

Machine-state embeddings as an operational reference space for accelerator operation

Chris Tennant^{*}

*Thomas Jefferson National Accelerator Facility,
12000 Jefferson Ave, Newport News, Virginia 23606, USA*

Jundong Li[✉] and Song Wang[✉]

*Department of Electrical and Computer Engineering, University of Virginia,
Charlottesville, Virginia 22903, USA*



(Received 7 April 2026; accepted 4 September 2026; published 1 October 2026)

We demonstrate that graph neural network embeddings of injector configurations provide a practical operational reference space for the Continuous Electron Beam Accelerator Facility injector at Jefferson Lab. Using 137,389 snapshots spanning January 2022 through March 2023, we show that injector operation occupies a small number of persistent, well-separated neighborhoods in a 16-dimensional learned state space rather than a featureless continuum. Density-based clustering identifies ten persistent operating regimes with strong operational run alignment, and regime persistence statistics confirm that these regimes are stable over timescales of hours to weeks. Large relocations between neighborhoods are rare and episodic; 99.6% of 1-h operating windows fall within an empirically derived jitter baseline. Geometric outlier screening narrows a year-long dataset to a small set of intervals warranting operational review, and nearest-neighbor retrieval enables case-based reasoning over the historical archive. A controlled beam study validates that deliberate injector reconfiguration traces coherent, interpretable trajectories in embedding space. Together these capabilities demonstrate that machine-state embeddings support holistic operational monitoring in ways that single-channel inspection cannot.

DOI: [10.1103/d1b7-qdcm](https://doi.org/10.1103/d1b7-qdcm)

I. INTRODUCTION

Modern particle accelerator operation produces a machine state that is simultaneously high-dimensional and highly coupled. Magnets, rf systems, diagnostics, and beam-dependent readbacks evolve together under a web of nonlinear dependencies and hidden operational constraints. Inspecting individual signals in isolation offers an incomplete and often misleading picture of how the machine is actually behaving, and tracking the operational history of a complex beamline through time is correspondingly difficult. The fundamental challenge is representation: how should one encode the global state of a machine with hundreds of channels in a form that exposes the operational structure?

This paper explores whether a learned low-dimensional embedding can serve as a practical operational reference space for the Continuous Electron Beam Accelerator

Facility (CEBAF) injector. By an operational reference space we mean a metric space, built from historical operating data, in which operational similarity is encoded geometrically. It is a compressed index of past machine states rather than a fixed physical coordinate system, and its validity is tied to the epoch spanned by the data used to construct it. If similar machine states map to nearby points in embedding space, then the geometry of the state space becomes a proxy for operational similarity, and a wide range of analysis tasks—identifying recurring modes, detecting unusual conditions, retrieving historical precedent—become questions about the geometry rather than about individual channels. Such an embedding does not simplify the machine by ignoring complexity; rather, it organizes that complexity into a form that can be reasoned about holistically.

The embedding analyzed here was produced by prior work in which CEBAF injector configurations were represented as heterogeneous directed graphs, with nodes corresponding to beamline elements and features encoding settings and readbacks [1]. A graph neural network (GNN) was trained on historical data using self-supervised contrastive learning, yielding a 16-dimensional latent representation in which operational similarity is encoded geometrically [2]. The present study takes that embedding

^{*}Contact author: tennant@jlab.org

Published by the American Physical Society under the terms of the [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/). Further distribution of this work must maintain attribution to the author(s) and the published article's title, journal citation, and DOI.

as its foundation and investigates whether the resulting state space supports the operational analyses listed above at scale, over a 14-month operational archive.

The contributions of this paper are threefold. First, we characterize the large-scale structure of the CEBAF injector state space and show that it consists of a small number of persistent operating regimes with strong operational run alignment and quantifiable persistence statistics. Second, we develop and validate a suite of operational analysis tools—stability assessment, outlier screening, and historical analog retrieval—that operate on the geometry of the embedding and require no hand-crafted anomaly definitions. Third, we examine where embeddings add value over single-channel monitoring and discuss the practical conditions under which alternative representations—from direct feature-space projections to transformer-based encoders—may be appropriate.

II. RELATED WORK

The embedding model analyzed here builds directly on earlier work by Wang *et al.* [1], in which the CEBAF injector was modeled as a heterogeneous directed graph and a GNN was trained to produce latent representations of operational configurations. That work established that the learned space encodes meaningful operational structure, motivating the present deeper analysis of what that structure implies for monitoring and operations. In the present application, the CEBAF injector was represented as a heterogeneous directed graph whose node and edge definitions were chosen to reflect the machine’s operational topology and available settings and readbacks. The contrastive training objective follows Chen *et al.* [2]. Graph-specific extensions of contrastive learning—in which domain-appropriate augmentations (node dropping, feature perturbation) are applied to create positive pairs—have since been developed and shown to yield superior representations for graph-structured data [3].

More broadly, the use of latent representations has become an important organizing principle of machine learning applications for particle accelerators. Prior studies have used an encoder-decoder architecture with uncertainty quantification for virtual diagnostics [4], encoder-decoder networks for high-fidelity prediction of longitudinal phase-space images from noninvasive machine inputs [5], conditional variational autoencoders combined with recurrent models to learn reduced latent representations for generation and forecasting of beam dynamics [6], conditional Siamese architectures for errant-beam prognostics under changing accelerator configurations [7], conditional latent diffusion models for inverse reconstruction of phase-space projections from noninvasive beam current and beam loss measurements [8], and convolutional autoencoders for reconstructing distorted diagnostic signals into transverse beam profiles [9]. In these works, the latent space typically serves as a

compressed intermediate representation: a bottleneck for mapping noninvasive measurements to beam images, a reduced state space for learned beam evolution, a generative manifold for inverse diagnostics, or a similarity space for detecting abnormal behavior. The present work is aligned with this broader trend but uses the embedding differently. Rather than targeting a specific diagnostic image, forecast, or fault label, we treat the learned 16-dimensional representation itself as an operational reference space for the whole injector state. This shifts the emphasis from producing a particular downstream artifact to analyzing the geometry of historical operation in the embedding space.

Beyond the specific use of latent representations, machine learning has been identified as a high-priority opportunity across virtually all domains of accelerator science—from beam diagnostics and fault detection to tuning and surrogate modeling [10]. In the context of accelerator physics, the problem of making sense of high-dimensional diagnostic data has been approached in several ways. At the LHC, Fol *et al.* [11,12] applied unsupervised learning to beam diagnostic data for BPM fault detection using Isolation Forest [13]—a method closely related to the k -nearest-neighbor (k NN) outlier scoring used in Sec. VI below—and demonstrated that learned representations can identify faulty instruments and reveal patterns in operational behavior without requiring labeled examples. Condition monitoring in industrial machinery has seen analogous development. Wójcik and Barszcz [14] showed that training a variational autoencoder [15] on healthy data alone produces a latent space whose geometry supports anomaly detection through distance-based measures, without requiring any labeled fault data—a paradigm that underlies the present work as well. Once a meaningful latent representation has been learned, a natural next question is how to inspect and interpret its geometry. In that context, principal component analysis (PCA) [16] and the broader toolkit of nonlinear dimensionality reduction—including t-SNE [17] and UMAP [18]—have established that low-dimensional projections can expose cluster structure, transitional states, and outliers in high-dimensional scientific datasets, motivating their use as diagnostic tools across domains. The insight that low-dimensional projections of complex process data can reveal actionable operational structure has also been demonstrated in chemical process control by Joswiak *et al.* [19], who benchmarked dimensionality reduction methods for visualizing high-dimensional plant data and showed that projections expose operating regimes, transitions, and outliers even in strongly nonlinear, highly coupled processes. The CEBAF injector shares those properties, and the results below can be understood as an accelerator-specific realization of this broader paradigm.

III. EMBEDDING AND DATASET

A. Graph neural network embedding

Each injector snapshot is represented as a heterogeneous directed graph with 207 nodes drawn from 12 device types (correctors, quadrupoles, solenoids, srf and warm rf cavities, BPMs, BLMs, ion pumps, and current monitors, among others), with directed edges encoding the physical relationships among them. Each node carries a feature vector built from the 393 process variables associated with that snapshot (267 settings and 126 readings), standardized feature-by-feature using means and standard deviations computed from the historical operational dataset. No timestamp or other time-derived quantity is included in the graph or provided to the model, so cluster assignments in Sec. IV A reflect machine configuration alone.

Node-type-specific linear input layers first map these heterogeneous node features into a common 16-dimensional hidden space. Three graph attention layers [20] and a final linear projection then update the node representations using the graph structure, after which mean pooling over all nodes yields a single 16-dimensional embedding for the complete machine state.

The GNN was trained with the Adam optimizer using a self-supervised objective that combines contrastive learning on feature-masked views of each graph with an alignment loss that preserves distances between machine states in the embedding space [1,2]. The architecture and self-supervised training procedure follow Ref. [1]; the model used in this work was not fine-tuned as in that earlier study, but trained from scratch, following the same procedure, on the full 137,389-snapshot archive analyzed in this paper, spanning January 2022 through March 2023. The training procedure encourages similar operational configurations to map to nearby points in the embedding space, while dissimilar configurations are pushed apart. The resulting embedding is not a compression of individual channel values but an encoding of the relational structure of the whole machine state.

For visualization, PCA is applied to the 16-dimensional embeddings to obtain a 2D projection. Except where explicitly noted, all quantitative analyses in this paper—distances, clustering, outlier scoring, and retrieval—are performed in the full 16-dimensional space. PCA serves solely as a tool for human-interpretable visualization of results computed in the higher-dimensional space.

B. Dataset

The dataset consists of 137,389 injector snapshots spanning January 4, 2022 through March 20, 2023, covering 439 days of operation. Snapshots were collected every 2 min whenever the injector beam current was at least 0.05 μA . Beam current across the dataset ranged from 0.05 to 151.37 μA . Each snapshot carries a 16-dimensional

GNN embedding, 267 setting features, 126 reading features, and a 393-dimensional normalized feature vector constructed from the full machine state. Index alignment and timestamp consistency were applied as preprocessing steps to ensure that all retained records represent valid, consistent injector configurations.

IV. INJECTOR OPERATIONS AS A STRUCTURED STATE-SPACE LANDSCAPE

A natural first question is whether the embedding organizes injector history into interpretable structure or merely redistributes noise in a lower-dimensional space. The month-colored PCA projection in Fig. 1 provides an immediate answer: the 14-month operational history does not fill the 2D plane uniformly but instead occupies several clearly separated islands. The January–February 2022 data concentrate in a compact region in the upper left. The mid-2022 datasets (June–August) populate a central-left region with visible internal substructure. The late-2022 data form a separate lower region, and early-2023 data cluster predominantly in the right-hand portion. These islands are not mere visualization artifacts; they are consistent across the density and beam current views shown in Fig. 2. Each island contains one or more high-occupancy cores—the most frequently visited operating neighborhoods—surrounded by lower-density halos and sparse satellite patches consistent with brief excursions or transitional states. Beam current varies within these regions rather than mapping one-to-one onto a single global direction, implying that the embedding encodes a richer operational signature than beam current alone could provide.

A. Persistent operating regimes

To move from a visual impression to a quantifiable characterization, we applied HDBSCAN [21,22] directly to the 16-dimensional embedding space. HDBSCAN is a density-based clustering algorithm that identifies groups of points in dense regions while leaving isolated or weakly connected points unlabeled as noise. It does not require the number of clusters to be specified in advance and is therefore well suited to an embedding space that may contain groups of different size and density.

The analysis identifies ten persistent operating regimes and a noise class. Only 4694 points (3.4%) are classified as noise, indicating that the great majority of operational history falls cleanly within one of the identified regimes. These ten clusters are best understood as a descriptive summary of persistent operational modes in this dataset. The annotated cluster map is shown in Fig. 3.

A small number of clusters dominate dwell time. Cluster 9 is the largest (43,094 points, mean beam current 30.96 μA), followed by cluster 1 (31,616 points, 42.44 μA), cluster 7 (18,216 points, 50.53 μA), cluster 2 (16,142 points, 38.16 μA), and cluster 6 (12,816 points,

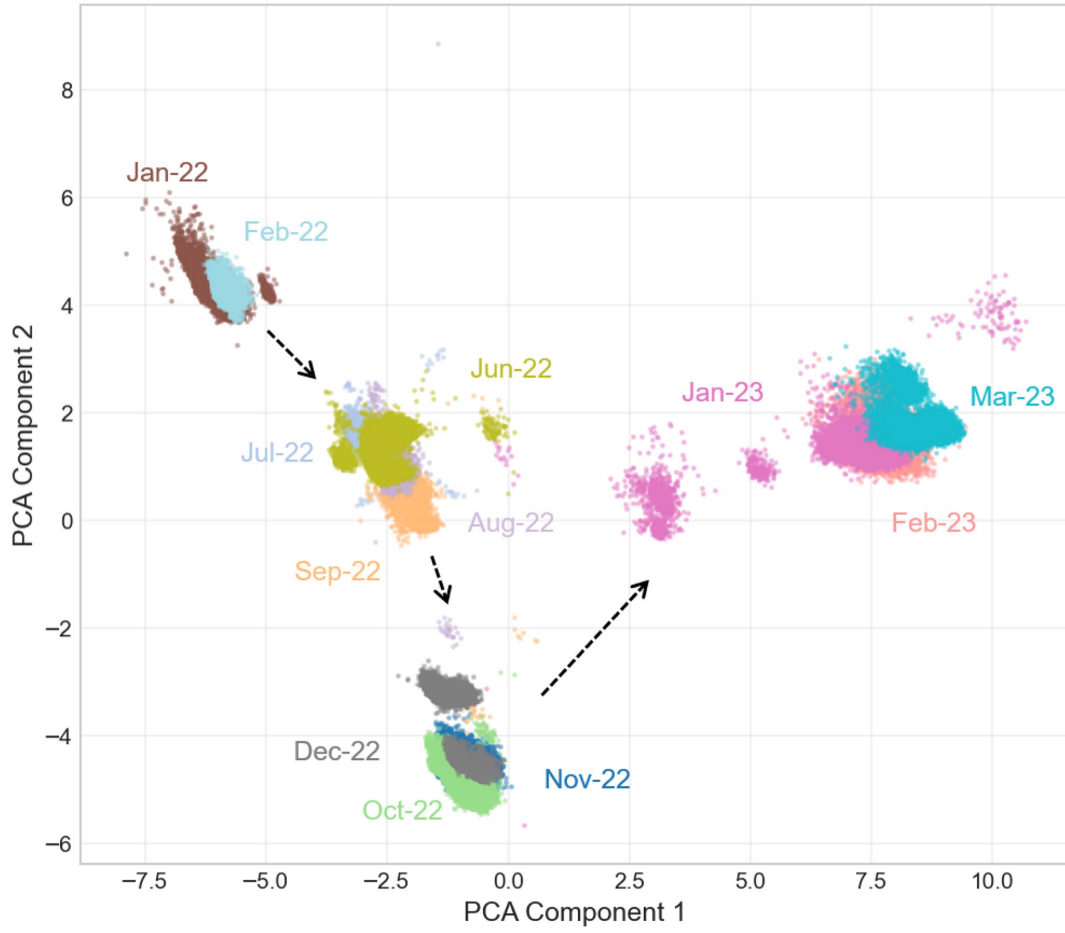


FIG. 1. PCA projection of the GNN embedding colored by calendar month. The operational history from January 2022 through March 2023 occupies several discrete islands rather than a single contiguous cloud. Dashed arrows indicate the general progression of time.

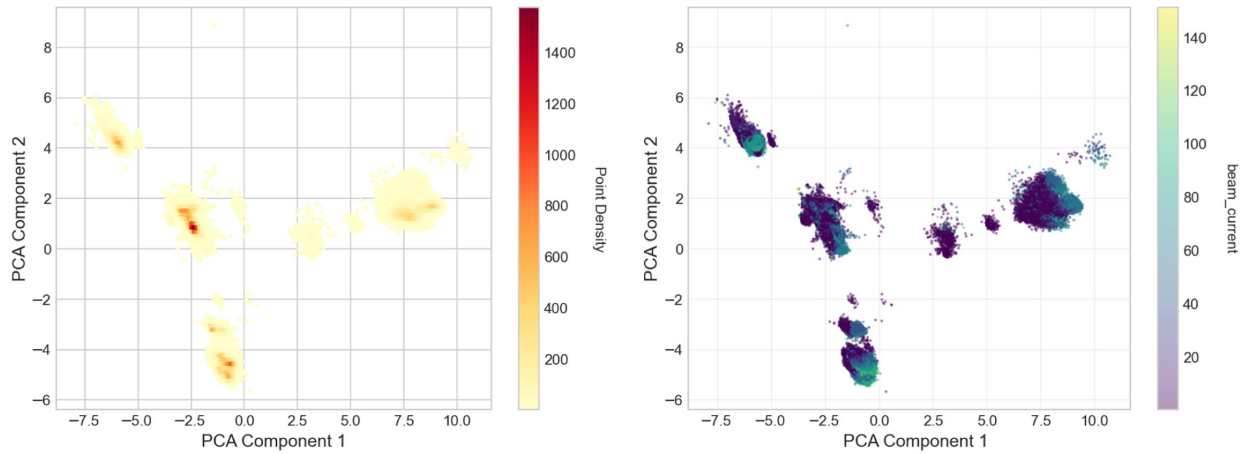


FIG. 2. PCA projection colored by point density (left) and beam current (right). Dense cores mark frequently occupied operating neighborhoods; beam current varies within neighborhoods rather than aligning with a single global direction, indicating that embedding structure reflects more than current alone.

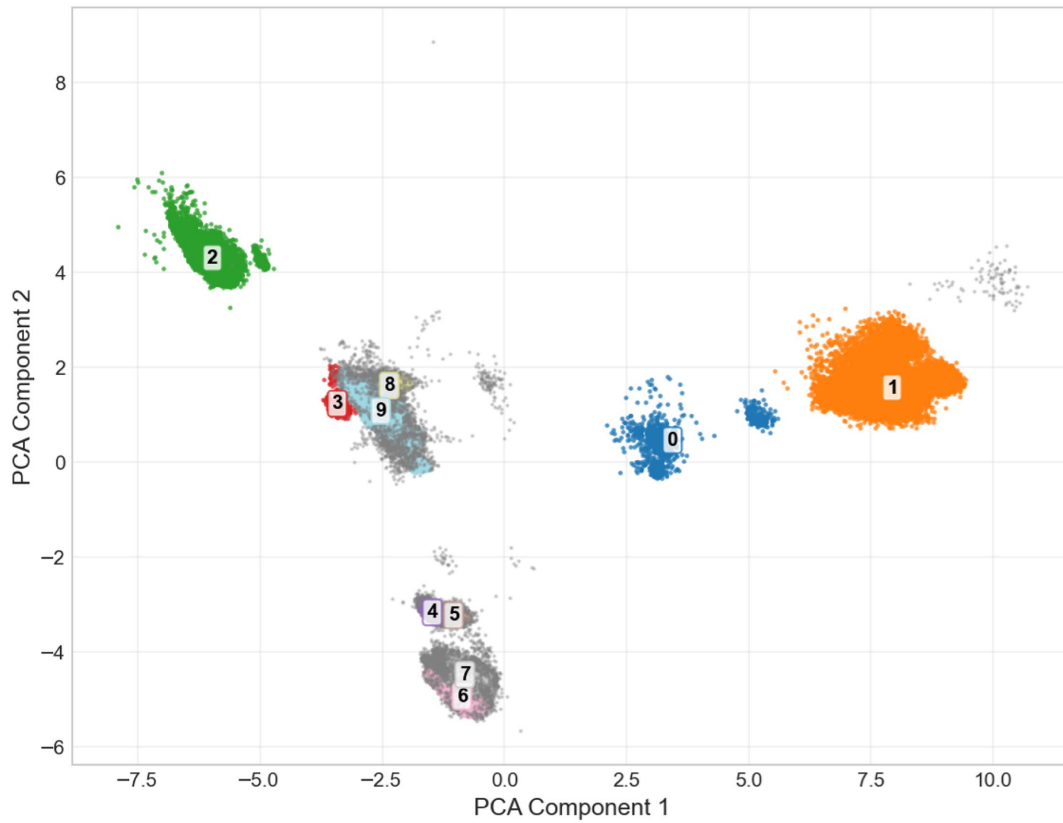


FIG. 3. HDBSCAN clusters identified in the 16D embedding space and visualized in the 2D PCA projection. Cluster labels mark persistent operating regimes and gray points denote noise or weakly assigned states.

40.21 μA). Together, these five clusters account for the large majority of operational time.

B. Regime persistence

That clusters exist does not by itself establish that they are persistent—they could in principle correspond to frequently revisited but short-lived excursions. To quantify persistence, we computed, for each cluster, the number of times the machine entered it (defined at the 8-h shift level), the median contiguous dwell time in hours, and the 90th-percentile dwell time. Results are summarized in Table I. A dwell run is a maximal sequence of consecutive shifts in the same cluster; a run ends when the cluster assignment changes or when one or more shifts are absent from the data (i.e., a gap of at least 8 h with no samples).

The dominant clusters are entered only a handful of times but maintain median dwells measured in days: cluster 1 entered 4 times with a median dwell of 396 h (~ 16.5 days) and a 90th-percentile dwell of 581.6 h (~ 24 days). Cluster 7 entered 3 times with a median of 160 h (~ 6.7 days). This persistence is also visible in the stacked timestamp histogram of Fig. 4, which shows clusters dominating long, contiguous run periods rather than being uniformly interleaved in time. The noise class (−1), by contrast, enters 24 times with a median dwell of

only 8 h, confirming that unclassified periods are typically transitional episodes rather than an alternative stable operating mode.

One notable feature of Fig. 4 is that the noise label is absent during January–February 2022. A plausible interpretation is that early-2022 operation was dominated by a

TABLE I. Summary of operating regimes identified by HDBSCAN clustering in the 16D embedding space. Each 8-h shift is assigned its dominant cluster (by mode). Persistence statistics are computed across all such runs. Rows are sorted by descending median dwell time.

Cluster ID	No. entries	Median dwell (h)	p_{90} dwell (h)
1	4	396	581.6
7	3	160	448.0
9	15	136	275.2
6	5	104	236.8
2	7	88	280.0
0	1	88	88.0
8	3	64	83.2
5	2	44	66.4
3	2	24	30.4
4	4	16	116.8
−1	24	8	13.6

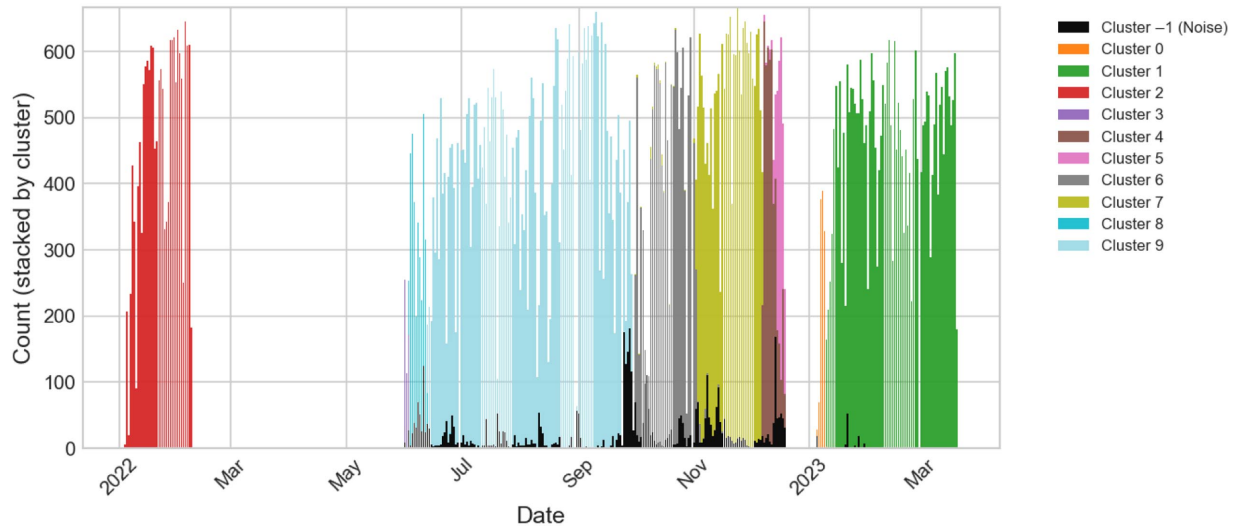


FIG. 4. Time distribution of cluster membership shown as a stacked histogram. Long contiguous blocks indicate that operating regimes persist over extended periods rather than being uniformly interleaved in time.

single especially dense and homogeneous operating region, with all nearby states absorbed into clusters. Beginning in June 2022, the embedding space becomes more fragmented as distinct operating modes multiply, creating sparse intercluster regions that HDBSCAN naturally assigns to noise.

This homogeneity is precise enough to be worth stating explicitly: cluster 2 matches, point for point, the January 4 through February 8, 2022 stretch of the dataset (16,142 points). This is not surprising. That window opens immediately after a holiday shutdown in December 2021 and closes ahead of an extended, multimonth maintenance period beginning in February 2022. The machine ran continuously between those two boundaries without a substantial reconfiguration, so a single dense operating regime spanning the full window is the expected outcome rather than a coincidence. Not every cluster is confined to one continuous run in this way. Four of the ten clusters—1, 6, 7, and 9—reappear across separate, noncontiguous stretches of the dataset, indicating regimes the machine returns to independently of any single run. The remaining clusters, including cluster 2, are observed during only one continuous run.

The GNN encoder was trained on the full 137,389-snapshot archive, so the absence of noise during January–February 2022 cannot be attributed to that period being disproportionately represented during training. With 16,142 snapshots it is one of the smaller continuous runs in the dataset, well below several other multiweek stretches (for example, 26,516 snapshots in November–December 2022 and 26,191 in September–October 2022). Evidence that does not depend on the embedding, computed entirely from the raw feature arrays, also corroborates the “quieter operation” reading directly. Computed entirely from the raw feature arrays, without reference to

the embedding, January–February 2022 shows lower reading-space dispersion (8.11 vs 9.39, a 16% reduction) and lower total variance of the 393-dimensional normalized feature vector (197 vs 268, a 26% reduction) than June–August 2022. Because the two periods have very different sample sizes (16,142 vs 35,575 snapshots), we checked that this is not an artifact of that imbalance by repeatedly subsampling the larger June–August period down to the smaller period’s size and recomputing both statistics. The observed January–February values fall below the resulting 95% confidence interval in both cases, confirming that the reduced variability is a genuine property of that period’s operation rather than a sample-size effect.

C. Cluster fingerprints

Geometric regimes are operationally useful only if they connect back to the underlying machine. For each cluster, we compute a fingerprint by comparing that cluster’s mean settings and readings to the corresponding full-dataset means. The top-ranked features by absolute delta are those that most strongly distinguish the regime from typical operation. To illustrate the interpretive value of this approach, we focus on two clusters whose fingerprints map cleanly onto recognizable injector subsystems: cluster 2 and cluster 6.

Table II lists the top four settings and top four readings for each of these two clusters. Cluster 2 is defined primarily by four warm chopper cavities—CHOP1Y and CHOP1X (the vertical and horizontal settings of the first chopper), and CHOP2X and CHOP2Y (the corresponding settings of the second chopper). The choppers control the pulse structure of the beam extracted from the electron gun, and their dominance in the fingerprint indicates that cluster 2 corresponds to a mode defined by a specific configuration during the beam formation stage.

TABLE II. Top four setting and reading features distinguishing cluster 2 and cluster 6 from the full-dataset average, based on largest absolute cluster-minus-global-mean deltas. Element types are noted in parentheses.

	Top 4 settings	Top 4 readings
Cluster 2	CHOP1Y (chopper 1—vertical), CHOP1X (chopper 1—horizontal), CHOP2X (chopper 2—horizontal), CHOP2Y (chopper 2—vertical)	VIP0L0450 (ion pump), IPM0L10 (beam position monitor), VIP0L04B (ion pump), VIP0L05A (ion pump)
Cluster 6	0L04-8 (srf cavity), 0L04-6 (srf cavity), 0L04-5 (srf cavity), 0L04-3 (srf cavity)	VIP0L00 (ion pump), IPM0R04 (beam position monitor), IPM0L02 (beam position monitor), IPM0L04 (beam position monitor)

Cluster 6, by contrast, is defined by four srf cavity setpoints in the 0L04 cryomodule—the last module in the injector before the beam enters the main linac. Specifically, 0L04-8, 0L04-6, 0L04-5, and 0L04-3 are individual cavities within that module. The associated readings are ion pump currents and beam position monitors along the medium-energy transport line. Together, the cluster 6 fingerprint points to a mode in which the energy delivered by the final injector cryomodule is the primary organizing variable, distinguishing it geometrically from other clusters that share similar chopper or gun settings but differ in 0L04 configuration.

These two examples illustrate the broader point: fingerprints do not establish causality, but they do provide an interpretable bridge from abstract geometric structure back to specific subsystems, enabling an operator or subject matter expert to reason about what physical configuration each persistent regime represents.

V. DYNAMICS OF MOTION THROUGH STATE SPACE

Having established that the embedding is organized into persistent neighborhoods, we turn to how the machine moves through that space over time. The analysis reveals a characteristic pattern across multiple timescales: most motion is small and routine, while large relocations are rare and episodic.

A. Shift-to-shift step sizes

To quantify motion at an operationally meaningful timescale, we aggregate snapshots into owl, day, and swing shifts and represent each shift by its centroid in the 16-dimensional embedding space. The Euclidean distance between consecutive shift centroids then measures shift-to-shift reconfiguration. Across 885 shifts (884 transitions), the step-size distribution is strongly

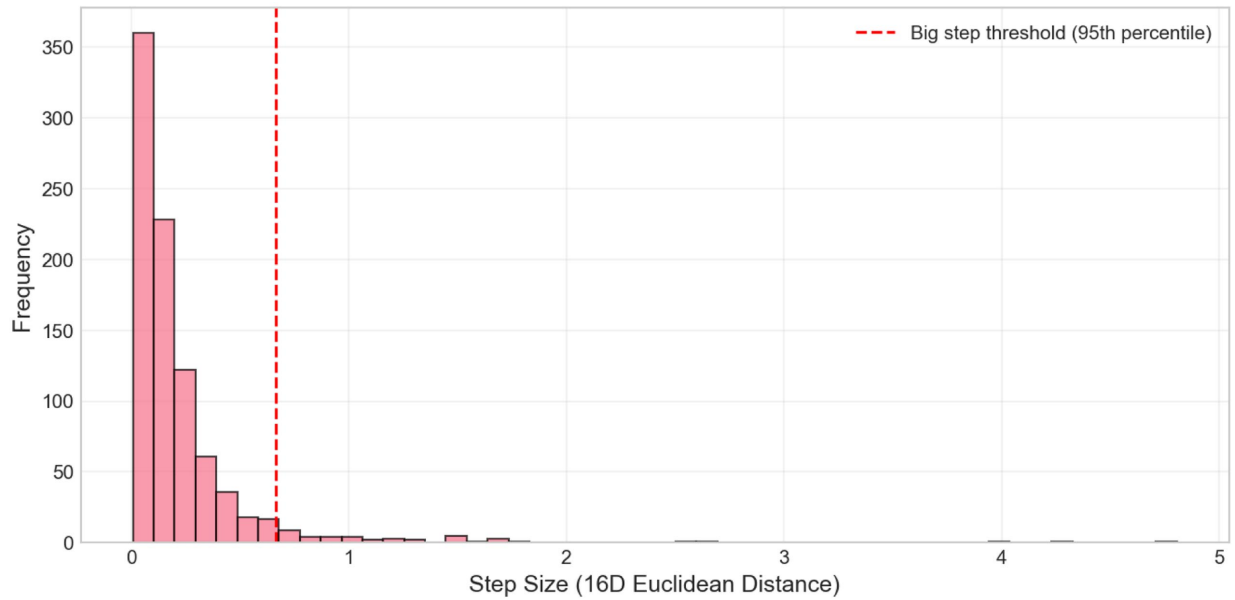


FIG. 5. Distribution of step sizes for the 884 shift-to-shift transitions, computed as the Euclidean distance between consecutive shift centroids in 16D embedding space. The distribution is strongly right-skewed; the dashed line marks the 95th-percentile “big step” threshold of 0.67.

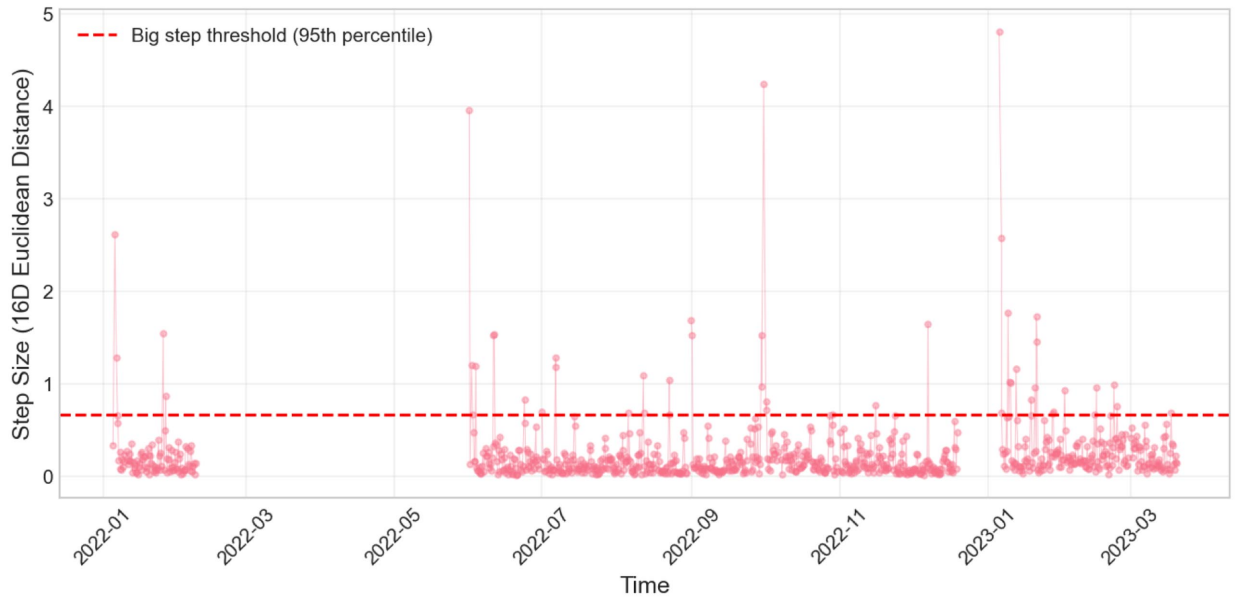


FIG. 6. Shift-to-shift embedding-space step sizes as a function of time. The dashed line marks the 95th-percentile threshold; large relocations appear as isolated spikes rather than sustained elevations.

right-skewed. The distribution peaks sharply near zero, with most transitions clustered well below 0.3 in 16D Euclidean distance, while a long tail of rare large moves extends toward values above 1.0 (Fig. 5).

Using the 95th-percentile value of 0.67 as a threshold for large relocations, these events appear as isolated spikes in time (Fig. 6) rather than sustained elevations. The largest jumps cluster on a few specific dates: January 5, 2023, September 30, 2022, and June 1, 2022. Table III lists the ten largest transitions and their associated beam current changes. Several large steps correspond to returns from extended downtime.

Within-shift variability also differs by shift type. Mean within-shift path lengths are 32.28 for day shifts, 36.96 for swing shifts, and 41.28 for owl shifts, suggesting that day operations are typically more stable. When examining

the shift handoff—the cumulative embedding-space step associated with entering each shift type, excluding long-downtime gaps—the smoothest routine handoff is swing-to-owl (cumulative step 54.33), whereas the largest reconfiguration tends to accompany the owl-to-day transition (67.91).

B. Long-horizon drift

The drift-to-reference metric measures the distance of each shift centroid from the final shift of the dataset (March 20, 2023). As shown in Fig. 7, this view reinforces the picture of discrete relocations. Early 2022 lies far from the final reference, mid- to late-2022 occupies an intermediate plateau with a visible jump around early October, and early 2023 collapses to much smaller distances after the abrupt January transition. The machine spends long stretches at approximately constant distance from the reference—confirming extended residence in stable neighborhoods—punctuated by rapid steps.

C. The September 30, 2022 case study

Among the large relocation events, the September 30, 2022 transition (step size 4.24, second largest in the dataset) provides an opportunity to connect embedding-space motion to specific operational context. To investigate, logbook entries from September 28–October 1, 2022 were extracted, filtered for human-authored narrative content, and analyzed using large language models. The logbook analysis workflow is described in Sec. XI.

Both models converge on a compound operational event rather than a single isolated cause. A full maintenance day on September 29 was followed by rf recovery and setup

TABLE III. The ten largest shift-to-shift step sizes, with associated beam current change and gap since the previous shift. Step size is the Euclidean distance between consecutive shift centroids in the 16D embedding space.

Date-time	Step size	ΔI (μA)	Gap (days)
2023-01-05 16:04	4.81	−0.57	17.30
2022-09-30 13:10	4.24	−6.57	0.58
2022-06-01 00:50	3.95	−24.25	112.73
2022-01-05 18:14	2.62	18.86	0.47
2023-01-06 15:48	2.57	33.36	0.83
2023-01-09 08:28	1.77	−0.12	0.02
2023-01-21 00:16	1.72	23.59	0.01
2022-08-31 12:30	1.69	−26.99	1.52
2022-12-07 08:00	1.65	−48.40	0.00
2022-01-25 17:08	1.54	−33.89	0.40

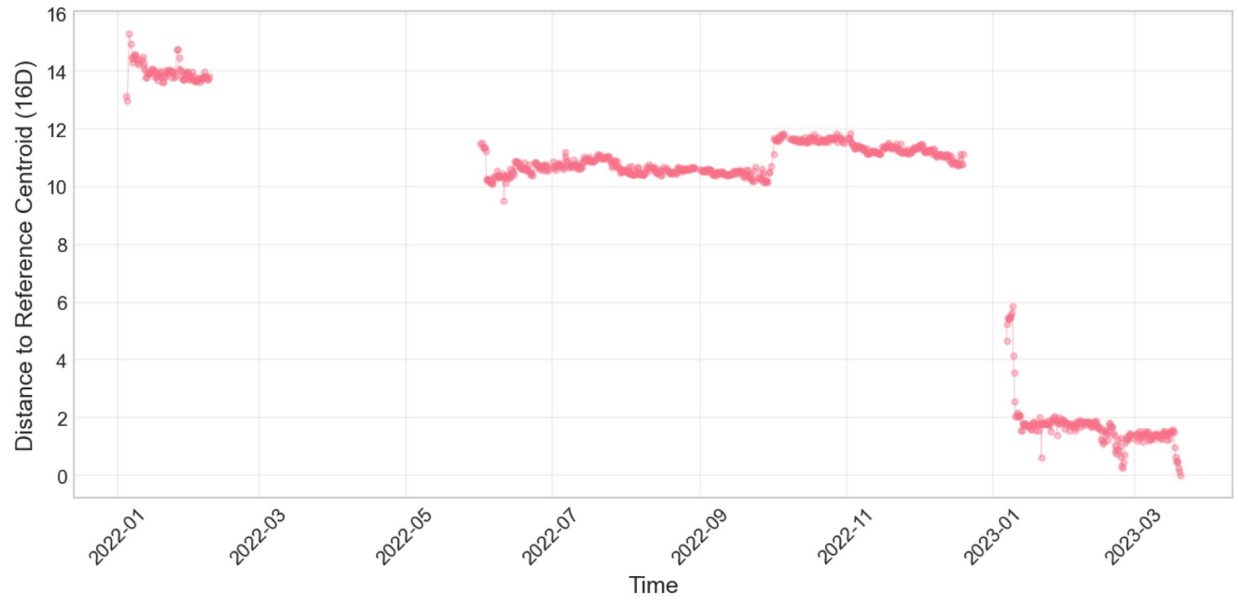


FIG. 7. Distance of each shift centroid from the final shift (March 20, 2023), plotted over time. Breaks indicate extended beam-off periods. The structure of stable plateaus and abrupt jumps reinforces the picture of persistent regimes with rare large transitions.

changes intended to support multihall running, including gangphase and path-length adjustments. Recovery was then disrupted by cascading faults: a North Linac 2 K cold-box trip, cavity instabilities and deratings, magnet over-temperature faults, and additional control issues. The resulting large embedding-space jump reflects not a single parameter change but a substantial, partially degraded reconfiguration of the injector following maintenance and multisystem recovery. This is precisely the kind of event that is difficult to detect from any single channel but is visible as a large step in the holistic machine-state representation.

D. Regime transition dynamics

Downsampling to the shift level, the regime-to-regime transition graph reduces to a sparse set of pathways. Across the full run there are 50 shift-scale transitions spanning 19 unique directed pairs. The most common exchanges involve the noise label: transitions from noise (-1) to cluster 9 and the reverse each occur 12 times, while other pairings occur only a handful of times (Table IV). This sparsity confirms that the machine rarely changes

TABLE IV. Most frequent regime transitions at the shift level. The dominance of noise-mediated transitions indicates that inter-regime moves typically pass through unclassified states.

From cluster	To cluster	Count
-1	9	12
9	-1	12
8	-1	3
-1	8	2

regimes at the shift scale. When it does, the change often passes through transitional noise states rather than jumping directly between clusters.

When the machine enters the noise label, it typically does not remain there long: 651 separate noise runs are identified, with a median dwell time of 8.0 min and a mean of 14.8 min, though the distribution is heavy-tailed with a maximum of 268 min. Note that the fifth longest noise dwell occurs in late September (Table V), which aligns with the degraded machine state documented in the September 30 case study above.

VI. OPERATIONAL UTILITY

A. Stability baseline and anomaly screening

Identifying genuinely unusual intervals requires an empirical baseline for expected variability under stable operation. We define this baseline using anchor windows—multihour periods during which settings remained essentially constant—computing the L_2 reading-space dispersion

TABLE V. Five longest contiguous dwell periods in the HDBSCAN noise class (-1). The concentration in late September–early October 2022 is consistent with the compound operational event documented in the September 30 case study.

Start	End	No. points	Dwell (min)
10/06/2022 17:06	10/06/2022 21:34	95	268
10/07/2022 21:28	10/08/2022 01:04	105	216
12/14/2022 06:10	12/14/2022 09:06	62	176
06/11/2022 01:06	06/11/2022 03:50	72	164
09/28/2022 01:30	09/28/2022 04:10	48	160

within each window as a reference distribution. Concretely, for a window containing N snapshots with reading vectors $\mathbf{r}_1, \dots, \mathbf{r}_N \in \mathbb{R}^d$ ($d = 126$), the dispersion is defined as the mean pairwise Euclidean distance from each point to the window centroid $\bar{\mathbf{r}}$:

$$D = \frac{1}{N} \sum_{i=1}^N \|\mathbf{r}_i - \bar{\mathbf{r}}\|_2. \quad (1)$$

Five anchor windows were identified (901 samples total) and their pooled 95th-percentile dispersion of 12.44 defines the jitter baseline. The stability ratio for any 1-h operating window is then

$$s = D_{\text{window}} / D_{p95}, \quad (2)$$

where D_{window} is the reading-space dispersion for that window and D_{p95} is the pooled baseline. Values of $s \leq 1$ are consistent with routine jitter whereas values above 1 indicate dispersion exceeding the empirical noise floor.

Of the 6263 1-h windows evaluated across the full dataset, 6241 (99.6%) have stability ratio $s \leq 1$, with a median of 0.38. Only 22 windows (0.4%) exceed the jitter baseline, appearing as isolated spikes (Fig. 8) rather than sustained periods of elevated variability. The maximum ratio observed is 2.36. As shown in Fig. 9, elevated stability ratios are concentrated at very low beam current, whereas moderate- and high-current operation clusters uniformly below $s = 1$. This pattern suggests that departures from the baseline are associated with low-current setup or recovery periods rather than with instability during established beam delivery.

This baseline is deliberately representation-independent. Equation (1) operates on the raw 126-dimensional reading vectors and does not reference the embedding at all. That is

a design choice rather than an oversight. It establishes, on grounds no embedding could dispute, that aggregating variability across all simultaneous readbacks is itself a meaningful operational quantity, and it provides a control against which the embedding's added value can be measured elsewhere in the paper.

An embedding-space analog of the same construction can also be built: an anchor-window jitter baseline defined the same way, but using the 16-dimensional embedding vectors $\mathbf{e}_1, \dots, \mathbf{e}_N$ in place of the raw reading vectors $\mathbf{r}_1, \dots, \mathbf{r}_N$, and likewise for the two raw-feature representations compared in Sec. IX B. These three representations agree that stable operation dominates: 90%–99% of 1-h windows fall below their own jitter floor, depending on representation. However, the specific windows each one flags as unstable overlap only partially with one another. We quantify this overlap using the Jaccard index, the number of windows two representations both flag as unstable divided by the number flagged as unstable by either one; a value of 1 means two representations flag exactly the same windows, and a value of 0 means they share none. The pairwise Jaccard overlap among these three representations is only 0.08–0.23. This indicates that “instability” is not a single representation-independent quantity; it is defined relative to whichever variability the representation happens to be sensitive to. The reading-space construction used in this section remains the most directly interpretable of the four representations considered, since it is defined on physical units with no intervening transformation.

B. Geometric outlier screening

As a complement to the density-based clustering of Sec. IV A, we screen for outliers using a second, independent method: a distance-based isolation score

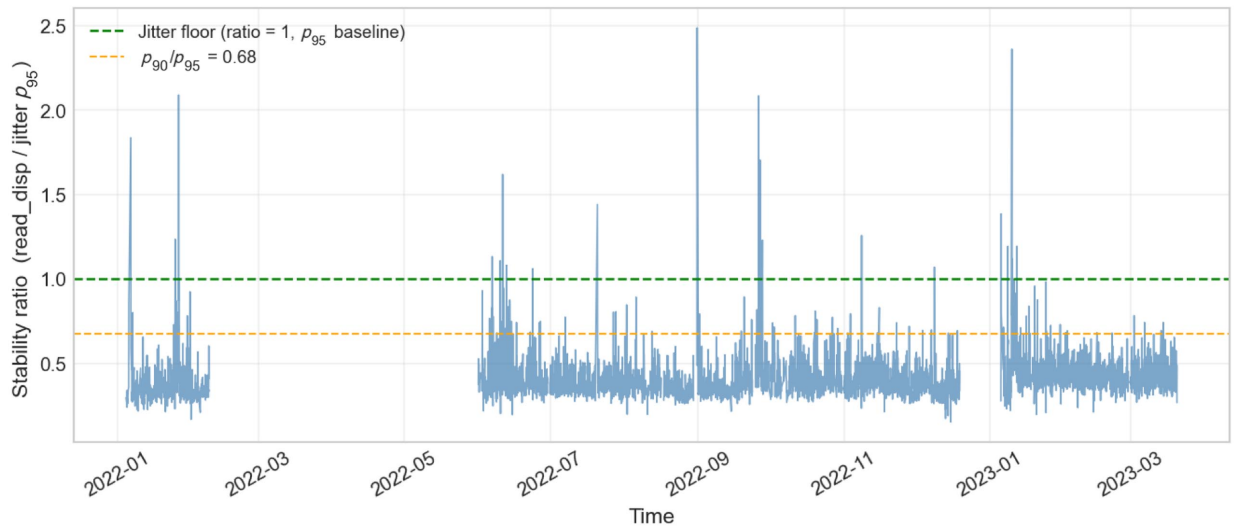


FIG. 8. Stability ratio (reading-space dispersion normalized to the pooled anchor-window p_{95} baseline) for all 1-h windows across the 14-month dataset. Only 22 of 6263 windows (0.4%) exceed the jitter threshold ($s = 1$, dashed green line).

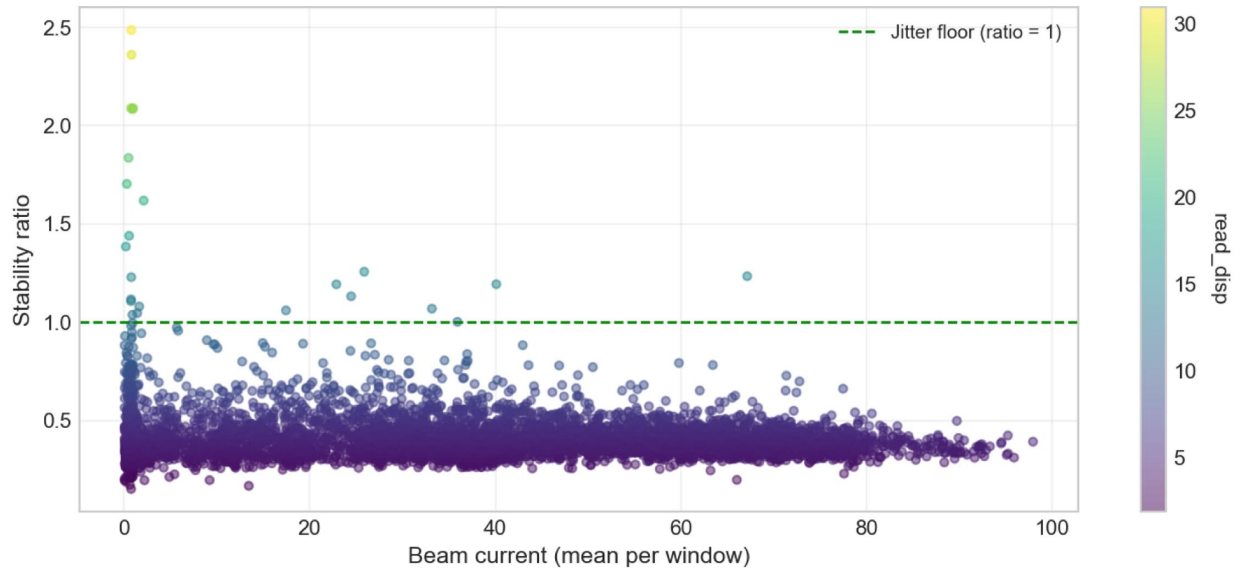


FIG. 9. Stability ratio vs mean beam current per 1-h window. Elevated ratios occur primarily at very low current, consistent with setup or recovery activity rather than instability during normal beam delivery.

computed for every point from its local neighborhood in the full 16D embedding space, using the distances to its k NN [23,24]. Larger scores indicate that a point is geometrically more isolated from the historical cloud of operating states. We flag the 1374 points (1.0%) whose isolation score exceeds the 99th percentile as k NN outliers.

The k NN outlier count and the HDBSCAN noise fraction of 4694 points (3.4%) capture related but distinct notions of unusual behavior. HDBSCAN labels a point as noise when it falls in a region too sparse to be assigned confidently to any of the identified clusters, whereas the k NN score is not a classification step but a continuous isolation measure computed for every point. Table VI summarizes the key distinctions between the two methods.

Because the k NN count is fixed by construction at the 99th-percentile threshold, it should not be read as a discovered result. The HDBSCAN noise fraction is not fixed in this way, since HDBSCAN determines which points are too sparse to join a cluster from the data’s own

density structure rather than from a preset quantile. The substantive findings of the k NN screen therefore concern where and when flags land, not how many there are. In the PCA overlay (Fig. 10), flagged points concentrate near the boundaries of major operating islands and in low-density intercluster regions, indicating that outliers are typically excursions away from dominant neighborhoods rather than evidence for additional undetected regimes. Merging temporally adjacent outliers with a 300-s gap threshold yields 806 distinct intervals, confirming that anomalous episodes are usually brief and fragmented. The three longest intervals are: 5640 s on January 21, 2023 (01:28–03:02), 2520s on July 20, 2022 (20:24–21:06), and 2400s on January 27, 2023 (20:18–20:58).

To check that this concentration into relatively few intervals is a genuine pattern and not an artifact of the merging procedure, we compared it against a null model. In this null model, the same number of outlier timestamps (1374) are drawn uniformly at random from the full

TABLE VI. Comparison of the two anomaly-related analyses. The methods are complementary: noise dwell asks how long the machine remained outside the main regimes while k NN scoring asks how geometrically isolated each state was from the full operational history. For the k NN-based interval analysis, temporally adjacent outlier points were merged into a single interval when they were separated by no more than 300 s; this is the “merge gap” referred to in the text.

	Noise cluster (–1)	k NN outliers
Source	HDBSCAN density clustering	k NN mean distance score
Scope	Low-density unclassified points	All 137,389 points scored
Threshold	HDBSCAN internal density	99th percentile of k NN score
Count	4694 points (3.4%)	1374 points (1.0%)
Merge gap	10-min gap	5-min (300 s) merge gap

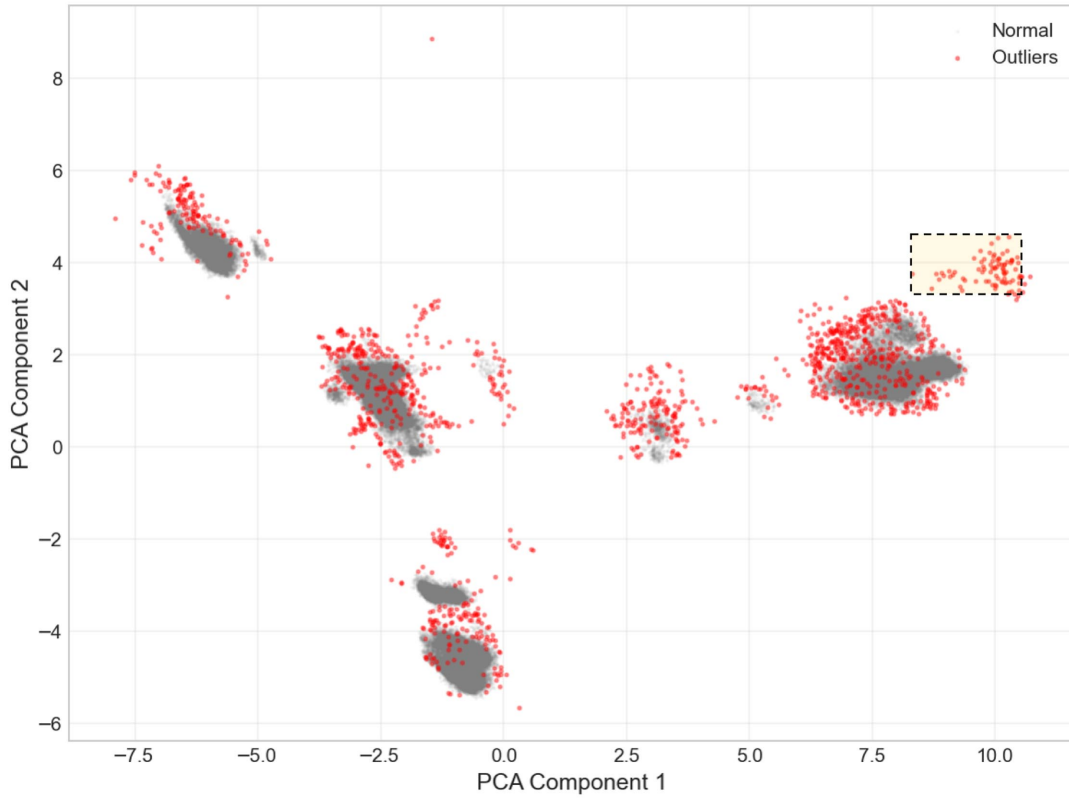


FIG. 10. Outliers (red) overlaid on the PCA projection. Flagged points lie preferentially at regime boundaries and in sparse interisland regions. The dashed box marks the area associated with the longest outlier interval (January 21, 2023).

14-month archive and merged using the same 300-s rule. Repeating this random draw 1000 times gives an average of 1351 intervals, close to the raw count of 1374 points, since timestamps scattered uniformly at random rarely fall within 5 min of one another by chance. A smaller interval count means more of the original 1374 points were merged together, i.e., a more temporally clustered set of timestamps; the observed count of 806 is therefore the more extreme outcome if outliers are genuinely clustered in time rather than scattered independently through the archive. The observed value of 806 is lower than all 1000 random draws (equivalent to an empirical permutation $p < 0.001$), confirming that the flagged outliers are genuinely clustered in time.

Cross-referencing the two screens provides a stronger operational signal than either alone: a sustained noise dwell that also contains many high-scoring k NN outliers is more likely to reflect a genuine machine excursion than either indicator by itself. To test this, we computed the overlap between the ten longest noise dwells and the k NN outlier intervals. Results are presented in Table VII. The overlap is generally sparse, with most of the longest noise dwells having little or no concurrent k NN outlier activity. One case stands out—the January 21, 2023 dwell of 94 min overlaps with outlier intervals at 100%. This episode is worth examining in detail precisely because

it demonstrates the value of the combined screen in an unexpected way. Manual inspection revealed that multiple beam position monitors and corrector setpoints returned undefined values in the archiver during this period, pointing to a data-quality failure rather than a beam-physics event. This outcome illustrates its utility. The two independent geometric screens—one based on cluster membership, the other on nearest-neighbor isolation—converged on the same 94-min window without any prior knowledge of the archiver fault. A screening tool that surfaces both unusual machine states *and* unusual data states is operationally valuable, since corrupted inputs to an embedding model can produce misleading embeddings regardless of the underlying machine behavior. The January 21 case therefore motivates a practical refinement: applying data-quality masks prior to embedding or anomaly scoring, i.e., flagging and excluding samples or channels with undefined, missing, or otherwise invalid archived values so that instrumentation failures are not misinterpreted as genuine operational excursions.

C. Historical analog retrieval

A practical advantage of any well-structured embedding is that it enables case-based reasoning. Given a query state, one can search the historical record for the geometrically

TABLE VII. Overlap analysis between the ten longest HDBSCAN noise dwells and k NN outlier intervals. Most long noise dwells have little or no concurrent outlier activity.

Noise start	Noise end	Dwell (min)	Overlap	Overlap (min)	Overlap (%)
10/06/2022 17:06	10/06/2022 21:34	268	No	0	0
10/07/2022 21:28	10/08/2022 01:04	216	No	0	0
12/14/2022 06:10	12/14/2022 09:06	176	No	0	0
06/11/2022 01:06	06/11/2022 03:50	164	Yes	8	4.9
09/28/2022 01:30	09/28/2022 04:10	160	No	0	0
08/11/2022 22:10	08/12/2022 00:32	142	Yes	6	4.2
12/14/2022 02:04	12/14/2022 03:52	108	No	0	0
11/07/2022 10:24	11/07/2022 12:04	100	Yes	10	10.0
11/03/2022 16:04	11/03/2022 17:38	94	No	0	0
01/21/2023 01:28 ^a	01/21/2023 03:02 ^a	94	Yes	94	100

^aThe January 21, 2023 entry shows 100% overlap and was traced to an archiver data-quality failure affecting multiple process variables.

nearest states and ask when the injector last occupied a comparable configuration [25,26]. To avoid trivial self-matches, neighbors within a ± 12 -h window around the query are excluded.

Figure 11 shows retrieval results for three query types. The neighbor search itself is performed in the full 16D embedding space; the PCA map is used only to visualize

the query and its retrieved neighbors in two dimensions. A query drawn from the largest cluster (cluster 9, July 12, 2022 at 20:26) returns 19 neighbors that remain tightly co-located with the query even in the 2D projection—the expected behavior for a highly repeatable operating state. A query near a large shift-change boundary (September 30, 2022 at 13:10) returns 18 neighbors that appear more

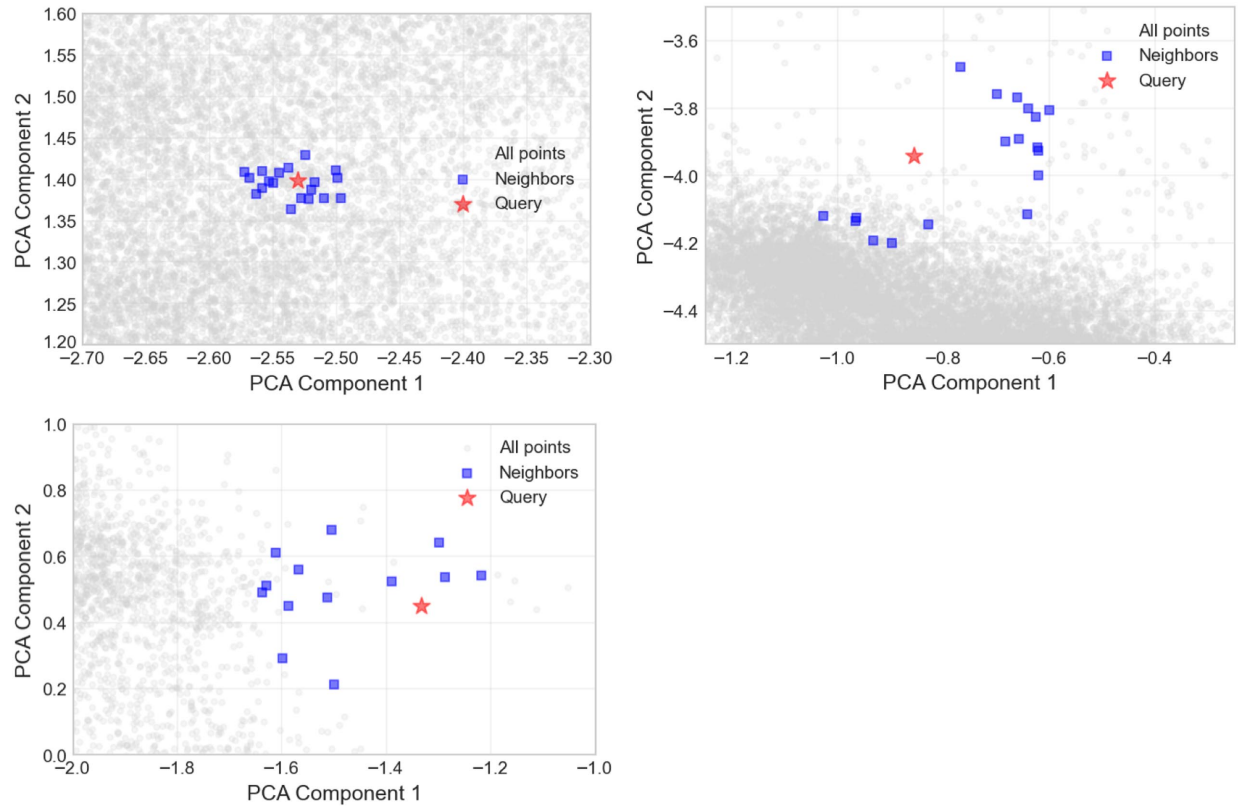


FIG. 11. Historical-neighbor retrieval for three query types: a point from a large persistent cluster (upper left), a point near a large shift-scale transition (upper right), and a k NN outlier (lower left). Retrieval is performed in the full 16D embedding space while the plot shows the query and retrieved neighbors in a 2D PCA projection. Blue markers show retrieved neighbors after excluding the ± 12 -h window around the query; stars mark query locations.

spatially dispersed in the PCA view, consistent with that state lying near a transition corridor rather than a cluster core. An outlier query (September 27, 2022 at 16:58) returns 13 neighbors in a compact set, demonstrating that geometric outlier status does not mean the state is unique: it may be rare relative to the dominant modes but still recur with enough frequency to have historical precedent.

This case-based lookup is a qualitatively different operational capability from threshold-based alarming. It does not require a predefined definition of “normal” and instead leverages the full operational archive as a reference. Section IX B extends this demonstration with a quantitative retrieval-fidelity metric, scored against ground truth independent of the embedding itself, and repeats it for two alternative representations.

VII. PHYSICAL INTERPRETATION

The fingerprint analysis (Table II) and feature-variability ranking together connect the geometric structure of the embedding back to specific machine subsystems. Two complementary variability metrics provide different perspectives. The global mean-variance ranking—which asks which features fluctuate most on an hour-to-hour basis across the entire dataset—identifies superconducting cavity settings 0L02-7 and 0L02-8 as tied for the largest mean variance (2.40), followed by dipole MBF0L06 and quadrupole MQJ0L01 among settings. Reading variability is dominated by ion pumps VIP0L03E, VIP0L03D, and VIP0L03C. The stable-vs-unstable contrast—which selects windows by the dispersion of the embedding itself, then asks which raw features expand most strongly in the windows of largest embedding-space spread relative to those of smallest spread—produces a consistent picture: 0L02-7 and 0L02-8 again rank first (variance increase 6.54), followed by dipoles MFD0I04A and MFD0I04, and corrector MQS0L01. Unlike the representation-independent dispersion analysis of Sec. VI A, this particular contrast is embedding-derived. The window selection depends on the embedding, though the resulting feature ranking is computed directly from the raw settings and readings.

The recurrence of 0L02-7 and 0L02-8 in both rankings makes them especially compelling candidates for further investigation. These cavities are part of the quarter cryomodule (0L02), which sits early in the injector lattice and directly sets the initial beam energy. Their prominence in both routine variability and in the most dynamically evolving intervals suggests that changes to their gradients or phases are a primary driver of machine-state relocations across the dataset.

The cluster fingerprints extend this interpretation to specific regimes. As discussed in Sec. IV C, cluster 2 is organized by chopper cavity settings, indicating a mode defined by early beam formation, while cluster 6 is organized by 0L04 srf cavities, pointing to downstream

configurations as the primary organizing variable. Cluster 9—the largest by point count, with a median dwell exceeding 5 days—is characterized by solenoid, capture, and corrector settings in the gun/low-energy region, suggesting a well-established, stable operating point in which the early capture optics are the primary differentiators.

These interpretations are suggestive, not definitive. Fingerprinting identifies correlates, not causes. But the consistency between the global variability rankings and the cluster-level differentiators strengthens the picture: the dominant directions of machine-state change in the embedding space correspond to physically interpretable subsystems.

VIII. CONTROLLED BEAM-STUDY VALIDATION

To test whether the embedding also reflects deliberate, operator-imposed changes, we examine a controlled beam study in which the injector was reconfigured from one stable setup to another through 62 incremental changes. These changes were applied sequentially and spanned horizontal correctors, vertical correctors, quadrupoles, and—in one multielement step—rf gradients, phases, and quadrupoles simultaneously.

The PCA projection of this study is shown in Fig. 12. Rather than appearing as disconnected scattered points, the sequence traces a smooth, interpretable trajectory through state space from the initial to the final configuration. Most steps produce local movement within a shared neighborhood: horizontal correctors, vertical correctors, and quadrupole changes all fall within a coherent region, consistent with their relatively limited influence on the global machine state. The single compound “various” step—involving simultaneous rf and magnet changes—produces a distinctly larger relocation between two separated regions.

This result serves as validation that the learned embedding is not merely organizing historical data into *post hoc* structure: it behaves as a genuine operational reference space in which deliberate reconfiguration appears as structured motion rather than as noise.

IX. DISCUSSION

A. What embeddings add beyond single-channel monitoring

Traditional single-channel monitoring detects anomalies only when a specific channel crosses a predefined threshold. This approach is well suited to hardware faults and out-of-range readings, but it cannot detect operational conditions that are unusual only in the aggregate—where individual channels each appear normal but their combination is unusual relative to the machine’s history. The embedding-based approach inverts this: it first asks whether the overall machine state is typical, and only then asks which channels are responsible for any detected

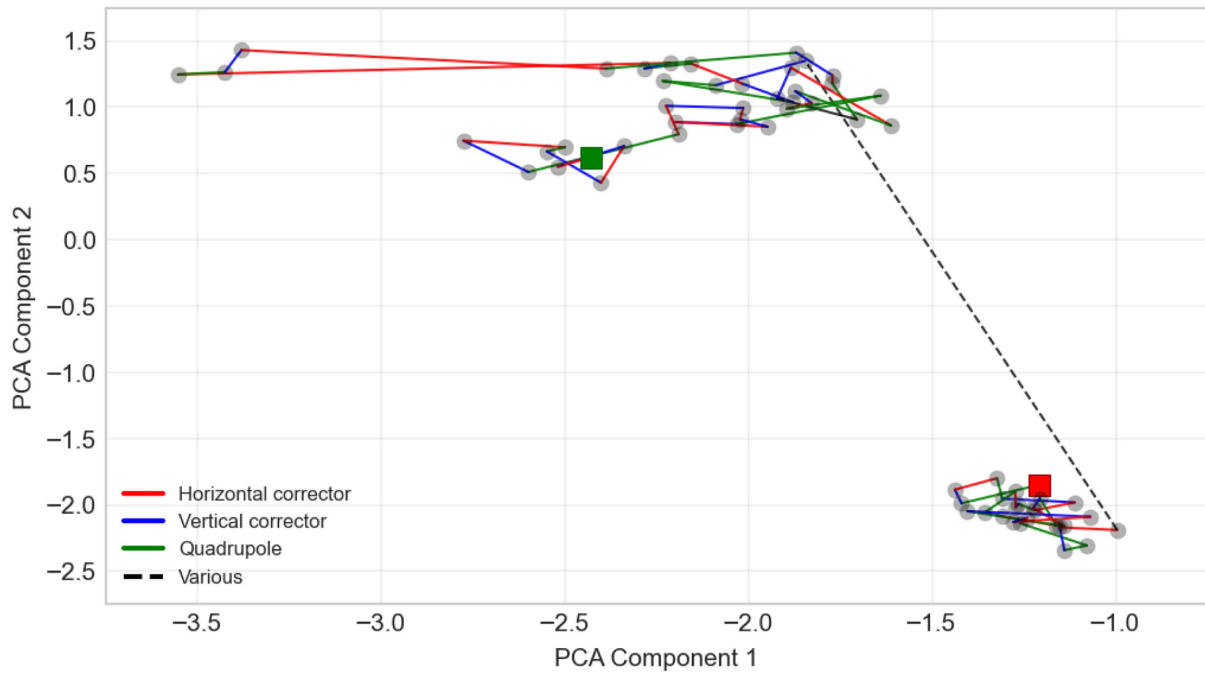


FIG. 12. PCA projection of the 62-step controlled beam study. Start and end configurations are marked by red and green squares, respectively. Edge colors denote the class of element changed; the dashed various step involves simultaneous rf gradient, phase, and quadrupole changes and produces the largest single relocation. The coherent trajectory demonstrates that the embedding responds meaningfully to deliberate reconfiguration.

deviation. The 99.6% jitter compliance rate established in Sec. VI A is not a result that could be obtained from any single channel; it reflects a reading-space dispersion computed across 126 simultaneous readbacks. Similarly, retrieval of historical analogs is inherently a whole-machine question. There is no single channel whose value uniquely specifies the injector's operational configuration, so case-based lookup requires an embedding that condenses the full state into a searchable representation.

B. Alternative representations

The GNN is not necessarily the only viable representation for organizing injector state space, so we quantified the alternative directly. We built two baselines from the standardized 393-dimensional raw feature vector (267 settings plus 126 readings): a 16-component PCA projection, capacity-matched to the GNN embedding, and the full, unreduced 393-dimensional vector. We reran every analysis in this paper against all three representations under identical hyperparameters and a common 40,000-point subsample, including clustering, regime persistence, run-period alignment, historical-analog retrieval, beam-study trajectory coherence, and stability. The subsample, rather than the full 137,389-snapshot archive, is necessary to keep the 393-dimensional baseline tractable across all three representations. Because it is smaller than the archive used for the clustering of Sec. IV A, the absolute cluster counts reported below are not directly comparable to the

ten regimes identified there; the comparison across representations, which uses the same subsample and hyperparameters throughout, is unaffected.

Two of these comparisons use metrics not introduced elsewhere in this paper. Run-period alignment is scored with the Adjusted Rand Index, which measures how well the cluster assignments produced by a representation agree with the known archive time period of each snapshot; a value of 1 indicates perfect agreement and 0 indicates no better than chance. Retrieval fidelity is scored against settings-space ground truth using two further metrics. Lift is the factor by which a representation's retrieved neighbors are closer in settings space than a random sample of the same size. Recall@20 is the fraction of a snapshot's 20 true nearest neighbors in settings space that are also among its 20 nearest neighbors under the representation being tested.

On these two comparisons, the raw-feature baselines match or exceed the GNN. Run-period alignment gives an Adjusted Rand Index of 0.58–0.61 for the PCA and raw-feature baselines, vs 0.52 for the GNN. Retrieval fidelity gives a Lift of 40 for the baselines vs 7 for the GNN, and a Recall@20 of 0.05 vs 0.00. This result is expected. A Euclidean metric over the settings vector, or a rotation of it, trivially recovers settings-space neighbors better than a representation trained for a different objective.

The comparison is decisive, however, in the one place with a ground truth that is physically caused and does not depend on any representation: the August 31, 2022 beam

study of Sec. VIII. Only the GNN embedding identifies the largest change, a compound, multicomponent various step, as its largest representation-space displacement. The raw and PCA representations both rank that same step only tenth of 62. The GNN's trajectory is also fourfold to fivefold more path-efficient, with a net displacement over path length of 0.20 vs 0.04–0.05 for the baselines, tracking the physically commanded reconfiguration far more directly. Its cluster structure is markedly cleaner as well, with 0.2% HDBSCAN noise vs 5%–6% for the baselines, concentrating the archive into five well-separated regimes rather than spreading it across 14 to 16.

Taken together, raw-feature PCA preserves nearby setpoint geometry, which is unsurprising since it is built directly from the setpoints. That is enough to explain its edge on run-period alignment and retrieval fidelity above. It does not, however, reproduce the GNN's advantage on the beam study, the one comparison with a ground truth that is physically caused and independent of any representation. This gap is the paper's central empirical argument, namely, an encoder trained to organize machine state captures something about deliberate reconfiguration that a fixed linear projection of the raw features does not.

This does not make the specific GNN architecture used here privileged. The GNN's graph is fixed in advance, so any relationship between elements that the model can exploit must already be encoded as an edge in that graph. A transformer-based encoder [27] replaces this fixed topology with self-attention, letting the model learn directly from data which elements condition one another rather than relying on a graph supplied by hand. That flexibility is the main motivation for exploring transformer-based encoders next. Incorporating non-numeric metadata, such as element type or capability, may also be more natural in that setting [28], though a hand-designed categorical encoding could in principle extend PCA the same way.

Looking further ahead, the GNN represents one point on a spectrum of representation complexity. Transformer-based encoders offer a middle ground and have seen increasing adoption in scientific machine learning [29]. Exploring such architectures as alternatives or complements to the GNN is a natural direction for future work, particularly as injector configurations evolve and the cost of maintaining a fixed-graph representation grows.

C. Limitations

Several limitations of the present study should be acknowledged. First, the GNN model is tied to a specific graph representation of the injector beamline. When the beamline configuration changes substantially—as it did during the scheduled maintenance period following March 2023, when elements were removed, added, and rearranged—the fixed-graph representation may no longer be valid, requiring retraining or fine-tuning before new

data can be analyzed consistently with historical embeddings. This retraining burden is a practical constraint on the method's long-term applicability.

Second, the clustering and outlier results are sensitive, to some degree, to the choice of hyperparameters: the HDBSCAN minimum cluster size, the k NN neighbor count, and the merge gap for interval construction. The quantitative results reported here reflect a specific set of choices, and the cluster count, dwell statistics, and outlier counts would shift under different settings. We have aimed to report robustly supported qualitative conclusions—that regimes are persistent, that large relocations are rare, that outlier episodes are brief—that are not sensitive to small parameter variations.

Third, the jitter baseline was estimated from five anchor windows chosen for their operational stability. The observed window-to-window variation in p_{95} (from 5.52 on January 4, 2022 to 19.57 on June 13, 2022) indicates that the effective noise floor is not stationary across operating contexts. In the present analysis, all 1-h windows are normalized against a single pooled baseline. If that baseline were chosen more conservatively—for example, by normalizing against a larger reference value drawn from the upper end of the anchor-window distribution—the resulting stability ratios would be smaller, and fewer windows would be flagged as unstable. That would reduce sensitivity to modest departures from the nominal jitter floor. Conversely, a context-dependent baseline, in which the reference jitter level is conditioned on operating regime, beam current range, or another state descriptor, would make the comparison more local and could improve specificity by reducing false positives caused by comparing dissimilar operating conditions to a single global reference. Determining how best to define such adaptive baselines is an important direction for future work.

Finally, the overlap analysis between noise dwells and k NN outliers demonstrated that the combined screen is sensitive not only to unusual machine states but also to data-quality failures—in this case, archiver gaps that caused multiple PVs to return undefined values. This is a useful property for a screening tool, but it underscores that manual follow-up is necessary before any flagged interval can be interpreted operationally. Data-quality filtering prior to embedding or anomaly scoring would reduce the noise in such screens.

D. Path to real-time deployment and model lifecycle

All analyses in this paper were performed retrospectively on a stationary historical dataset. The long-term goal is a real-time operational capability in which new machine snapshots are embedded continuously and the state-space view is updated live. New embeddings could be generated every 2 min, alongside updates to regime assignment, stability metrics, outlier flags, and the nearest-neighbor reference pool. Lighter-weight representations, whether

direct feature-space projections or transformer-based encoders, may offer a more maintainable path to continuous deployment than the present GNN, and benchmarking the operational tradeoffs among these approaches is an important next step. Realizing this goal requires a clear answer to how the system behaves as the machine evolves beyond its training epoch, and when the encoder itself needs to change.

The encoder analyzed throughout this paper was trained once, on the full 137,389-snapshot archive. Two properties of a new observation distinguish a genuine new operating regime from an anomalous excursion. A new but valid operating point produces snapshots that are mutually close to one another and that persist over time. A fault or transient instead produces isolated snapshots that are far from history and do not cohere with one another. Concretely, we flag a novel-regime candidate when consecutive snapshots exceed the k NN novelty threshold relative to the existing reference pool, have low mutual dispersion, and sustain that pattern for at least one shift. We flag an anomaly when novelty is high but the flagged points are also mutually dispersed, or the episode is brief.

We demonstrate this rule with a pipeline-level temporal replay. Every downstream artifact—the HDBSCAN cluster model, the k NN reference pool, and the novelty threshold—is fit using only the first 51,781 snapshots (January–August 2022). The remaining 85,608 snapshots (September 2022–March 2023) are then scored against those fitted artifacts as unseen data, exactly as they would be by a pipeline that had never been updated since the January–August fit. This bounds the temporal generalization of the deployed pipeline, not of the encoder, since the encoder itself saw the full archive. The pipeline-level quantity is nevertheless the one that governs deployment behavior.

The replay produces a sustained pattern rather than a transient one. The weekly novelty rate against the January–August-only reference pool rises from 30% in the last week of August to a sustained $\geq 95\%$ from the week of September 26, 2022 onward and remains at essentially 100% through the end of the archive in March 2023, 6 months of continuous novelty. Of the four clusters fit on the early period, three (16,142, 2920, and 505 snapshots) never reappear in the later data at all, and the fourth captures only 6.0% of the later archive. Correspondingly, 94.0% of later snapshots are left unassigned by the early-fit cluster model. Every snapshot that the k NN threshold flags as novel is also left unassigned by HDBSCAN, so the two signals never disagree in that direction. A pipeline built this way does not silently misclassify the machine’s subsequent evolution as familiar. It correctly and durably flags that evolution as outside the reference envelope, which is the desired, conservative behavior. This is exactly what motivates refreshing the fitted artifacts on a regular

cadence rather than treating a single historical fit as permanent.

The workflow has exactly one trained component, the GNN encoder. The deployed system also carries a second tier of fitted artifacts built on top of the encoder’s output space: the HDBSCAN cluster model, the k NN reference pool, the PCA projection, and the anchor-window jitter baseline. All of these can be refit in minutes as new data accumulate. In practice this means the fitted artifacts should be refreshed often and cheaply, while the encoder itself is retrained only rarely. This substantially narrows the maintenance burden of the overall system.

Encoder retraining is mandatory when the beamline topology changes, for example when elements are added, removed, or rearranged, as occurred at the March 2023 boundary discussed in Sec. IX C. It is also mandatory when the set of instrumented channels changes. Retraining is indicated, though not strictly required, when the novelty rate against the reference pool stays persistently elevated even after the fitted artifacts have been refreshed. That pattern means the learned geometry itself, not merely the downstream models built on it, no longer resolves the machine’s new operating regions. Absent these triggers, no retraining is needed.

Finally, embeddings produced by different encoder versions are not directly comparable, since each version defines its own geometry. One way to preserve continuity across retrains would be to maintain a small set of representative machine snapshots (“anchor configurations”) retained specifically for this purpose and re-embedded under each new encoder version. The resulting coordinates could then be aligned to the previous version’s frame using Procrustes analysis [30], a standard technique that finds the rotation, reflection, and rescaling that best superimposes one set of points onto a corresponding reference set. This would preserve continuity of the historical index across retrains and is a natural extension of the present framework.

X. CONCLUSION

We have shown that GNN-based machine-state embeddings provide a practical operational reference space for the CEBAF injector. The 14-month operational history organizes into a small number of persistent, well-separated neighborhoods in a 16-dimensional learned space, with ten persistent regimes identified by density-based clustering and confirmed by persistence statistics. The dominant clusters are entered only a handful of times per year and maintain median dwells measured in days. Large relocations between neighborhoods are rare and episodic—detectable as isolated spikes in the shift-to-shift step-size distribution—while 99.6% of 1-h operating windows fall within an empirically derived jitter baseline. Geometric outlier screening narrows the operational archive to a small, auditable set of intervals, and

nearest-neighbor retrieval enables case-based reasoning over the historical record. A controlled beam study validates that the embedding responds meaningfully to deliberate reconfiguration. Together, these capabilities demonstrate that machine-state embeddings support holistic, data-driven operational monitoring in ways that single-channel inspection cannot achieve. The present work establishes an analysis framework; its full value will be realized when combined with real-time data streams and integrated with supporting information sources such as logbooks, downtime records, and operational documentation.

XI. SUPPLEMENTARY METHODS: LOGBOOK ANALYSIS WORKFLOW

For the September 30, 2022 case study, logbook entries were analyzed to connect a large embedding-space relocation event to operator-recorded machine context. The goal of this workflow was not to automate root-cause attribution, but to assemble a chronologically ordered, human-authored narrative record that could be reviewed efficiently with the aid of a large language model.

Entries were retrieved via a Streamlit application that wraps the Jefferson Lab logbook REST API. For this study, the retrieval window spanned September 28–October 1, 2022, centered on the transition identified in the embedding analysis. Queries were restricted to entries of type ELOG and explicitly excluded AUTOLOG records. This distinction was important because AUTOLOG entries comprise the majority of logbook traffic but are machine generated and often contain little narrative context, whereas the purpose here was to recover operator descriptions, interventions, and observed symptoms.

Only the body text of each retained entry was used for downstream analysis. HTML markup was converted to plain text using BeautifulSoup, with paragraph and list structure preserved where possible to maintain readability. The raw collection was then filtered to retain entries likely to contain meaningful human-authored narrative. Two criteria were applied: the cleaned body text had to contain at least 50 words, and the fraction of recognizable English words had to exceed 0.70. These thresholds were chosen empirically to suppress entries consisting primarily of device names, numerical readbacks, or other low-context content, while preserving entries that described operational actions or observations.

The filtered results were normalized into JSON Lines (JSONL) format, with one record per entry. Each record contained a normalized ISO-8601 timestamp, the original timestamp string, the log number, the entry title, and the cleaned body text. Two lightweight enrichment steps were then added. First, heuristic flags were assigned by scanning the title and body for keywords associated with broad classes of operational activity, including configuration changes, fault or trip events, maintenance actions, procedural steps, and shift-summary content. These flags

were intended only as cues for later interpretation; they were not treated as authoritative labels. Second, CEBAF-specific entities were extracted using regular expressions targeting machine-relevant identifiers such as cavity and zone names, device identifiers, and hall references. As with the flags, these extracted entities were used as contextual aids rather than as ground truth. The enriched records were sorted chronologically before being passed to the language model.

The resulting JSONL file was supplied to an OpenAI-compatible chat interface. The system prompt instructed the model to act as a CEBAF operations analyst, to ground its claims in the cleaned log text, to treat flags and extracted entities as heuristic aids rather than evidence, and to reason explicitly about temporal order. The prompt also directed the model to distinguish direct textual evidence from inference and to state clearly when the available record was insufficient to support a conclusion. In this work, the analysis was run using both Claude Haiku and GPT-4o.

This workflow has several limitations. Most importantly, the filtering thresholds are heuristic. Brief but consequential entries may be excluded if they fall below the word-count cutoff, while some retained entries may still contribute little interpretive value. In addition, keyword-based flags and regular-expression entity extraction can both miss relevant context and introduce false positives. The workflow should therefore be understood as a retrieval-and-organization aid for human review, not as a source of definitive operational labels.

ACKNOWLEDGMENTS

This material is based on work supported by the U.S. Department of Energy, Office of Science, Office of Nuclear Physics under Contract No. DE-AC05-06OR23177. AI-based language tools (ChatGPT-5.4, OpenAI) were used to improve clarity and organization of the manuscript. All AI-assisted text was reviewed, edited, and verified by the authors, who take full responsibility for the accuracy and content of the manuscript.

DATA AVAILABILITY

The data that support the findings of this article are not publicly available because they contain commercially sensitive information. The data are available from the authors upon reasonable request.

-
- [1] S. Wang, C. Tennant, D. Moser, T. Larrieu, and J. Li, Graph learning for particle accelerator operations, *Front. Big Data* **7**, 1366469 (2024).
 - [2] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, A simple framework for contrastive learning of visual representations, in *Proceedings of the 37th International*

- Conference on Machine Learning* (PMLR, Cambridge, MA, USA, 2020), pp. 1597–1607, <https://proceedings.mlr.press/v119/chen20j.html>.
- [3] Y. You, T. Chen, Y. Sui, T. Chen, Z. Wang, and Y. Shen, Graph contrastive learning with augmentations, in *Advances in Neural Information Processing Systems* (Curran Associates, Inc., Red Hook, NY, USA, 2020), Vol. 33, pp. 5812–5823, <https://proceedings.neurips.cc/paper/2020/hash/3fe230348e9a12c13120749e3f9fa4cd-A-bstract.html>.
 - [4] O. Convery, L. Smith, Y. Gal, and A. Hanuka, Uncertainty quantification for virtual diagnostic of particle accelerators, *Phys. Rev. Accel. Beams* **24**, 074602 (2021).
 - [5] J. Zhu, Y. Chen, F. Brinker, W. Decking, S. Tomin, and H. Schlarb, High-fidelity prediction of megapixel longitudinal phase-space images of electron beams using encoder-decoder neural networks, *Phys. Rev. Appl.* **16**, 024005 (2021).
 - [6] M. Rautela, A. Williams, and A. Scheinker, A conditional latent autoregressive recurrent model for generation and forecasting of beam dynamics in particle accelerators, *Sci. Rep.* **14**, 18157 (2024).
 - [7] K. Rajput, M. Schram, W. Blokland, Y. Alanazi, P. Ramuhalli, A. Zhukov, C. Peters, and R. Vilalta, Robust errant beam prognostics with conditional modeling for particle accelerators, *Mach. Learn.:Sci. Technol.* **5**, 015044 (2024).
 - [8] A. Scheinker and A. Williams, Latent diffusion can map beam loss to two-dimensional phase-space projections, *Phys. Rev. Accel. Beams* **28**, 094602 (2025).
 - [9] Q. Xu, H. D. Zhang, G. Trad, A. Hill, F. Roncarolo, and C. P. Welsch, Development of radiation-tolerant beam imaging via multimode fiber and synthetic data-driven machine learning, *Phys. Rev. Accel. Beams* **29**, 032801 (2026).
 - [10] A. Edelen and X. Huang, Machine learning for design and control of particle accelerators: A look backward and forward, *Annu. Rev. Nucl. Part. Sci.* **74**, 557 (2024).
 - [11] E. Fol, R. Tomás, J. Coello de Portugal, and G. Franchetti, Detection of faulty beam position monitors using unsupervised learning, *Phys. Rev. Accel. Beams* **23**, 102805 (2020).
 - [12] P. Arpaia, G. Azzopardi, F. Blanc, X. Buffat, L. Coyle, E. Fol, F. Giordano, M. Giovannozzi, T. Pieloni, R. Prevete, S. Redaelli, B. Salvachua, B. Salvant, M. Schenk, M. S. Camillocci, R. Tomas, G. Valentino, F. F. Van der Veken, and J. Wenninger, Beam measurements and machine learning at the CERN large hadron collider, *IEEE Instrum. 'I&' Meas. Mag.* **24**, 47 (2021).
 - [13] F. T. Liu, K. M. Ting, and Z.-H. Zhou, Isolation forest, in *Proc. 8th IEEE Int. Conf. on Data Mining (ICDM)* (IEEE, New York, 2008), pp. 413–422.
 - [14] P. Wójcik and T. Barszcz, Condition monitoring using a latent space of variational autoencoder trained on healthy data, *Sensors* **24**, 6670 (2024).
 - [15] D. P. Kingma and M. Welling, Auto-encoding variational Bayes, in *Proceedings of the 2nd International Conference on Learning Representations (ICLR)*, Ithaca, NY, USA (2014), [arXiv:1312.6114](https://arxiv.org/abs/1312.6114).
 - [16] I. T. Jolliffe, *Principal Component Analysis*, 2nd ed. (Springer, New York, NY, USA, 2002).
 - [17] L. van der Maaten and G. Hinton, Visualizing data using t-SNE, *J. Mach. Learn. Res.* **9**, 2579 (2008).
 - [18] L. McInnes, J. Healy, N. Saul, and L. Großberger, UMAP: Uniform manifold approximation and projection for dimension reduction, *J. Open Source Softw.* **3**, 861 (2018).
 - [19] M. Joswiak, Y. Peng, I. Castillo, and L. H. Chiang, Dimensionality reduction for visualizing industrial chemical process data, *Control Eng. Pract.* **93**, 104189 (2019).
 - [20] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, Graph attention networks, [arXiv:1710.10903](https://arxiv.org/abs/1710.10903).
 - [21] R. J. G. B. Campello, D. Moulavi, and J. Sander, Density-based clustering based on hierarchical density estimates, in *Advances in Knowledge Discovery and Data Mining* (Springer, Berlin, Heidelberg, Germany, 2013), pp. 160–172, [10.1007/978-3-642-37456-2_14](https://doi.org/10.1007/978-3-642-37456-2_14).
 - [22] R. J. G. B. Campello, D. Moulavi, A. Zimek, and J. Sander, Hierarchical density estimates for data clustering, visualization, and outlier detection, *ACM Trans. Knowl. Discovery Data (TKDD)* **10**, 1 (2015).
 - [23] S. Ramaswamy, R. Rastogi, and K. Shim, Efficient algorithms for mining outliers from large data sets, in *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data* (Association for Computing Machinery, New York, NY, USA, 2000), pp. 427–438, [10.1145/342009.335437](https://doi.org/10.1145/342009.335437).
 - [24] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, LOF: Identifying density-based local outliers, in *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data* (Association for Computing Machinery, New York, NY, USA, 2000), pp. 93–104, [10.1145/342009.335388](https://doi.org/10.1145/342009.335388).
 - [25] A. Aamodt and E. Plaza, Case-based reasoning: Foundational issues, methodological variations, and system approaches, *AI Commun.* **7**, 39 (1994).
 - [26] J. Johnson, M. Douze, and H. Jegou, Billion-scale similarity search with GPUs, *IEEE Trans. Big Data* **7**, 535 (2021).
 - [27] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, Attention is all you need, in *Advances in Neural Information Processing Systems (NeurIPS)* (Curran Associates, Inc., Red Hook, NY, USA, 2017), Vol. 30, <https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-A-bstract.html>.
 - [28] F. Mayet *et al.*, Toward particle accelerator machine state embeddings as a modality for large language models, in *Proc. 20th Int. Conf. Accel. Large Exp. Phys. Control Syst. (ICALEPCS'25)* (JACoW Publishing, Geneva, Switzerland, 2025), pp. 1233–1238.
 - [29] Y. Gorishniy, I. Rubachev, V. Khrulkov, and A. Babenko, Revisiting deep learning models for tabular data, in *Advances in Neural Information Processing Systems (NeurIPS)* (Curran Associates, Inc., Red Hook, NY, USA, 2021), Vol. 34, pp. 18932–18943, <https://proceedings.neurips.cc/paper/2021/hash/9d86d83f925f2149e9edb0ac3b49229c-Abstract.html>.
 - [30] P. H. Schönemann, A generalized solution of the orthogonal procrustes problem, *Psychometrika* **31**, 1 (1966).