use of a TFT implemented in PyTorch Lightning [43] with the PyTorch forecasting package [44]. A 75/25 training split identical to that of the linear regression model in Sec. III was employed with one caveat: for both the training and validation sets, the first $\mathbf{k}$ instances (with k being the length of the context window) do not have corresponding predictions. The predictors were the same as those used for the linear regression model in Sec. III A. Finally, a robust normalizer was applied to the data, which scales and centers it with regard to but without transforming the target. Additional details of the TFT model can be found in Appendix B.

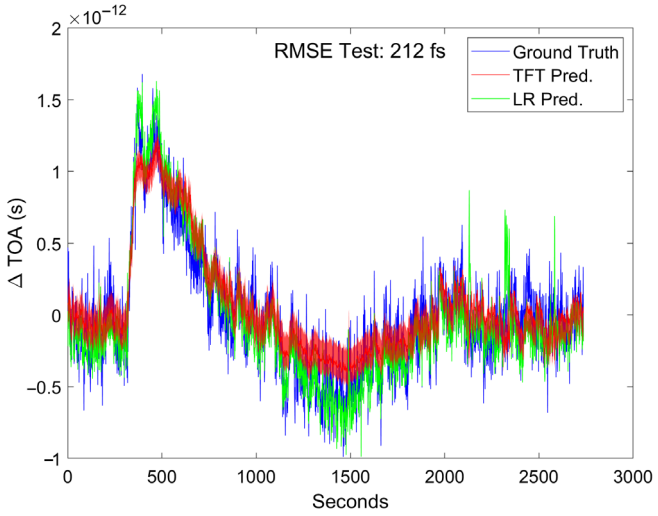

FIG. 9. TFT TOA median and interquartile range (shaded) predictions in the validation set. Predictions are compared with linear regression (LR) predictions and the ground truth.

C. Results and discussion

The results shown in Fig. 9 are the quantile predictions trained on ground-truth observations of the beam TOA. Note that the RMSE is approximately 6% better than that of linear regression, as the residuals of the model are in general closer to zero, as shown in Fig. 10.

This improvement can be explained by noting that despite state-of-the-art stability at HiRES, Fig. 10 demonstrates that the processes causing long-term TOA drift introduce correlations of error over time that could be reduced via the use of a forecasting model. As shown in Fig. 10, residuals of the TFT show little autocorrelation

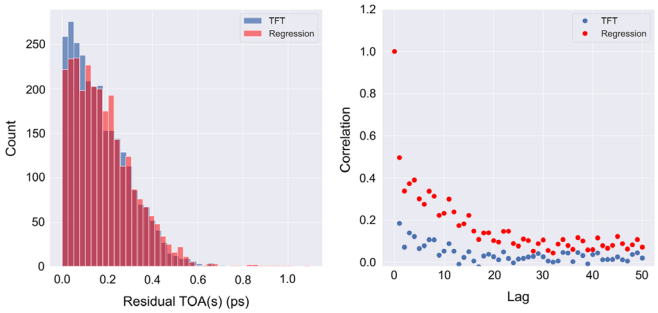

FIG. 10. Left: error histogram comparing the absolute value of the residuals from the linear regression and TFT models. Note that the error is bunched closer to zero for the TFT. Right: residual correlation for linear regression and TFT models. Note that the use of a TFT predictor reduces the correlation in the residuals between time steps.

between time steps relative to the linear regression results. This demonstrates that the approach of incorporating historical machine parameter information despite the challenges of online forecasting detailed above is a reasonable one.

It should be noted that because of the small number of observations, only a validation set was used in this work, and a test set was omitted. Although the validation set was used for hyperparameter optimization and as a metric for training, the generalization to a validation set is still notable, under these conditions. In the future, pretraining on data from different accelerator runs prior to fine-tuning on related data should be investigated to further reduce residual error and allow for increased observation of anomalous states in the machine parameters and their correlated effect on the target beam parameters. Every state-of-the-art use of the transformer model used in natural language processing since 2015 has relied on pretraining to increase performance [45] and the use of transformers for forecasting would likely benefit from similar methods. Greater diversity of observations from better exploration of the parameter space of predictors would lead to better generalization and better model performance. The inability of the predictions to capture the full variation of the ground-truth beam TOA exhibited in Fig. 10 could be explained by the fact that while relationships between machine parameters and beam TOA have been learned in training, anomalous transitions in machine parameters have not been observed before, forcing the model to extrapolate. While anomalous observations are, by definition, rare, a larger number of observations of similar transitions in other experiments would allow for greater predictive power during these periods. Training with more raw data and possibly with lagged data could help capture temporal relationships between changes in the variables.

Another way to improve accuracy with training data would be training with a greater diversity of examples prior to any aggregation. In addition to learning the between time-step error, learning the variance of the sensor readings within a time step and training with data taken with the rf cavities' PID controllers disabled would allow the model to learn a more robust embedding space.

V. CONCLUSION

In this work, a novel application of virtual diagnostics has been explored, toward enhancing UED temporal resolution by predicting electron beam TOA—or the main contributor to TOA in this energy regime, beam energy. Linear-regression-based models can be used to greatly reduce uncertainty in machine parameters. For energy stamping, linear-regression-based virtual diagnostics were shown to mitigate the long-term drift to a level comparable to what can be done with the PID feedback loops. For time stamping, linear-regression-based virtual diagnostics were shown to work in concert with traditional feedback to mitigate long-term drift and lower the uncertainty to 225 fs, which is on the same scale as the shot-to-shot fluctuations and a significant reduction from the uncompensated standard deviation uncertainty of 600 fs. Further, state-of-the-art forecasting models were applied to mitigate the temporal correlation of residuals of the model predictions, resulting in a nominal reduction in prediction uncertainty to 212 fs.

There are several ways to realize benefits from reducing the uncertainty in prediction error. For example, one could use a virtual diagnostic for feedback, in order to remove the long-term drift. Another method is to make use of the knowledge provided by the virtual diagnostic, without direct feedback. Working under the paradigm of "measurement is easier than control" has been shown to be effective (e.g., [17,18]) and has several advantages; rather than working to control further the natural parameter drift of the machine in an already state-of-the-art stability environment, the remaining drift and jitter can be harnessed to improve measurements. For example, in UED experiments, if for each shot, the virtual diagnostics showcased in this work are applied to retrieve the relative time of arrival within the shot-to-shot error, one would be able to reorder the data using the shot-tag information with a corresponding improvement in the temporal resolution as well as a significant reduction of acquisition times.

ACKNOWLEDGMENTS

This material is based partly upon work supported by the U.S. Department of Energy, Office of Science, Office of Workforce Development for Teachers and Scientists, Office of Science Graduate Student Research (SCGSR) program. The SCGSR program is administered by the Oak Ridge Institute for Science and Education for the DOE under Contract No. DE-SC0014664. This work was also partially supported by the DOE Office of Basic Energy Sciences under Contract No. DE-AC02-05CH11231, and by the DOE Office of Science, Office of High Energy Physics under Contract No. 89233218CNA000001. F. C. also acknowledges support from NSF PHY-1549132, Center for Bright Beams.

APPENDIX A: MODEL PARAMETERS

Parameters for the model in Sec. III A are listed in Table I. Parameters marked as "Async" were taken asynchronously, as described in Sec. II A. For the gun and buncher, PID-type feedback loops were engaged, based on some of these readings. If PID-type feedback was engaged based on one of the parameters, it is designated as "IL" or "in-loop." If an independent measurement of the parameter exists, it is denoted as "OL" or "out-of-loop."

TABLE I. List of predictors for TOA used in models described in Secs. III A and IV.

Parameter name	Description
Gun probe Amp (IL)	In-loop gun rf probe amplitude
Gun probe Phs (IL)	In-loop gun rf probe phase
Gun probe Amp (OL)	Out-of-loop gun rf probe amplitude
Gun probe Phs (OL)	Out-of-loop gun rf probe phase
Laser Phs	Phase difference between laser and rf
Laser crosstalk	"Amplitude" of the above signal—measures channel crosstalk
Gun Rev Amp	Amplitude of reverse power
Gun Rev Phs	Phase of above
Gun Fwd Amp	Amplitude of forward power
Gun Fwd Phs	Phase of above
Buncher probe Amp (OL)	Out-of-loop buncher rf probe amplitude
Buncher robe Phs (OL)	Phase of above
Laser position (x)	x-coordinate of the virtual cathode image centroid
Laser position (y)	y-coordinate of the virtual cathode image centroid
Buncher Rev Amp 2 (Async)	Amplitude of reverse power at buncher coupler 2
Buncher Rev Phs 2 (Async)	Phase of the above
Buncher Fwd Amp (Async)	Amplitude of buncher forward power
Buncher Fwd Phs (Async)	Phase of above
Buncher Rev Amp (Async)	Amplitude of buncher reverse power
Buncher Rev Phs (Async)	Phase of above
Buncher Probe Amp (IL-Async)	Amplitude of in-loop buncher probe
Buncher Probe Phs (IL-Async)	Phase of above
TCAV Rev Amp (Async)	Amplitude of TCAV reverse power
TCAV Rev Phs (Async)	Phase of above
TCAV Fwd Amp (Async)	Amplitude of TCAV forward power
TCAV Fwd Phs (Async)	Phase of above
TCAV Probe Amp (Async)	Amplitude of TCAV probe
TCAV Probe Phs (Async)	Phase of above

TABLE II. List of hyperparameters for the model shown in Sec. IV.

Hyperparameter	Value
Hidden size	89
Dropout	0.276
Attention head size	2
Learning rate	0.0012
Max encoder length	25

APPENDIX B: TFT HYPERPARAMETERS

The TFT architecture is outlined in detail in [41]. The hyperparameters for the TFT model are listed in the Table II.

[1] A. Scheinker, S. Gessner, C. Emma, and A. L. Edelen, Adaptive model tuning studies for non-invasive diagnostics and feedback control of plasma wakefield acceleration at FACET-II, Nucl. Instrum. Methods Phys. Res., Sect. A **967**, 163902 (2020).

[2] C. Emma, A. Edelen, M. 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).

[3] O. Convery, L. Smith, Y. Gal, and A. Hanuka, Uncertainty quantification for virtual diagnostic of particle accelerators, Phys. Rev. Accel. Beams **24**, 074602 (2021).

[4] V. Yakimenko, L. Alsberg, E. Bong, G. Bouchard, C. Clarke, C. Emma, S. Green, C. Hast, M. Hogan, J. Seabury *et al.*, FACET-II facility for advanced accelerator experimental tests, Phys. Rev. Accel. Beams **22**, 101301 (2019).

[5] A. Hanuka, C. Emma, T. Maxwell, A. S. Fisher, B. Jacobson, M. J. Hogan, and Z. Huang, Accurate and confident prediction of electron beam longitudinal properties using spectral virtual diagnostics, Sci. Rep. **11**, 2945 (2021).

[6] A. Scheinker and S. Gessner, Adaptive method for electron bunch profile prediction, Phys. Rev. ST Accel. Beams **18**, 102801 (2015).

[7] C. Scoby, P. Musumeci, J. Moody, and M. Gutierrez, Electro-optic sampling at 90 degree interaction geometry for time-of-arrival stamping of ultrafast relativistic electron diffraction, Phys. Rev. ST Accel. Beams **13**, 022801 (2010).

[8] M. Othman, A. Gabriel, M. Hoffmann, F. Ji, E. Nanni, X. Shen, E. Snively, and X. Wang, Terahertz driven compression and time-stamping technique for single-shot

ultrafast electron diffraction, in *Proceedings of Particle Accelerator Conference, IPAC'21, Campinas, SP, Brazil* (JACoW, Geneva, Switzerland, 2021), https://jacow.org/ipac2021/papers/mopab141.pdf.

[9] D. Filippetto and H. Qian, Design of a high-flux instrument for ultrafast electron diffraction and microscopy, J. Phys. B **49**, 104003 (2016).

[10] L. Zhao, Z. Wang, C. Lu, R. Wang, C. Hu, P. Wang, J. Qi, T. Jiang, S. Liu, Z. Ma *et al.*, Terahertz Streaking of Few-Femtosecond Relativistic Electron Beams, Phys. Rev. X **8**, 021061 (2018).

[11] M. G. Minty and F. Zimmermann, *Measurement and Control of Charged Particle Beams* (Springer, Berlin, 2003), https://cds.cern.ch/record/629879?ln=en.

[12] E. Prat and M. Aiba, Four-dimensional transverse beam matrix measurement using the multiple-quadrupole scan technique, Phys. Rev. ST Accel. Beams **17**, 052801 (2014).

[13] D. Marx, J. G. Navarro, D. Cesar, J. Maxson, B. Marchetti, R. Assmann, and P. Musumeci, Single-shot reconstruction of core 4D phase space of high-brightness electron beams using metal grids, Phys. Rev. Accel. Beams **21**, 102802 (2018).

[14] D. Filippetto, P. Musumeci, R. Li, B. J. Siwick, M. Otto, M. Centurion, and J. Nunes, Ultrafast electron diffraction: Visualizing dynamic states of matter, Rev. Mod. Phys. **94**, 045004 (2022).

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

[16] J. Duris, D. Kennedy, A. Hanuka, J. Shtalenkova, A. Edelen, P. Baxevanis, A. Egger, T. Cope, M. McIntire, S. Ermon *et al.*, Bayesian Optimization of a Free-Electron Laser, Phys. Rev. Lett. **124**, 124801 (2020).

[17] S. Li, F. Cropp, K. Kabra, T. Lane, G. Wetzstein, P. Musumeci, and D. Ratner, Electron Ghost Imaging, Phys. Rev. Lett. **121**, 114801 (2018).

[18] K. Kabra, S. Li, F. Cropp, T. J. Lane, P. Musumeci, and D. Ratner, Mapping photocathode quantum efficiency with ghost imaging, Phys. Rev. Accel. Beams **23**, 022803 (2020).

[19] A. Scheinker, A. Edelen, D. Bohler, C. Emma, and A. Lutman, Demonstration of Model-Independent Control of the Longitudinal Phase Space of Electron Beams in the Linac-Coherent Light Source with Femtosecond Resolution, Phys. Rev. Lett. **121**, 044801 (2018).

[20] Z. Zhang, X. Yang, X. Huang, J. Li, T. Shaftan, V. Smaluk, M. Song, W. Wan, L. Wu, and Y. Zhu, Accurate prediction of mega-electron-volt electron beam properties from UED using machine learning, Sci. Rep. **11**, 13890 (2021).

[21] Z. Zhang, X. Yang, X. Huang, T. Shaftan, V. Smaluk, M. Song, W. Wan, L. Wu, and Y. Zhu, Toward fully automated UED operation using two-stage machine learning model, Sci. Rep. **12**, 4240 (2022).

[22] S. Li and A. Adelmann, Time series forecasting methods and their applications to particle accelerators, Phys. Rev. Accel. Beams **26**, 024801 (2023).

[23] F. Sannibale, D. Filippetto, C. Papadopoulos, J. Staples, R. Wells, B. Bailey, K. Baptiste, J. Corlett, C. Cork, S. De Santis *et al.*, Advanced photoinjector experiment photogun commissioning results, Phys. Rev. ST Accel. Beams **15**, 103501 (2012).

[24] D. Durham, K. Siddiqui, F. Ji, J. G. Navarro, P. Musumeci, R. Kaindl, A. Minor, and D. Filippetto, Relativistic ultrafast electron diffraction of nanomaterials, Microsc. Microanal. **26**, 676 (2020).

[25] K. M. Siddiqui, D. B. Durham, F. Cropp, C. Ophus, S. Rajpurohit, Y. Zhu, J. D. Carlstrom, C. Stavrakas, Z. Mao, A. Raja, P. Musumeci, L. Z. Tan, A. M. Minor, D Filippetto, and R. Kaindl, Ultrafast optical melting of trimer superstructure in layered 1T′-$TaTe_2$, Commun. Phys. **4**, 152 (2021).

[26] F. Ji, D. B. Durham, A. M. Minor, P. Musumeci, J. G. Navarro, and D. Filippetto, Ultrafast relativistic electron nanoprobes, Commun. Phys. **2**, 54 (2019).

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

[28] F. Sannibale, D. Filippetto, H. Qian, C. Mitchell, F. Zhou, T. Vecchione, R. K. Li, S. Gierman, and J. Schmerge, High-brightness beam tests of the very high frequency gun at the Advanced Photo-injector EXperiment test facility at the Lawrence Berkeley National Laboratory, Rev. Sci. Instrum. **90**, 033304 (2019).

[29] G. Huang, L. R. Doolittle, Y. L. Xu, and J. Yang, Low noise digitizer design for ICLS-II llRF, in *Proceedings of North American Particle Accelerator Conference, NAPAC'16, Chicago, IL, USA* (JACOW, Geneva, Switzerland, 2016), TUPOA40, pp. 364–366, 10.18429/JACoW-NAPAC2016-TUPOA40.

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

[31] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, Learning internal representations by error propagation, California University San Diego La Jolla Institute for Cognitive Science, ICS Report No. 8506, 1985.

[32] M. Schuster and K. K. Paliwal, Bidirectional recurrent neural networks, IEEE Trans. Signal Process. **45**, 2673 (1997).

[33] S. Hochreiter and J. Schmidhuber, Long short-term memory, Neural Comput. **9**, 1735 (1997).

[34] J. Pérez, J. Marinković, and P. Barceló, On the turing completeness of modern neural network architectures, *International Conference on Learning Representations* (2019), https://openreview.net/forum?id=HyGBdo0qFm.

[35] I. Sutskever, O. Vinyals, and Q. V. Le, Sequence to sequence learning with neural networks, in *Advances in Neural Information Processing Systems*, edited by Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger (Curran Associates, Inc., 2014), Vol. 27, https://proceedings.neurips.cc/paper_files/paper/2014/file/a14ac55a4f27472c5d894ec1c3c743d2-Paper.pdf.

[36] S. Sodhani, S. Chandar, and Y. Bengio, Toward training recurrent neural networks for lifelong learning, Neural Comput. **32**, 1 (2020).

[37] M. McCloskey and N. J. Cohen, Catastrophic interference in connectionist networks: The sequential learning problem, *Psychology of Learning and Motivation* (Elsevier, New York, 1989), Vol. 24, pp. 109–165.

[38] J. Pouget-Abadie, D. Bahdanau, B. Van Merrienboer, K. Cho, and Y. Bengio, Overcoming the curse of sentence length for neural machine translation using automatic segmentation, arXiv:1409.1257.

[39] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, and M. Shah, Transformers in vision: A survey, ACM Comput. Surv. **54**, 1 (2022).

[40] T. Wolf *et al.*, Transformers: State-of-the-art natural language processing, in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* (Association for Computing Machinery, New York, NY, United States, 2020), pp. 38–45, 10.18653/v1/2020.emnlp-demos.6.

[41] B. Lim, S. Ö. Arık, N. Loeff, and T. Pfister, Temporal fusion transformers for interpretable multi-horizon time series forecasting, Int. J. Forecast. **37**, 1748 (2021).

[42] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, Attention is all you need, in *Advances in Neural Information Processing Systems*, edited by I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Curran Associates, Inc., 2017), Vol. 30, https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[43] W. Falcon *et al.* PyTorch lightning, GitHub Note, https://github.com/PyTorchLightning/pytorch-lightning **3** (2019).

[44] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, PyTorch: An imperative style, high-performance deep learning library, in *Advances in Neural Information Processing Systems*, edited by H. Wallach, H. Larochelle, A. Beygelzimer, F. d' Alché-Buc, E. Fox, and R. Garnett (Curran Associates, Inc., 2019), Vol. 32, https://proceedings.neurips.cc/paper_files/paper/2019/file/bdbca288fee7f92f2bfa9f7012727740-Paper.pdf.

[45] A. M. Dai and Q. V. Le, Semi-supervised sequence learning, *Advances in Neural Information Processing Systems*, edited by C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett (Curran Associates, Inc., 2015), Vol. 28, https://proceedings.neurips.cc/paper_files/paper/2015/file/7137debd45ae4d0ab9aa953017286b20-Paper.pdf.