This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2026.3753456

IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Digital Object Identifier

GRAPE-UPS: Direction-Constrained Peer Anomaly Detection for Weak Batteries in Data-Center UPS Strings

SHIBO LUO[1], BIN SHEN[1,2], (Member, IEEE), XIAONING CHEN[1], YIQUAN FANG[1], and HUIFENG WANG[2]

[1]Informatization and Data Management Center, East China University of Science and Technology, Shanghai 200237, China
[2]School of Information Science and Engineering, East China University of Science and Technology, Shanghai 200237, China

Corresponding authors: Bin Shen (e-mail: shenbin@ecust.edu.cn) and Huifeng Wang (e-mail: whuifeng@ecust.edu.cn).

This work was supported by the Special Project for Improving Basic Conditions of Central Universities.

ABSTRACT The weakest unit limits the usable capacity of a series-connected uninterruptible power supply (UPS) battery string, but load and environmental variation can obscure its degradation signature. Group-Relative Anomaly Detection with Physics-Aware Evidence for UPS (GRAPE-UPS) uses each shared load event to define a robust peer reference and retains departures in physically adverse directions. Its directed peer score determines maintenance ranking, while one-class reconstruction and labeled validation support a separate, weakly supervised alert pathway. Across 15 independent simulated realizations per difficulty, GRAPE-UPS achieved mean area under the precision–recall curve (AUPRC) values of 0.904, 0.839, and 0.823 under easy, medium, and hard conditions. Its mean AUPRC exceeded that of all 18 comparators, with paired differences significant after correction across 54 comparisons. Directional and reference controls established the contribution of combining event-conditioned peers with adverse-direction evidence. On an independently generated electrochemical simulation benchmark, directed peer scoring achieved AUPRC of 0.848 under nominal conditions and 0.850–0.877 under operating shifts. Label-budget and operational stress tests characterized the calibration, sampling, and peer-comparability conditions required for screening. The results support task-separated maintenance prioritization and alerting from shared-event responses in battery strings with comparable units and a healthy majority.

INDEX TERMS UPS battery monitoring, group-relative anomaly detection, matched-control framework, direction-constrained scoring, physics-informed anomaly detection.

I. INTRODUCTION

DATA-center uninterruptible power supply (UPS) systems depend on series-connected battery strings to maintain service continuity during grid disturbances. The same discharge current passes through every unit. Consequently, the first unit to reach an unsafe voltage constrains the usable capacity of the entire string. Identifying weak units before a demand event is therefore central to preventive maintenance and backup-capacity assurance.

Online monitoring systems commonly measure unit voltage, temperature, and resistance-related indicators [1], [2]. These measurements provide direct evidence of condition, but their interpretation changes with load, state of charge, temperature, and recent cycling. An absolute threshold may therefore respond more strongly to a string-wide operating transition than to localized degradation. It also provides no general mechanism for distinguishing adverse and benign departures of equal magnitude.

Pack inconsistency and anomaly-detection studies address part of this problem by comparing units within a battery cohort. Group-relative residual monitoring suppresses collective voltage and temperature variation [3], while longitudinal Gaussian-process analysis models localized anomalies and degradation knees [4]. Their dominant decision variables, however, remain residual magnitude or predictive departure rather than a unified adverse-direction hypothesis.

For comparable lead–acid units during a shared discharge, a relative voltage depression is treated as adverse evidence, whereas an equally large elevation does not contribute to this cue. High resistance and excessive temperature rise provide adverse evidence; equally unusual low values do not indicate the same forms of weakness. A peer comparison also requires comparable excitation. For a series-connected UPS string, the shared current transition supplies this condition naturally.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2026.3753456

General multivariate time-series anomaly detectors provide increasingly expressive alternatives. Isolation methods and autoencoders model sparse outliers and nonlinear normality [5], [6]. Transformer detectors model association discrepancy, self-conditioned reconstruction, contrastive attention, reconstruction trends, and sub-adjacent relationships [7], [8], [9], [10], [11]. Frequency-aware, multi-scale, and pre-trained models further broaden the representation space [12], [13], [14], [15]. Benchmark evidence nevertheless shows that architectural complexity alone does not ensure reliable anomaly detection [16]. A generic temporal objective does not automatically construct the comparison set imposed by a common current, nor does it convert physical sign into a one-sided decision rule.

Battery-specific representation learning narrows this gap through cell-to-pack features, clustering, multi-scale analysis, variational representations, and probabilistic risk estimates [17], [18], [19], [20], [21], [22]. Structure-aware general detectors introduce temporal-topological, causal, or generative constraints [23], [24], [25]. These methods strengthen either the battery representation or the anomaly objective, but they still leave open how a shared UPS event should define the reference group, adverse alternative, and maintenance order.

Batteries in the same string receive nearly identical current excitation, so contemporaneous units form an event-conditioned matched-control group. When degraded units remain a minority, robust peer statistics estimate the expected response without a dedicated reference battery. The diagnostic question then becomes whether a battery departs from that reference in a direction consistent with weakness.

This paper develops Group-Relative Anomaly Detection with Physics-Aware Evidence for UPS (GRAPE-UPS) around that question. The method aligns responses to the shared event, builds a robust peer reference, and retains only adverse-direction evidence. The resulting directed peer score ranks batteries for maintenance. A separate reconstruction pathway contributes to alert screening only when validation data demonstrate added discriminative value.

The principal contributions are:

1) **A matched-control formulation for shared load events.** The common series current defines an adaptive comparison population for each event, converting operating context from a confounder into a control.
2) **Direction-constrained peer evidence.** Robust within-event normalization is coupled with one-sided battery physics to produce an interpretable maintenance-ranking score.
3) **A task-separated decision architecture.** Maintenance ranking remains tied to directed physical evidence, while validation-gated reconstruction supplies optional alert evidence without changing the ranking.

II. RELATED WORK

A. THRESHOLD-, TREND-, AND MODEL-BASED BATTERY MONITORING

The reference definition limits both fixed-threshold and historical-trend monitoring. Stationary battery monitoring commonly begins with voltage, temperature, conductance, or internal-resistance indicators [1], [2]. Historical trend analysis detects persistent deterioration that may not cross a fixed limit. An absolute limit must cover variation caused by load, temperature, state of charge, and battery history. A historical baseline can absorb some of this variation, but it may drift with operating practice and does not necessarily distinguish a string-wide event from a localized weak unit.

Physics-based state estimation provides a stronger reference by relating observations to an equivalent-circuit or electrochemical model. Such models can support state-of-health estimation and fault interpretation, but their reliability depends on parameter identifiability and on whether the available monitoring signals excite the modeled dynamics. GRAPE-UPS addresses the narrower screening problem of ordering units that experienced the same excitation.

Generative modeling provides a complementary route to evaluating diagnostic methods under controlled variation. Luder et al. [26] coupled electrical, thermal, aging, and fault models for lithium-ion packs and varied manufacturing, environment, loading, and service conditions.

B. STATISTICAL AND LEARNED PACK-RELATIVE DIAGNOSIS

Pack-relative methods define normality from other cells or batteries rather than from an absolute limit. Residual-based monitoring combines group-relative voltage or temperature deviations with principal-component and cumulative-sum statistics [3]. Piecewise dimensionality reduction and outlier deletion quantify cell inconsistency from voltage trajectories [27]. Robust principal-component analysis (RPCA) separates low-rank charging behavior from sparse cell-level outliers [28]. Gaussian-process monitoring adds a longitudinal model and has identified localized anomalies and degradation knees in field resistance trajectories [4].

These methods use peer structure efficiently, but their decision statistics are principally residual magnitude, distance, or outlier likelihood. They suppress common mode without generally asking whether the remaining residual has the adverse sign expected for a weak unit. RPCA-based methods also rely on a low-rank collective pattern. This is useful when the group is coherent, but the sparse component is not itself a weak-battery hypothesis. A positive voltage excursion and a negative excursion can both appear as sparse deviations.

Representation-learning methods move the comparison into a latent space. A spatio-temporal autoencoder incorporates cell-to-pack voltage differences [17]. Multi-feature clustering, two-tier screening, multi-scale decomposition with local-outlier scoring, and variational representations extend the ability to model nonlinear and time-varying inconsistency [18], [19], [20], [21]. Deep embedded clustering has been evaluated using operational battery-energy-storage data and a public dataset [29]. Probabilistic modeling further converts deviations into a risk measure [22].

Recent work broadens the diagnostic target. Minor-anomaly methods differentiate micro-short-circuit, low-capacity, and low-electric-quantity behavior by combining normalized volt-

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2026.3753456

TABLE 1. Comparison of Reference Definitions, Directional Semantics, and Outputs in Weak-Battery Screening

Method family and representative studies	Reference definition	Directional semantics	Typical output	Boundary relative to shared-event screening
Fixed threshold and historical trend [1], [2]	Fixed limit or unit-specific history	Limit sign may be explicit; peer-relative direction is absent	Threshold breach or condition trend	Operating-state variation is not controlled by contemporaneous peers
Statistical pack-relative diagnosis [3], [4], [27], [28]	Contemporaneous or historical pack statistics	Usually based on residual magnitude, distance, or sparse departure	Inconsistency score, outlier flag, or degradation trend	A unified adverse-direction hypothesis and maintenance order are not generally defined
Learned battery-pack diagnosis [17], [18], [19], [20], [21], [22], [29], [30], [31], [32]	Learned cell-to-pack, cluster, or normative representation	Implicit in the learned objective or specified for an individual task	Fault class, anomaly score, or risk estimate	Reference comparability, calibration, and ranking semantics depend on the model and task
General time-series anomaly detection [5], [6], [7], [8], [9], [10], [11], [12], [13], [14], [15], [16]	Historical normal sequences or representations	Typically uses reconstruction, association, or embedding departure	Point-level or sequence-level anomaly score	Shared-current equivalence and battery-specific adverse direction are not explicit
Multi-model battery fusion [33]	Healthy reconstruction and contemporaneous pack distributions	Reconstruction, density, and distributional departure	Fused score and hierarchical or voting alarms	No explicit common-current adverse-direction projection
Structure-regularized learning [23], [24], [25]	Structural constraints within learned representations or reconstruction objectives	Directional meaning depends on the learned task	Anomaly score or root-cause representation	Prior information enters the model or objective, rather than a fixed battery-specific peer rule
Generative battery modeling [26]	Coupled electrical, thermal, aging, and fault models	Fault behavior follows physical parameter changes	Simulated operating and fault trajectories	Provides evaluation data rather than a maintenance-ranking rule
GRAPE-UPS	Peers within the same aligned load event	Explicit one-sided projection toward weak-battery signatures	Within-event maintenance ranking and validation-gated alert	Requires comparable units, a healthy majority, and event-level directional contrast

age trends, filtering, clustering, and charge-stage persistence [30]. A multi-perspective framework integrates electrical, thermal, and aging inconsistency in an in-service battery-energy-storage system [31]. These studies show that pack diagnosis benefits from mechanism-aware features and multiple evidence channels. They also solve different tasks from weak-battery maintenance ranking. Classification or inconsistency grading does not by itself define how contemporaneous units should be ordered during a shared load event.

Zhang et al. [32] combined Gramian angular-field representations with a multiscale convolutional network for transfer diagnosis across nickel–cobalt–manganese (NCM) and lithium iron phosphate (LFP) battery data, followed by cell-level localization. Their classifier uses labeled fine-tuning on fault data generated from a background derived from real-world operation. Wang et al. [33] instead combined sparse-autoencoder reconstruction, local outlier factor (LOF), and Wasserstein-distance (WD) evidence through hierarchical and parallel rules, with vehicle and laboratory case studies.

Across the statistical and learned literature, three unresolved issues remain relevant to UPS screening. First, a peer group must be conditioned on a comparable excitation; otherwise, group differences may reflect operation rather than health. Second, the diagnostic meaning of a deviation depends on its sign. Third, alert discrimination and maintenance ranking are not necessarily served by the same score. GRAPE-UPS addresses these issues through event-conditioned peers, one-sided physical projection, and task-separated outputs.

C. GENERAL MULTIVARIATE TIME-SERIES ANOMALY DETECTION

General anomaly detection provides powerful models when the normal process can be learned from historical sequences. Isolation Forest partitions sparse observations without a parametric density model [5]. Autoencoders identify departures from a learned nonlinear manifold [6]. Transformer detectors extend this paradigm through association discrepancy, adversarial self-conditioning, dual-attention contrast, reconstruction trends, and sub-adjacent neighborhoods [7], [8], [9], [10], [11]. Frequency-patched and multi-resolution models capture heterogeneous temporal structure [12], [13]. DADA and MOMENT extend anomaly detection through general pre-training and foundation-model representations [14], [15].

Recent methods also impose structural or generative constraints. ScatterAD combines temporal and topological scattering with contrastive learning [23]. OracleAD aligns causal representations for joint detection and root-cause analysis [24]. IGAD constrains reconstruction through idempotent generation [25].

The unresolved issue for a series string is not only representation capacity. A generic detector asks whether one trajectory departs from a learned normal distribution. UPS screening instead asks whether one unit responds worse than contemporaneous units receiving the same current. If the shared event is not represented explicitly, the model must infer this equivalence from data and may use capacity to model common-mode variation that can be removed directly. General methods remain useful as auxiliary evidence, particularly when absolute scale or within-event temporal shape is informative. In this work, they are evaluated as external comparators; within GRAPE-UPS, reconstruction evidence is used only in the validation-gated alert branch.

D. RESEARCH GAP AND POSITIONING OF GRAPE-UPS

The gap addressed here is an explicit connection between a shared-event reference, an adverse-direction hypothesis, and a maintenance ranking (Table 1). In GRAPE-UPS, physical knowledge enters the comparison set and scoring rule:

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2026.3753456

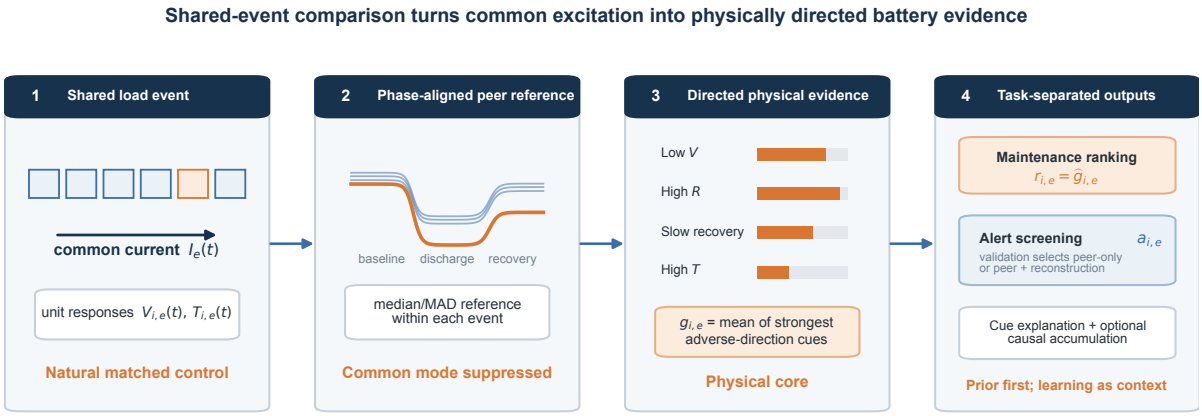


FIGURE 1. Schematic of the GRAPE-UPS architecture. Event-aligned peer normalization produces the directed physical score $g_{i,e}$, whose healthy-reference normalized component $\hat{g}_{i,e}$ defines the maintenance ranking. A decoupled reconstruction pathway contributes auxiliary alert evidence under validation-gated selection.

common current defines the peers, robust centering removes event-level shifts, and the one-sided projection retains adverse deviations. It does not impose an electrochemical equation residual on the reconstruction loss or learn a graph to define the peer group. This differs from the representation-level structural constraints described above. The directed score retains a fixed meaning for maintenance ranking, while reconstruction affects alerting only when selected by validation. The reference and direction controls examine these design choices separately from model capacity.

III. GRAPE-UPS METHOD

A. PROBLEM FORMULATION AND DESIGN RATIONALE

Consider a string of N batteries observed during load event e. The record contains the shared current $I_e(t)$ and unit-level voltage and temperature responses:

$$\mathcal{X}_e = \{I_e(t), V_{i,e}(t), T_{i,e}(t)\}_{i=1}^{N}. \tag{1}$$

The battery–event pair (i, e) constitutes the atomic decision unit. The method makes three explicit assumptions. First, units within a string experience the same series current during an aligned event. Second, the units are sufficiently comparable for a common response to be meaningful. Third, degraded units remain a minority in the peer group. Under these conditions, contemporaneous peers estimate the expected event response without requiring a dedicated reference battery.

Physical knowledge enters the detector through three fixed elements: the shared-event reference, phase-resolved descriptor windows, and prespecified adverse directions. Event-conditioned peer normalization aligns each battery to the current transition, extracts phase-resolved electrical and thermal descriptors, and standardizes them against robust peer statistics computed within the same event. This operation reduces common-mode variation induced by load and thermal state. Direction-constrained evidence selection then applies one-sided physical constraints that admit only deviations in the adverse direction.

The resulting raw directed statistic $g_{i,e}$ has a training-normalized component $\hat{g}_{i,e}$ that ranks units for maintenance prioritization (Section III-C). Reconstruction-based alerting is a separate supplementary pathway and is used only when validation data support its discriminative contribution. Fig. 1 summarizes this task-separated architecture and information flow.

B. DIRECTION-CONSTRAINED GROUP-RELATIVE EVIDENCE

Each event is aligned to the shared current transition, with discharge onset at $t = 0$ and duration d. Voltage, temperature, and current are linearly interpolated onto the 128-point grid: 32 points over $[-5, 5]$ s, 32 over $[5.5, 60]$ s, and 64 over $[61, 240]$ s. Interpolation uses finite observations in time order and holds the nearest endpoint outside the observed interval. Table 2 defines the scalar inputs; window means and extrema use this grid. Interpolation can make a descriptor numerically available without resolving its original observation window, a distinction examined in Section IV-D.

RSSI and UPS load percentage are generated communication and operating covariates in the primary simulator. They enter both feature reconstruction models, but not the directed peer score. The electrochemical stress suite has no wireless-link or UPS-load-percentage model; these two inputs therefore use the respective primary-simulator healthy-training medians.

Missing scalar descriptors are also median-imputed, and feature and sequence standardization parameters are fitted on healthy training samples. Empty averaging windows and an under-supported slope fit give missing descriptors; the resistance and recovery-ratio denominators are numerically floored as specified in the table. A very small current step or sag can still make these proxies unreliable, even when a finite value is returned.

Let $x_{i,e,k}$ denote feature k for battery i in event e. The event-

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2026.3753456

TABLE 2. Scalar Feature Dictionary and Reconstruction Inputs. Time Is in Seconds Relative to Discharge Onset.

Feature (unit)	Definition or source	q_k
Pre-event voltage (V)	V_{b}: mean voltage over $[-5,-1]$	-1
Early voltage (V)	V_0: mean voltage over $[0,3]$	–
Voltage sag (V)	$S = V_{\mathrm{b}} - V_0$	$+1$
Current-step resistance (mΩ)	$1000S/\max(\lvert I_{\mathrm{b}} - I_0\rvert, 10^{-6})$; current means use the same baseline and early windows	$+1$
Discharge slope (V/s)	Linear least-squares slope on $[10, \min(d-5, 60)]$, requiring at least three grid points	-1
Minimum voltage (V)	$V_{\min}$: minimum over $[0,d]$	-1
Recovery voltage (V)	V_{r}: mean over $[d+20, d+40]$	-1
Recovery ratio	$(V_{\mathrm{r}} - V_{\min})/\max(S, 10^{-6})$	-1
Pre-event temperature (°C)	T_{b}: mean over $[-5,-1]$	–
Temperature rise (°C)	Maximum temperature over $[0, \min(240, d+60)]$, minus T_{b}	$+1$
Mean RSSI (dBm)	Event mean of the received signal strength indicator (RSSI) for the unit's wireless measurement link	–
Missing-sample fraction	$\max(0, 1 - n_{\mathrm{received}}/n_{\mathrm{expected}})$ for the unit's battery records	–
Four event covariates	$\lvert I_0\rvert$ (A), d (s), ambient temperature (°C), and UPS load percentage from event metadata	–
Sixteen peer features	Robust deviation and within-event percentile rank for each of the eight signed features above; tied ranks are averaged	–
Three historical descriptors	Current-step resistance, pre-event voltage, and temperature rise minus each unit's own preceding-three-event median	–

The first 12 descriptors and four event covariates enter both reconstruction models (16 inputs); the contextual model also uses the 16 peer features and three historical descriptors (35 inputs). Only the eight features with a signed q_k directly enter the directed peer score.

specific peer center and robust scale are

$$m_{e,k} = \mathrm{median}_j\, x_{j,e,k}, \qquad \sigma^{\mathrm{rob}}_{e,k} = \mathrm{MAD}^*_j(x_{j,e,k}), \quad (2)$$

where MAD^* is the median absolute deviation multiplied by 1.4826, the conventional factor that matches its scale to the standard deviation under a normal reference distribution. Scalar-descriptor scales have a numerical lower bound of 10^{-6}. The peer-relative deviation is

$$z^{\mathrm{rel}}_{i,e,k} = \frac{x_{i,e,k} - m_{e,k}}{\sigma^{\mathrm{rob}}_{e,k}}. \quad (3)$$

The median–MAD pair adapts to the load and thermal state of each event while limiting the influence of a minority of degraded units. The same normalization extends to aligned voltage and temperature sequences at each time point, with scale floors of 0.003 V and 0.03 °C, respectively.

The direction map specifies adverse departures for comparable lead–acid units during a shared discharge and was fixed before held-out evaluation. The signs are working hypotheses for the phase-resolved responses in Table 2; their dependence on physical conditions is examined in Section V-H. Increased internal resistance produces a larger instantaneous voltage depression and accompanies capacity loss. Degraded UPS cells have also been observed to combine higher resistance with lower float voltage and capacitance [2]. Discharge-related heating increases with ohmic and polarization losses [34], [35]. Each of the eight directed peer features carries an adverse-direction coefficient $q_k \in \{-1, +1\}$. Low values of pre-event voltage, discharge slope, minimum voltage, recovery voltage, and recovery ratio are adverse; high values of voltage sag, resistance, and temperature rise are adverse. The positive evidence retained for feature k is

$$d_{i,e,k} = \max\left(0, q_k z^{\mathrm{rel}}_{i,e,k}\right). \quad (4)$$

This directional filter discards deviations opposite to the specified adverse direction, irrespective of their magnitude. The directed peer score aggregates the three most adverse cues:

$$g_{i,e} = \frac{1}{3} \sum_{k \in \mathrm{Top3}(d_{i,e})} d_{i,e,k}. \quad (5)$$

Top-K aggregation with $K = 3$, fixed before held-out evaluation, keeps the score focused on a small set of dominant cues while avoiding a single-cue maximum. It also reduces dilution by features that are inactive for a particular mechanism. The selected cues form the reported explanation set for the ranking decision. The aggregation-width analysis reported below evaluates this design choice while retaining the primary $K = 3$ configuration.

The additive model clarifies the role of event conditioning:

$$x_{i,e,k} = b_{e,k} + \delta_{i,e,k} + \epsilon_{i,e,k}, \quad (6)$$

where $b_{e,k}$ is the event-level common component, $\delta_{i,e,k}$ is the unit-specific departure, and $\epsilon_{i,e,k}$ is residual variation. Robust peer centering reduces $b_{e,k}$. The one-sided projection then retains alternatives whose signs agree with the weak-battery hypothesis. The transformation changes an absolute measurement into an event-conditioned directed statistic.

C. DECOUPLED RECONSTRUCTION AND DECISION ARCHITECTURE

Certain degradation patterns manifest as subtle temporal distortions or depend on the absolute operating scale rather than on peer-relative departure alone. To capture such evidence, GRAPE-UPS trains two compact autoencoders exclusively on healthy battery–event samples. The contextual model uses five sequence channels (voltage, temperature, current, and peer-normalized voltage and temperature) and 35 features combining absolute descriptors and event covariates, peer-relative statistics, and cross-event self-history. The absolute model uses the first three sequence channels and the 16 absolute descriptors and event covariates, without peer-relative or historical inputs (Table 2).

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2026.3753456

The contextual feature branch additionally receives deviations of dynamic (current-step) resistance, pre-event voltage, and temperature rise from the median of the same battery's preceding three events. Only earlier events contribute to this reference; if fewer than three are available, all available events are used, and missing history gives a zero deviation. These deviation features are termed historical descriptors in the replacement analysis. They are not inputs to the absolute autoencoder or the sequence channels, and their state is separate from the score accumulator.

Reconstruction errors and the directed peer score are standardized against their healthy training distributions and clipped below zero before fusion. The peer component is $\widehat{g}_{i,e} = \max(0, (g_{i,e} - m_g)/s_g)$, where m_g is the median of the healthy-training scores and $s_g = \max(1.4826\ \mathrm{MAD}(g), 10^{-8})$ is their robust scale. The numerical floor prevents division by zero or a near-zero scale. Unlike the raw statistic $g_{i,e}$, this component can tie values below the training median. Unless explicitly identified as the raw statistic g, directed peer scoring below refers to $\widehat{g}$; the raw-statistic evaluations are identified in their protocols. The five nonnegative components enter a convex combination:

$$s_{i,e} = \sum_{r=1}^{5} w_r c_{i,e,r}, \qquad w_r \geq 0, \qquad \sum_{r=1}^{5} w_r = 1. \tag{7}$$

The five components are contextual sequence reconstruction error, contextual feature reconstruction error, directed peer evidence, absolute sequence reconstruction error, and absolute feature reconstruction error. Fusion weights were selected to favor both high average validation AUPRC and consistent performance across the three difficulty levels. The selection objective subtracts 0.15 times the across-difficulty standard deviation from the mean AUPRC. The coefficient 0.15 is a fixed trade-off parameter; the standard deviation uses the three difficulty values with a divisor of three. A development ablation on realizations disjoint from the primary method-comparison data excluded a standalone drift component. The remaining weights were renormalized and fixed before held-out evaluation. The separate label-budget analysis in Section IV-C repeats selection with the same objective on dedicated development data. It does not replace these primary weights.

GRAPE-UPS enforces a structural separation between ranking and alerting. The maintenance ranking is defined exclusively by directed peer evidence:

$$r_{i,e} = \widehat{g}_{i,e}. \tag{8}$$

The alert score is selected through a validation-driven binary choice between the normalized peer component and the fused score:

$$a_{i,e} \in \{\widehat{g}_{i,e}, s_{i,e}\}. \tag{9}$$

The candidate maximizing validation AUPRC is retained, with ties resolved in favor of the peer score. A binary alert threshold is subsequently calibrated on the same validation partition. The fixed fused score $s_{i,e}$ and the validation-selected alert score $a_{i,e}$ are therefore distinct experimental quantities. The former tests unconditional use of reconstruction, whereas the latter tests whether validation gating controls negative transfer.

The autoencoders are trained under a one-class regime using only healthy samples. Labeled validation samples are used to choose the fusion, alert head, and threshold. GRAPE-UPS is therefore a weakly supervised framework combining one-class representation learning with labeled validation. Held-out external labels are used only for evaluation, after the architecture and primary fusion weights have been fixed; each realization retains its own validation partition for alert-head and threshold calibration.

D. TEMPORAL EVIDENCE ACCUMULATION

Gradual degradation mechanisms and intermittent faults produce weak single-event signatures whose directional evidence may fall below the detection threshold in any individual discharge. Temporal evidence accumulation draws on an information source distinct from that used by reconstruction: while the reconstruction pathway combines event responses with short self-history descriptors, accumulation smooths a specified alert score across successive discharges.

Let $h_{i,e}$ denote the input alert score, such as $a_{i,e}$ or $s_{i,e}$. The same causal recursion applies to either input:

$$\widetilde{h}_{i,e} = \alpha h_{i,e} + (1-\alpha)\widetilde{h}_{i,e-1}, \qquad \widetilde{h}_{i,1} = h_{i,1}. \tag{10}$$

The reported accumulation analyses use $h_{i,e} = s_{i,e}$, as specified in Table 3, holding the score pathway fixed to isolate the contribution of temporal smoothing from alert-head selection. The decay factor was selected as $\alpha = 0.80$ on a dedicated longitudinal calibration partition and then held fixed for the longitudinal held-out evaluations. The accumulator uses only the current and preceding events. It can strengthen persistent evidence, but it does not force a monotonic score; a sequence of lower subsequent scores can reduce the accumulated value.

E. END-TO-END DECISION PROCEDURE

Algorithm 1 summarizes the separation between offline calibration and online scoring. All decisions involving labels occur before held-out evaluation. The online stage requires the current event measurements and covariates in Table 2, each unit's preceding descriptor history and accumulated score, and the fixed preprocessing, reconstruction, fusion, and threshold parameters.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2026.3753456

TABLE 3. Experimental Data, Score Variants, and Evaluation Questions

Experiment (results section)	Data and replication	GRAPE-UPS score or evidence	Question
Method and head comparisons (V-A)	Primary suite: 15 realizations/difficulty for 19 methods; 5 for head analysis; pooled 20 for 18-method sensitivity	Selected alert a; head analysis also compares $\widehat{g}$ and s	Detection comparison and task separation
Reference–direction factorial (V-B)	Same 15 realizations/difficulty	Raw g and three reference/direction controls	Joint contribution of reference and sign
Direction and cross-simulator tests (V-C)	Separate primary suite: 5/difficulty; four-domain electrochemical suite: 5/domain	Raw g, with perturbations or a symmetric control	Directional specificity and generator transfer
Aggregation width (V-D)	Primary five-realization structural subset/difficulty	$\widehat{g}$ at $K = 1, 3, 5, 7$	Sensitivity to the number of cues
Label availability (V-E)	Structural subset; separate development pairs, 20 label draws/condition	Static head selection between $\widehat{g}$ and s; complete selection also evaluates EWMA(s)	Label size, composition, and selection dependence
Peer contamination (V-F)	Five development realizations with controlled contamination	Raw directed top-three scores with median/MAD or mean/SD references	Reference stability under contamination
Faults, cues, and accumulation (V-G)	Separate primary suite: 5/difficulty; separate longitudinal suite: 5/difficulty	Fault coverage: s; cues: top-three d_k; longitudinal: s versus EWMA(s)	Mechanism coverage, explanation, and persistence
Physical domains (V-H)	Full-model electrochemical suite: 10 paired realizations across 16 domains	Raw g and symmetric control	Fixed-direction transfer across physical conditions
Operational stress (V-I)	Full-model electrochemical suite: 10 realizations across four tests	$\widehat{g}$ and fixed s; replacement uses EWMA(s)	Sampling, telemetry, background, and replacement effects

g: raw directed statistic; $\widehat{g}$: training-normalized, zero-clipped peer component; s: fixed fusion; a: validation-selected alert. EWMA denotes an exponentially weighted moving average; EWMA(s) denotes $\widetilde{h}$ in (10) with $h = s$. Other methods retain their own scoring rules. Realization counts refer to the stated suites; label draws within a fixed development pair are not independent realizations.

Algorithm 1 GRAPE-UPS Calibration and Event-Level Inference

Require: Healthy events $\mathcal{D}_{\text{train}}$, labeled calibration events $\mathcal{D}_{\text{val}}$, ordered event record $\mathcal{X}_e$
Ensure: Maintenance ranking, alert decision, and top-three cue set
Fit feature imputers and feature/sequence standardizers on $\mathcal{D}_{\text{train}}$
Train contextual and absolute autoencoders on healthy samples
Compute directed peer and reconstruction scores on $\mathcal{D}_{\text{val}}$
Select nonnegative fusion weights by stability-regularized validation
Select the alert head from $\widehat{g}_{i,e}$ and fixed fusion; prefer peer on a tie
Calibrate the alert threshold and set $\alpha = 0.80$
Detect the shared current transition and align event phases
Extract descriptors and causal self-history for every unit (Table 2)
Compute within-event median and MAD references
Compute $d_{i,e,k} = \max(0, q_k z^{\text{rel}}_{i,e,k})$ and top-three mean $g_{i,e}$
Normalize and clip $g_{i,e}$ using healthy training scores; output $r_{i,e} = \widehat{g}_{i,e}$ for ranking
Compute $s_{i,e}$ and selected $a_{i,e}$ by (7) and (9)
For accumulation, set $h_{i,e} = s_{i,e}$ (Table 3) and update $\widetilde{h}_{i,e}$ by (10)
Apply the frozen threshold calibrated for the evaluated score

IV. EXPERIMENTAL DESIGN

Table 3 relates each experiment to its data, scoring rule, and purpose.

A. SIMULATION DATASETS

The study originated from monitoring practice on an operating UPS battery string. The available record covers one load event at 40 battery positions, with a median sampling interval of 10 s (0.1 Hz) and gaps up to 110 s, and contains no independent battery-health labels. Detection performance is therefore evaluated using controlled simulations. Nominal heterogeneity, physical domains, and mixed backgrounds specify controlled conditions rather than fitted distributions of field measurements, manufacturing batches, or aging. The primary suite uses 0.5 Hz sampling; the separate sampling analysis in Section IV-D spans 10 Hz down to 0.1 Hz and examines the loss of event-window information at this coarser interval.

Primary physics-informed suite: The primary dataset supports controlled comparison across degradation mechanisms and operating difficulty. Each string contains 40 battery units and 30 load events. Every realization uses separate strings for healthy-reference learning, labeled validation, and held-out testing. The suite comprises 20 independently seeded realizations at each of three difficulty levels, totaling 216,000 battery–event samples across the three strings and difficulty levels. For matched comparison, all 19 methods are evaluated on the same 15 realizations per difficulty (162,000 samples across training, validation, and testing). The remaining five realizations per difficulty (54,000 samples) support structural and fixed-score sensitivity analyses. An extended sensitivity analysis pools all 20 realizations for the 18 methods with complete coverage; SAE-LOF-WD is evaluated only on the common 15-realization subset.

All labels are synthetic. In both the common 15-realization comparison and the five-realization structural subset, each held-out realization contains 1,200 battery–event samples, including 55, 99, and 110 anomalies under easy, medium, and hard conditions. The corresponding all-anomaly prevalences are 4.58%, 8.25%, and 9.17%. The easy condition contains five core mechanisms: elevated resistance, capacity loss, slow recovery, thermal drift, and sensor offset. The medium and hard conditions additionally contain voltage drift, intermittent voltage drop, accelerated aging, and sulfation, giving nine mechanism families across the suite.

Table 4 summarizes the primary generator settings. The nominal values represent 12 V, 100 Ah valve-regulated lead–acid (VRLA) units, and the degradation directions follow established lead–acid monitoring evidence [1], [2], [34], [35]. The single-event record does not establish population distributions of unit variability, fault severity, or measurement disturbances. These factors were specified as controlled test conditions before generating the evaluation realizations.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2026.3753456

TABLE 4. Frozen Settings of the Primary Physics-Informed Synthetic Generator

Component	Frozen setting	Experimental role and provenance
Cohort and split	Three 40-unit strings and 30 events per realization; separate healthy-reference, labeled-validation, and held-out strings	Prevents unit- or event-level overlap among model fitting, configuration selection, and final evaluation
Nominal unit heterogeneity	Float voltage $13.63 + \mathcal{N}(0, 0.022^2)$ V; resistance $\mathrm{clip}[\mathcal{N}(9.10, 0.48^2), 7.7, 10.5]$ mΩ; capacity $\mathrm{clip}[\mathcal{N}(100, 2.8^2), 92, 106]$ Ah; temperature offset $\mathcal{N}(0, 0.18^2)$ °C	Fixed battery-to-battery variation at the nominal simulation scale
Shared load event	Load levels 25%, 45%, and 65%; discharge current $\{22, 38, 52\} + \mathcal{N}(0, 1.3^2)$ A, respectively; 120 s discharge	Creates repeated common-current excitation while varying operating level across events
Acquisition window	0.5 Hz samples from 30 s before the current transition through 120 s of discharge and 120 s of recovery; 136 nominal points	Matches the event-aligned input used by all methods before resampling to 128 points
Measurement and communication variation	Voltage noise $\mathcal{N}(0, 0.0045^2)$ V and temperature noise $\mathcal{N}(0, 0.025^2)$ °C; packet loss 2%, 5%, and 10%; nonnegative half-normal upload-latency scale 0.3, 0.5, and 1.0 s for easy, medium, and hard conditions	Separates sensor/communication difficulty from the underlying physical response
Core fault perturbations at severity γ	Resistance: $+28\%\gamma$; low capacity: capacity $-28\%\gamma$ and resistance $+25\%\gamma$; slow recovery: time constants multiplied by $1 + 3.2\gamma$; thermal drift: $+1.4\gamma$ °C; sensor offset: -0.13γ V	Prescribed mechanism-specific interventions; easy severity is capped at 0.70 and medium/hard severity at 1.00
Additional medium/hard mechanisms	Voltage drift: up to -0.08 V with event age; intermittent drop: $-\mathcal{U}(0.05, 0.25)\gamma$ V with probability 0.15γ; accelerated aging: capacity $-35\%\gamma$, resistance $+40\%\gamma$; sulfation: capacity $-22\%\gamma$, resistance $+20\%\gamma$, and recovery time constants multiplied by $1 + 4.5\gamma$	Expands the hard regimes with gradual, sparse, and coupled degradation signatures

Independent electrochemical benchmark: A second dataset reduces coupling between fault generation and detection. It is constructed with the PyBaMM lead–acid framework [36], the lead–acid cell parameters of Sulzer et al. [37] (PyBaMM parameter set `Sulzer2019`), and a leading-order quasi-static model [38]. Faults are instantiated by perturbing latent physical parameters for acid inventory, active material, interfacial kinetics, and electrode porosity. The generator does not use the GRAPE-UPS feature map or score definition. The benchmark spans nominal, load-shift, temperature-shift, and combined-shift domains, with five realizations per domain. Each has 120 anomalies among 1,200 held-out battery–event samples. Across all domains, 2,400 of 24,000 samples are anomalous (10%), giving a constant-score AUPRC reference of 0.1000. Its standalone peer evaluation uses the raw directed statistic g, without the component-level normalization or clipping used by the operational head. All four fault families are retained without ex post removal, including the low-performing electrode-porosity condition.

B. BASELINES AND EVALUATION DISCIPLINE

- **Classical anomaly detection and operational rules:** Fixed voltage and resistance thresholds, Robust-MAD, PCA reconstruction, and Isolation Forest [5] define the reference performance of comparatively simple detectors.
- **Nonlinear reconstruction:** A feature-level autoencoder (Feature-AE) [6] captures nonlinear normative structure in the physical descriptor space.
- **Battery-specific pack-relative methods:** STAE [17] and MSKLF [20] encode cell-to-pack relationships through spatio-temporal and multi-scale representations. SAE-LOF-WD [33] combines sparse-autoencoder reconstruction, local outlier factors, and Wasserstein distances. All three are adapted to the common battery–event interface.
- **Modern sequence and foundation models:** The evaluation includes eight configurations based on CATCH [12], RTdetector [10], Sub-Adjacent Transformer [11], DCdetector [9], TimeMixer++ [13], zero-shot DADA [14], and MOMENT-1-small [15] in zero-shot and frozen-embedding adapter forms. Methods requiring target-domain training use only the healthy-reference partition. Zero-shot methods retain fixed public checkpoints.
- **Regularized and topology-aware detectors:** IGAD-2025-AE [25] constrains reconstruction via idempotent generation; ScatterAD [23] integrates temporal and topological scattering with contrastive objectives.

All target-domain-trained methods use the same healthy-reference string. Hyperparameters, point-to-event aggregation, alert heads, and thresholds are selected on the labeled validation string before external-test evaluation. Point-level outputs are aggregated into battery–event scores. The directed peer score is a GRAPE-UPS ablation rather than an external comparator.

STAE combines cell-to-pack voltage differences, temporal convolutions, and cross-battery attention; MSKLF combines multi-scale voltage descriptors, kernel dimensionality reduction, event-wise group deviation, and local-outlier scoring. SAE-LOF-WD uses the common 40-unit, 128-point event input, a 64–32–64 sparse autoencoder (50 epochs, batch size 64, L1 coefficient 0.05; ReLU hidden units, linear output, Adam with a learning rate of 0.001), and 15-neighbor LOF. Voltage is standardized using healthy-training statistics.

LOF and Wasserstein distances use 50-sample windows with stride 20 and an end-aligned final window; the WD component compares unit voltage quantiles with the pack median quantile function. These windows span approximately 40–139 s on the event grid. Reconstruction residuals and window-averaged scores give unit-level event scores.

The three channels are averaged after $u/(u + \delta)$ normalization, where u is a component score, $\delta = \mu + k\sigma, k = 6, 20, 18$, and μ, σ are computed from healthy-training scores within three scaled MADs of their median (MAD scale 1.4826). This event-level fusion replaces the source method's hierarchical alarm logic. The adapted methods are compared on the UPS task rather than on their source benchmarks.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2026.3753456

C. EVALUATION METRICS AND SENSITIVITY ANALYSES

AUPRC is the primary metric for the minority anomaly class. The main comparison uses all-anomaly labels; weak-battery detection is evaluated separately. Event-wise normalized discounted cumulative gain (NDCG) measures maintenance ordering, with area under the receiver operating characteristic curve (AUROC) and threshold-dependent metrics as secondary measures. For a given mechanism, one-vs-normal AUPRC treats samples carrying that mechanism as positive and healthy samples as negative; samples from other fault mechanisms are excluded.

Independently seeded realizations are the replication units (Table 3), and difficulties are analyzed separately. Events, positions, and reused profiles are not independent replicates; repeated label draws are summarized separately.

Method means have descriptive 95% t-intervals. Paired AUPRC differences use 100,000-resample percentile bootstrap intervals and exact two-sided sign-flipping tests over 2^{15} assignments, assuming independent, null-symmetric differences. Holm correction covers the 54 GRAPE-UPS comparisons across methods and difficulties; a two-sided sign test checks sensitivity. The 18 methods with complete 20-realization coverage support an extended 51-comparison analysis. Intervals describe simulation variability and are not simultaneous effect intervals.

Online scoring measurements use in-memory events on a single thread of a consumer-grade CPU (Apple M5 Pro), excluding acquisition, I/O, initialization, training, and threshold comparison. Each reported latency is the median of three process medians (five warm-up calls, then 30 measured calls, per process); reported memory is the maximum whole-process resident memory across completed one- and eight-thread runs.

Reference and direction: A 2×2 comparison crosses the reference (healthy-training or contemporaneous within-string median/MAD) with the transformation ($|z|$ or $\max(qz, 0)$). All cells share the eight descriptors, 10^{-6} scale floor, and top-three averaging, with separately calibrated validation thresholds. The peer-plus-directed cell is raw g, without component normalization or clipping. All-anomaly and weak-battery outcomes on the 15 matched realizations are summarized with paired effects and reference-by-direction interactions, using descriptive t-intervals.

Directional selectivity: Paired perturbations inject adverse-aligned, reverse-direction, and common-mode deviations into otherwise healthy battery responses, with decision thresholds fixed at the 95th percentiles of their clean-response score distributions. The directional selectivity index, defined as aligned recall minus reverse response rate, quantifies the method's capacity to discriminate between diagnostically relevant and irrelevant departures of identical magnitude.

Cross-simulator transfer: The independent electrochemical benchmark evaluates whether the fixed direction map remains useful beyond the data-generating process of the primary suite. It introduces load, temperature, and combined operating shifts and instantiates faults through a separate generative process. Transfer performance tests portability of the directional prior; it does not establish field generalization.

Leave-one-family-out calibration: A fault family is removed from the labeled validation partition used for head selection and threshold calibration; the resulting configuration is then evaluated on an independent string containing that held-out family. This fixed-score analysis changes alert-head and threshold calibration only. The complete selection-stage analysis below additionally removes labels from weight and accumulation-factor selection.

Auxiliary analyses examine boundary conditions. Peer-reference robustness is tested by increasing the affected fraction from isolated units to 40% of the string. Median/MAD and mean/standard-deviation references use identical direction rules and healthy-validation thresholds. Cue consistency is assessed against a mechanism–cue map fixed before held-out analysis, with label permutation defining a chance-alignment baseline. Temporal accumulation compares fixed fusion with its EWMA on a dedicated longitudinal suite containing voltage drift, intermittent drops, and slow recovery.

Aggregation-width sensitivity: The normalized directed peer score was recomputed for $K \in \{1, 3, 5, 7\}$ on the structural subset in Table 3, holding all other fitted components and decision settings fixed. AUPRC and event-wise NDCG were computed within each realization and summarized by difficulty, with paired differences relative to the primary $K = 3$ configuration.

Label availability and selection: The fixed-score analysis varies labels for static head and threshold calibration on the five-realization structural subset, retaining trained models and weights without accumulation. Each realization uses 20 nested draws of 10%, 25%, 50%, or 100% of the 40 validation trajectories. A second branch supplies one, two, four, or all known-fault trajectories together with all healthy trajectories. Results are averaged within each realization; single-class draws count as calibration failures.

Complete selection is evaluated on a separate development pair at each difficulty. One realization selects weights and accumulation factors; the other calibrates heads and thresholds and supplies the external evaluation string. Twenty paired label draws are tested at each budget, with healthy training fixed at 720 samples. Weight selection uses the mean-minus-0.15-standard-deviation objective in Section III-C and 40,008 candidates. Accumulation factors $\{0.25, 0.30, 0.40, 0.50, 0.60, 0.70, 0.80, 0.85, 0.90, 0.95, 1\}$ are eligible only if all-anomaly validation AUPRC improves over $\alpha = 1$ at every difficulty. Selection then maximizes mean one-vs-normal AUPRC across voltage drift, intermittent faults, and slow recovery, breaking ties by all-anomaly AUPRC; otherwise $\alpha = 1$ is retained. Complete ordered score trajectories are accumulated before labels are masked. Nine-family leave-one-family-out analysis removes a family's labels from every selection step to test dependence on known mechanisms.

Composition tests compare natural, balanced, and 75%-high-resistance subsets of four fault trajectories with the

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2026.3753456

TABLE 5. Physical-Domain Protocol Parameters

Parameter group	Settings
Nominal event and history	38 A for 120 s at 25 °C; remove 20% of a nominal 17 Ah charge at 1.7 A, then rest for 600 s.
One-factor design/aging domains	Electrode thickness: 0.9 or 1.1 of nominal; acid/kinetics: fresh 1.00/1.00 or more-aged 0.95/0.80 (nominal 0.975/0.90).
One-factor operating domains	Prior charge removal: 0 or 40%; temperature: 15 or 35 °C; load: 22 or 52 A; discharge duration: 60 or 240 s.
Two severe combined domains	Thickness 0.9; acid/kinetics 0.95/0.80; prior charge removal 40%; load 52 A; duration 240 s. Temperature is 15 or 35 °C.
Mild combined domain	Thickness 1.1; acid/kinetics 1.00/1.00; no prior charge removal; load 22 A; duration 60 s; other settings nominal.
Background multipliers	Independent uniform draws: acid concentration 0.98–1.02; each electrode's thickness 0.95–1.05, maximum porosity 0.97–1.03, and exchange-current density 0.95–1.05.
Fault multipliers (mild/moderate/severe)	Acid inventory and pore blockage: 0.90/0.75/0.60; active-material thickness: 0.85/0.65/0.45; interfacial kinetics: 0.75/0.55/0.35. Applied after domain/background settings; electrode faults affect both electrodes.
Numerical solver	Negative-electrode/separator/positive-electrode mesh: 50/82/68; IDAKLU relative and absolute tolerances: 10^{-8}.
Numerical deadbands (approximate)	Pre-event voltage: 10^{-5} V; sag: 0.00411678 V; resistance: 0.120218 mΩ; slope: 7.19091×10^{-6} V/s; minimum voltage: 0.00527025 V; recovery voltage: 0.000797617 V; recovery ratio: 0.00104515; temperature rise: 0.0120607 °C.

same healthy set. The standard pool contains only one high-resistance trajectory, so a separate development pair with three such trajectories enables the 75% condition. Complete-selection, peer-only, and attainable-composition results are reported on their respective development pairs, separately from the matched method comparison; they do not change its frozen configurations.

D. PHYSICAL AND OPERATIONAL ROBUSTNESS

Physical domains: The fixed-direction test uses 16 controlled parameter domains (Table 5), with all eight q_k and $K = 3$ unchanged. The full PyBaMM lead–acid model with the same parameter set and lumped thermal dynamics resolves spatial concentration and potential variations. Increasing differences between model fidelities at higher loads in development comparisons motivated its use here and for the operational stress profiles, alongside the leading-order quasi-static model used in the four-domain electrochemical benchmark.

Ten physical realizations are paired across domains. Each shares one background draw across a healthy profile and four fault families at three severities, giving 13 profiles per domain. Independent observation-noise streams, specified below, generate 40-unit, 30-event healthy-calibration and test sequences; four faulty test positions per event cycle through the families and severities. The eight descriptors use native 2 s samples and duration-adjusted windows, including recovery at $[d + 20, d + 40]$ and temperature rise through $d + 60$, instead of the fixed 128-point reconstruction grid. The raw statistic g, without component normalization or clipping, is compared with the otherwise identical top-three absolute-deviation score. Thresholds are healthy-score 99th percentiles from the nominal or matching domain.

For matched profiles, $q_k(x_k^{\text{fault}} - x_k^{\text{healthy}})$ is classified as aligned or opposed when its positive or negative change exceeds the feature-specific deadband; smaller changes are unresolved. The deadbands in Table 5 use twice the largest feature change in development mesh/tolerance checks, with per-feature floors. Detection metrics require all 13 profiles to complete in every planned realization; incomplete-domain outcomes are retained without substituting easier faults or shorter events.

Operational stress tests: Four tests examine acquisition quality and within-string comparability using 10 independent full-model realizations, each with 40 units and 36 events. Autoencoders and fusion weights transfer without retraining from an independent medium-difficulty development realization of the primary simulator. Both normalized $\hat{g}$ and the fixed fused score retain their primary-development healthy-training references; RSSI and UPS load use training medians (Table 2). Stress-suite healthy sequences calibrate detection thresholds.

All tests use the nominal conditioning and 38 A, 120 s discharge in Table 5, with 0.5 Hz sampling except in the rate sweep. Position voltage/temperature offsets have standard deviations of 4 mV/0.05 °C, shared event voltage shifts 10 mV, and sample noise 2 mV/0.01 °C. Seven background multipliers are shared within each physical realization. Its profiles are reused with independent observation noise across events; calendar aging is not simulated. The first three tests cover four fault families at three severities; replacement uses their middle severities.

Sampling: Rates of 10, 2, 1, 0.5, and 0.1 Hz are drawn from 200 Hz model output at four quarter-period phases. Healthy-score 99th-percentile thresholds use either 10 Hz/zero-phase or the matching condition. Minimum raw support requires one point in each of $[-5, -1]$, $[0, 3]$, $[140, 160]$, and $[0, 120]$ s, three in the slope window $[10, 60]$ s, and a record reaching 160 s. Non-RSSI descriptors must be finite, the current step must exceed 0.1 A, and the voltage sag must exceed 1 mV. All 40 units must pass. These criteria define observation-support subsets for this simulated event without changing scoring or history updates. The 160 s endpoint does not ensure temperature-window coverage through 180 s.

Telemetry: The 23 conditions comprise a clean baseline; voltage offsets $\pm 5, \pm 20$ mV, gain $\pm 1\%$, quantization 1/10 mV, and drift ± 20 mV; temperature offset ± 0.5 °C and gain $\pm 1\%$; timestamps ± 2 s; and random/contiguous record loss of 2%, 5%, or 10%. Sensor errors affect 10 units independently of fault locations; loss affects all units while current remains complete. At 10% contiguous loss, 13 of 129 records are removed per unit from a uniformly selected one of 117 starts. Both loss modes use the same support checks and fixed clean-healthy thresholds.

Mixed backgrounds: Four, 10, or 16 units receive thickness factors 0.9/1.1 or fresh/more-aged acid/kinetics factors 1.0/1.0

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2026.3753456

TABLE 6. AUPRC of 19 Methods on 15 Matched Realizations per Difficulty

Method	Easy	Medium	Hard
Fixed-threshold	0.3309 [0.3224, 0.3394]	0.4035 [0.3989, 0.4080]	0.4225 [0.4147, 0.4303]
Robust-MAD	0.4825 [0.4589, 0.5062]	0.3553 [0.3457, 0.3648]	0.2615 [0.2503, 0.2726]
Isolation-Forest	0.4034 [0.3661, 0.4406]	0.4601 [0.4273, 0.4929]	0.4606 [0.4187, 0.5025]
PCA-reconstruction	0.7758 [0.7501, 0.8016]	0.7576 [0.7479, 0.7674]	0.7533 [0.7388, 0.7678]
Feature-AE	0.7758 [0.7459, 0.8056]	0.7822 [0.7702, 0.7942]	0.7890 [0.7760, 0.8020]
Task-adapted STAE	0.1876 [0.1697, 0.2056]	0.2739 [0.2418, 0.3061]	0.2706 [0.2413, 0.2999]
Task-adapted MSKLF	0.2102 [0.1787, 0.2417]	0.4867 [0.4355, 0.5379]	0.5861 [0.5433, 0.6289]
IGAD-2025-AE	0.4273 [0.3836, 0.4710]	0.4839 [0.4470, 0.5209]	0.4820 [0.4290, 0.5350]
ScatterAD-2025-adapted	0.6828 [0.6297, 0.7359]	0.5538 [0.5083, 0.5994]	0.5072 [0.4657, 0.5488]
CATCH-2025-adapted	0.3485 [0.3388, 0.3582]	0.3560 [0.3447, 0.3673]	0.3361 [0.3223, 0.3498]
RTdetector-2025-adapted	0.1468 [0.1306, 0.1630]	0.2214 [0.2021, 0.2406]	0.2124 [0.1914, 0.2333]
Sub-Adjacent-Transformer-2024-adapted	0.3471 [0.3012, 0.3930]	0.3426 [0.3015, 0.3837]	0.3112 [0.2812, 0.3412]
DCdetector-2023-adapted	0.0465 [0.0447, 0.0482]	0.0851 [0.0817, 0.0886]	0.0929 [0.0908, 0.0950]
TimeMixer++-2025-adapted	0.6610 [0.6241, 0.6979]	0.4790 [0.4542, 0.5038]	0.4320 [0.4095, 0.4545]
DADA-2025-zero-shot	0.1532 [0.1385, 0.1679]	0.2385 [0.2234, 0.2537]	0.2131 [0.2009, 0.2253]
MOMENT-1-small-2024-zero-shot	0.3299 [0.3197, 0.3400]	0.3382 [0.3261, 0.3502]	0.3049 [0.2914, 0.3185]
MOMENT-Stat+CAE-2024-adapted	0.4161 [0.3645, 0.4678]	0.5170 [0.4831, 0.5509]	0.5483 [0.5144, 0.5823]
SAE-LOF-WD-2026-adapted	0.6205 [0.6045, 0.6365]	0.5239 [0.5042, 0.5435]	0.5712 [0.5482, 0.5941]
GRAPE-UPS	**0.9039** [0.8927, 0.9150]	**0.8385** [0.8226, 0.8545]	**0.8232** [0.7978, 0.8486]

Values are means with descriptive 95% t-intervals over 15 independent simulation realizations per difficulty. Bold indicates the highest mean AUPRC in each column. GRAPE-UPS uses validation-selected alert scoring. Constant-score AUPRC equals test-set anomaly prevalence: 0.0458, 0.0825, and 0.0917 for easy, medium, and hard conditions, respectively. PCA and Feature-AE tie at the displayed precision under easy conditions; PCA has the higher unrounded mean. Method definitions and event-level adaptations are given in Section IV-B. All 54 advantages remained significant under the sign test (Holm-adjusted $p \leq 0.0352$); all 51 comparisons among the 18 methods with 20-realization coverage were significant under sign flipping ($p \leq 0.00113$, Holm-adjusted). The five-realization subset alone did not reach significance after correction.

or 0.95/0.8, versus nominal 1.0 and 0.975/0.9. Sequences contain no faults or four faults independent of background membership. Healthy-score 99th-percentile thresholds use nominal or matching backgrounds.

Replacement: Zero, four, 10, or 16 positions change between events 12 and 13 under the five backgrounds. Weak-to-healthy replacements include the four weak units; healthy-to-healthy and healthy-to-weak controls isolate replacement and new faults. Exactly four replaced units are weak in the latter. Policies retain history, reset only accumulation, or assign new identities to reset both historical descriptors and accumulation. For $\alpha = 0.80$ and 1, thresholds are nominal-healthy accumulated-score 99th percentiles and remain fixed after replacement. Outcomes cover the next 24 events, weighting the four fault families equally within each realization.

V. RESULTS

A. COMPARATIVE DETECTION AND THE ROLE OF TASK SEPARATION

Across the common 15-realization subset at each difficulty, GRAPE-UPS achieved mean AUPRC of 0.9039, 0.8385, and 0.8232 (Table 6). Against the strongest external comparator—PCA under easy conditions and Feature-AE under medium and hard conditions—the paired gains were 0.1280 [0.1051, 0.1520], 0.0564 [0.0412, 0.0704], and 0.0342 [0.0110, 0.0545], with Holm-adjusted sign-flip p-values of 0.00330, 0.00330, and 0.01343. All 54 corrected comparisons favored GRAPE-UPS at the 0.05 level. It led all external comparators in 15/15 easy realizations and exceeded Feature-AE in 14/15 medium and 12/15 hard realizations. Although anomaly prevalence increased across difficulty levels, mean GRAPE-UPS AUPRC declined. This decline therefore cannot be attributed to a shrinking positive class.

Fig. 2(a) visualizes the 19-method comparison in Table 6. Panels (b)–(d) and Table 7 examine task separation on the separate five-realization structural subset.

PCA reconstruction and Feature-AE were the strongest external comparators, supporting the predictive value of absolute multivariate structure in this task. The evaluated sequence and foundation-model adaptations did not surpass these simpler comparators under the common event-level interface.

On the 15-realization method-comparison subset at each difficulty, zero-shot MOMENT-1-small produced AUPRC values of 0.330, 0.338, and 0.305. MOMENT-Stat+CAE combines statistics of frozen embeddings with a convolutional autoencoder trained on healthy data. This adapter increased hard-condition AUPRC to 0.548. The remaining gap to the selected GRAPE-UPS score was 0.275.

The three GRAPE-UPS scoring choices also differ on the 15-realization comparison set. The raw directed peer statistic g achieved AUPRC of 0.9046/0.8102/0.7785 under easy/medium/hard conditions; its normalized and zero-clipped form $\widehat{g}$, used for maintenance ranking, achieved 0.9046/0.8098/0.7773. Relative to raw g, validation-selected alert scoring increased AUPRC by 0.0284/0.0448 under medium/hard conditions but reduced it by 0.0007 under easy conditions, where the five-realization subset showed a small gain. Thus, offering reconstruction-based fusion as an alert alternative mainly benefited the more difficult conditions.

Table 7 isolates the two GRAPE-UPS outputs on the five-realization structural subset at each difficulty. The directed peer score achieved 0.887, 0.808, and 0.772 without learned reconstruction. The fixed fused score used the validation-derived fusion weights in every test condition, whereas the selected alert score reflects a choice between the peer and fused alternatives based on validation AUPRC. Averaged across difficulty levels, the selected alert score improved AUPRC by 0.029 over directed peer scoring. The gain was only 0.002 under easy conditions but reached 0.061 under hard conditions.

The tested fixed fusion improved hard-condition detection at

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2026.3753456

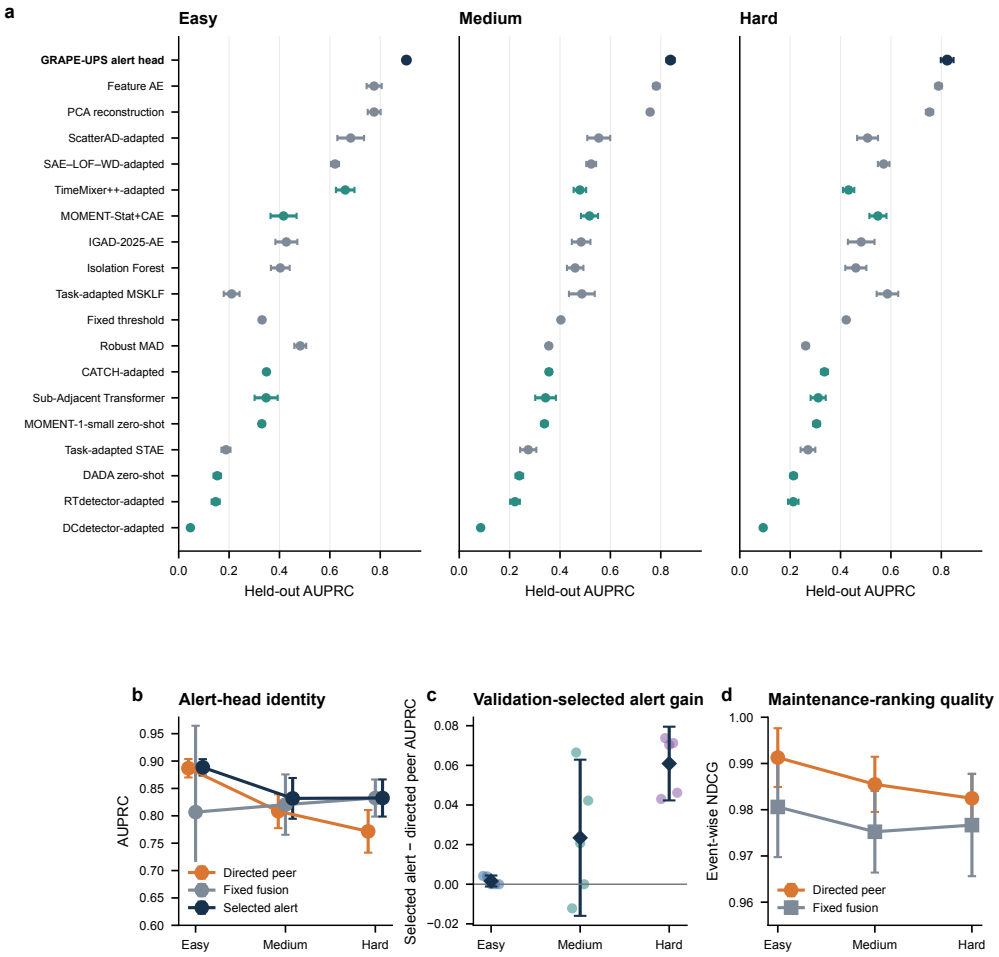


FIGURE 2. **Detection and ranking performance. (a) Held-out AUPRC of 19 methods over 15 realizations per difficulty, matching Table 6. (b)–(d) Task separation on five separate realizations per difficulty: (b) alert AUPRC under directed-peer, fixed-fusion, and validation-selected scoring; (c) per-realization selected-alert gain over directed peer scoring; and (d) NDCG for directed peer and fixed fusion. Error bars show descriptive 95% *t*-intervals. SAE-LOF-WD denotes the event-level adaptation.**

TABLE 7. **Task-Separated Performance on the Five-Realization Structural Subset per Difficulty**

Difficulty	Directed peer AUPRC	Fixed fused AUPRC	Selected alert AUPRC	Alert gain over peer	Directed-peer NDCG
Easy	0.887	0.807	**0.889**	+0.002	0.991
Medium	0.808	0.820	**0.832**	+0.023	0.985
Hard	0.772	**0.833**	**0.833**	+0.061	0.982

Bold indicates the highest AUPRC within each difficulty; tied best values are both bold.

the cost of easy-condition performance. Fixed fusion reduced easy-condition AUPRC from 0.887 to 0.807 even though its weights were determined on validation data. The validation-gated selector avoided most of this loss and returned 0.889. Under the hard condition, fixed fusion and the selected alert score both reached 0.833. Reconstruction was therefore useful, but its utility depended on the operating regime and realization.

Ranking provides a separate test of architectural purpose. On the five-realization structural subset, directed-peer NDCG reached 0.991, 0.985, and 0.982, compared with 0.981, 0.975, and 0.977 for the fixed fused score. The differences were small but consistent across difficulty levels. Reconstruction improved hard-condition detection while slightly weakening the order of units within an event.

B. SEPARATE CONTRIBUTIONS OF REFERENCE AND DIRECTION

The factorial comparison tests whether detection benefits from the choice of reference, the use of adverse directions, or their combination. Four scores combine either a healthy-training reference or contemporaneous peers with either two-sided deviations or adverse-direction deviations. Direction constraints improved all-anomaly AUPRC with both references. Their gains were larger with peers: the difference between the two directional gains was 0.2828, 0.3225, and 0.3505 under easy, medium, and hard conditions, respectively.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2026.3753456

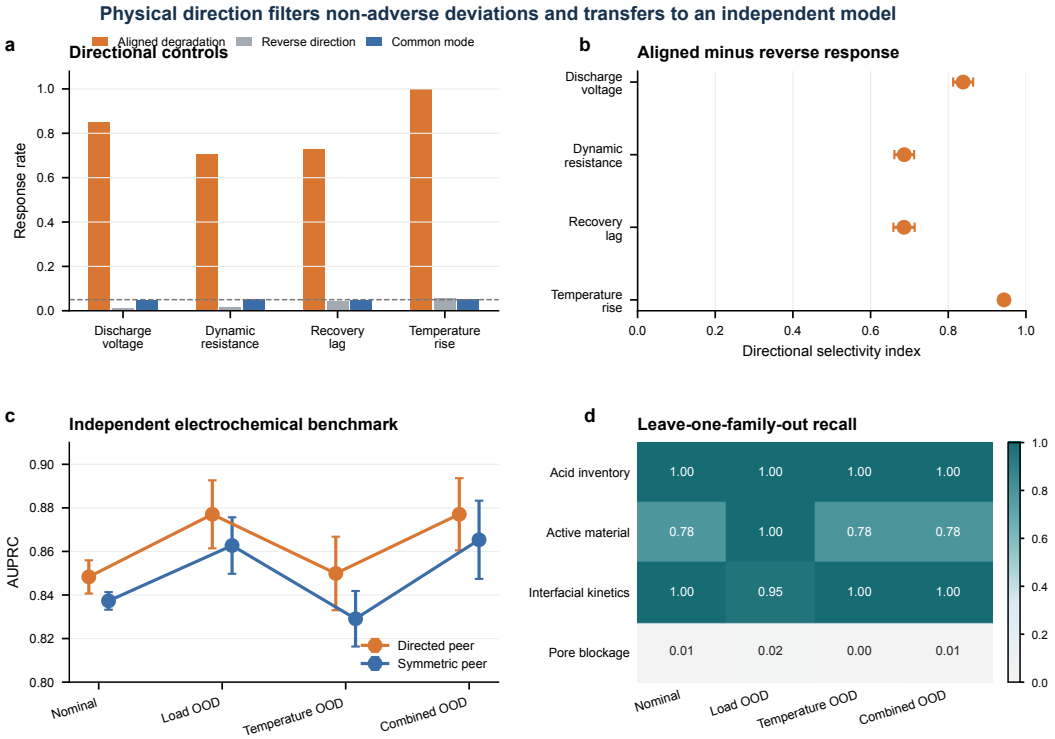


FIGURE 3. Directional selectivity and cross-simulator validation. (a) Response rates under directed, reversed, and common-mode perturbations. (b) Directional selectivity index across five primary-simulator realizations at each of three difficulties. (c) Directed versus symmetric peer AUPRC on the independent electrochemical benchmark under nominal and shifted operating domains. (d) Leave-one-family-out recall; each row omits one degradation family from labeled validation calibration.

When both scores used adverse directions, changing the reference from healthy training data to peers increased AUPRC by 0.0657 and 0.0140 under easy and medium conditions. For hard-condition all-anomaly detection, the change was -0.0035 with a 95% interval of $[-0.0184, 0.0114]$, providing no clear advantage for either reference. For weak-battery detection under the same hard conditions, peers improved AUPRC by 0.0175 [0.0031, 0.0319]. Peer referencing therefore benefits the intended weak-battery target even where it does not improve detection of all anomaly types.

C. DIRECTIONAL SELECTIVITY UNDER STRUCTURAL PERTURBATION AND DISTRIBUTION SHIFT

Reverse-direction controls test whether the score responds only to deviation magnitude. Common-mode controls test whether it reacts to a string-wide shift. The independent simulator tests dependence on the primary generator. Leave-one-family-out calibration then tests whether validation labels must contain the evaluated mechanism.

Directional perturbation controls: Adverse-direction, reverse-direction, and common-mode perturbations were applied to otherwise healthy responses. Adverse and reverse perturbations had matched amplitudes, and thresholds remained at their clean-response calibration. Across voltage, dynamic resistance, recovery, and temperature channels, adverse-direction recall ranged from 0.704 to 1.000. Reverse-direction response rates remained between 0.009 and 0.056, while common-mode responses did not exceed the 5% reference rate. The directional selectivity index ranged from 0.686 to 0.944 (Fig. 3).

These controls isolate the contribution of the sign constraint. A magnitude-only score would respond similarly to adverse and reversed perturbations. The observed separation shows that the implemented direction map changes the decision.

Cross-simulator transfer: Directed peer scoring achieved AUPRC of 0.848 in the nominal electrochemical domain and 0.850–0.877 under load, temperature, and combined shifts. Across five realizations in each of four domains, directed scoring exceeded the symmetric peer score in every domain aggregate. The pooled mean gain was 0.014, with a descriptive 95% interval of [0.010, 0.019].

The positive gain on a separately implemented generator supports transfer of the adverse-direction prior under the tested operating shifts. The shifted-domain values show that the tested shifts did not reduce aggregate AUPRC below the nominal value in this benchmark. They should not be interpreted as evidence that distribution shift improves detection.

Leave-one-family-out calibration: Acid-inventory loss achieved recall of 1.000 when withheld from calibration. Interfacial-kinetics loss achieved 0.950–1.000, and active-material loss achieved 0.775–1.000. Electrode-porosity reduction, referred to below as pore blockage, produced near-zero recall and one-vs-normal AUPRC of 0.066–0.133.

This contrast defines a mechanism boundary. The successful families produced event-level departures that remained

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2026.3753456

TABLE 8. Aggregation-Width Sensitivity of the Frozen Directed Peer Score

Difficulty	Metric	$K = 1$	$K = 3$ (primary)	$K = 5$	$K = 7$
Easy	AUPRC	0.851	0.887	**0.905**	0.903
	Event-wise NDCG	0.987	0.991	**0.992**	0.991
Medium	AUPRC	0.763	0.808	**0.817**	0.808
	Event-wise NDCG	0.976	**0.985**	0.985	0.983
Hard	AUPRC	0.696	0.772	**0.792**	0.786
	Event-wise NDCG	0.968	0.982	**0.984**	0.983

Values are means over five independently seeded realizations per difficulty. Bold indicates the highest mean in each row, determined before rounding. Only the number of aggregated directional cues changes; all fitted components and decision settings remain frozen.

compatible with the fixed direction map. In the benchmark, pore blockage provided insufficient acute directional contrast for the calibrated score. It shows that a mechanism can be physically adverse yet remain poorly detectable by a short-event directional screen.

D. SENSITIVITY TO THE NUMBER OF AGGREGATED CUES

Table 8 evaluates whether the directed peer result depends on the prespecified three-cue aggregation. Reducing the score to a single cue lowered mean AUPRC relative to $K = 3$ by 0.036, 0.046, and 0.075 under easy, medium, and hard conditions, respectively. The decrease occurred in four of five easy realizations and in all medium and hard realizations. A single extreme feature therefore did not capture the distributed evidence used by the peer score.

Increasing the width to $K = 5$ raised mean AUPRC by 0.018, 0.008, and 0.020 relative to $K = 3$; the paired change was positive in 5/5, 4/5, and 5/5 realizations, respectively. The corresponding NDCG changes were only +0.0002, -0.0009, and +0.0016 and changed sign across realizations. Extending the aggregation to $K = 7$ did not continue the AUPRC improvement: its mean values were lower than those for $K = 5$ at every difficulty, and medium-condition AUPRC returned to the $K = 3$ level. Thus, the ranking result reaches a broad plateau once several cues are combined, while wider aggregation can provide a modest detection gain without a consistent ranking gain. The primary $K = 3$ setting is retained as the prespecified sparse explanation width; the held-out sensitivity results are not used for post hoc reselection.

E. DEPENDENCE ON VALIDATION LABELS

Few labels limited calibration feasibility. With models and fusion weights fixed across five realizations per difficulty, four- and ten-trajectory budgets produced single-class validation sets in 120/300 and 26/300 draws. None occurred at 20 or 40 trajectories.

On the separate development pairs, the complete selection procedure at random 10%, 25%, 50%, and 100% budgets succeeded across all three difficulties in 3, 15, 19, and 20 of 20 draws (Fig. 4). Performance curves describe successful subsets. Calibration succeeded for every draw when one, two, four, or all known-fault trajectories were supplied together with all healthy trajectories.

Among these successful known-fault-label draws, hard-condition static F1 rose through 0.5387, 0.6286, 0.6646, and 0.7130 as the fault-label budget increased. AUPRC was nonmonotonic, distinguishing ranking performance from threshold calibration.

Label composition also mattered. A separate simulated dataset containing three high-resistance fault trajectories allowed their share of the selected fault labels to reach 75%. In this dataset, natural, balanced, and high-resistance-skewed (75%) four-fault-trajectory subsets yielded hard-condition AUPRC of 0.7051, 0.6720, and 0.7505, respectively. Their false-alarm rates were 59.83, 70.74, and 73.50 per 1,000 healthy samples. Balanced sampling did not dominate, and higher AUPRC under skew came with more false alarms. All three compositions selected the same easy-condition peer score, giving AUPRC of 0.7446. These are within-pool comparisons; differing data and training content preclude attributing differences between development pools solely to label composition.

In the complete-selection leave-one-family-out analysis, omitting each of nine fault families during selection changed static AUPRC by 0 to +0.0291. The accumulated-score changes ranged from -0.0524 to +0.0685. Static zeros reflect an unchanged peer score, including four easy omissions with no labels to remove. Other changes reflect weights, head selection, or accumulation decay. Threshold changes alone cannot alter AUPRC.

Aggregate performance in this analysis masked poor voltage-drift coverage: static hard-condition and accumulated medium-condition AUPRC values were 0.0577 and 0.0475, respectively, both with zero recall.

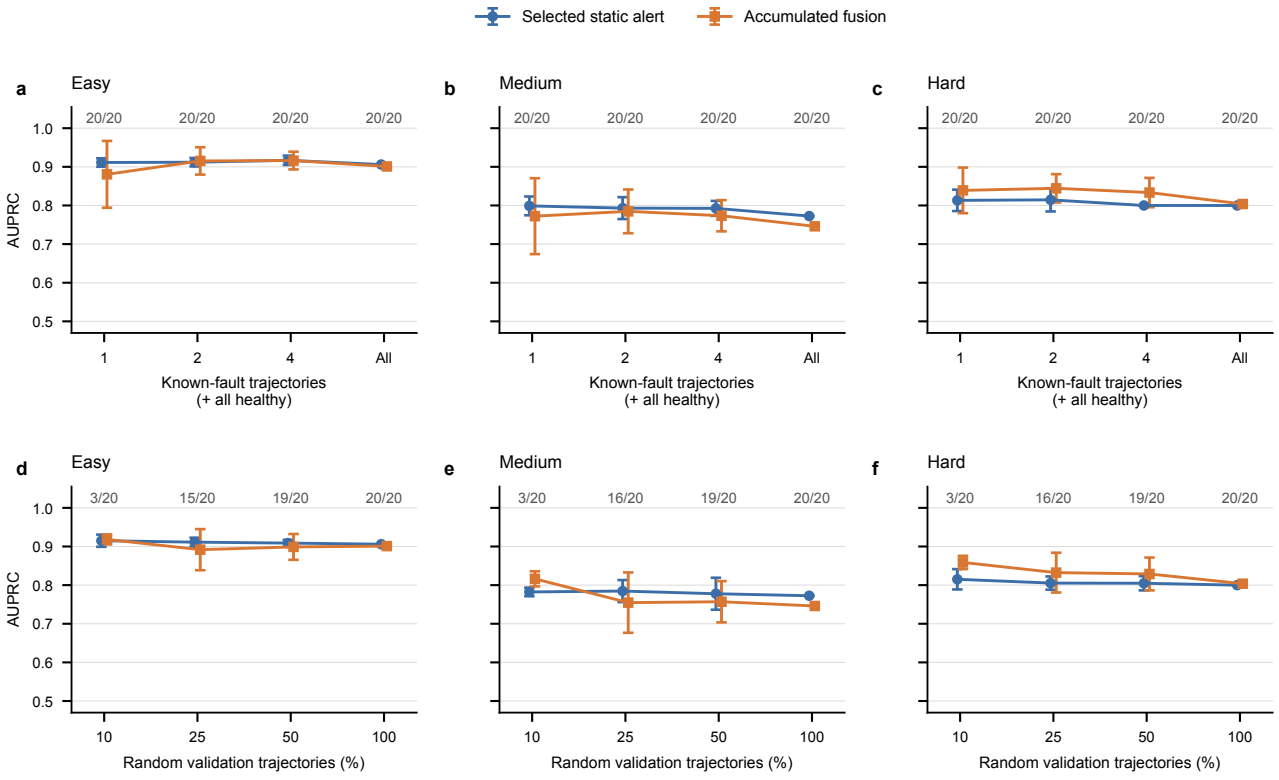
The directed peer score $\widehat{g}$ was also evaluated without fault labels, using a threshold calibrated at the 99th percentile of healthy scores. This comparison used the same independent test string as the full parameter-selection experiment at each difficulty. Easy/medium/hard AUPRC was 0.9059/0.7760/0.7999, with 21.83/14.53/5.50 false alarms per 1,000 healthy samples. This operating mode still requires a healthy reference set; the complete selected-alert framework remains weakly supervised.

F. ROBUSTNESS OF THE PEER REFERENCE UNDER STRING CONTAMINATION

In a controlled test on five development realizations, contamination began with isolated affected units. Their number was then increased until they accounted for 40% of the string. Median/MAD normalization preserved higher recall

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2026.3753456



Mean ± SD over successful label draws; *n*/20 shown in each panel. Fixed development realizations.

FIGURE 4. Complete-selection label-budget curves on fixed development realizations. The top row supplies one, two, four, or all known-fault trajectories together with all healthy trajectories; the bottom row randomly samples 10%, 25%, 50%, or 100% of validation trajectories. Points and error bars show mean and standard deviation across successful label draws, not independent-realization confidence intervals. Low-budget points describe the successful subset and can favor draws with more informative labels. Each point's successful calibration count out of 20 is shown inside the panel; at 25%, easy has 15 and medium/hard have 16 successful calibrations, whereas only 15 draws complete all three difficulties. The all-label masks are identical across draws. Curves show the selected static alert and accumulated fusion extension; neither replaces the frozen primary comparison.

than mean/standard-deviation normalization across the tested range under identical direction rules and thresholds (Fig. 5). At 40% contamination, recall was 0.241 for median/MAD and 0.109 for mean/standard deviation. Reference-center displacement from the uncontaminated baseline was 0.213 and 0.367 normalized units, respectively. The absolute recall of 0.241 at that level shows the remaining effect of reference movement as the affected fraction grows.

The failure mechanism is a moving reference. As affected units occupy more ranks, the event median shifts toward the degraded group and the robust deviations of those units shrink. At the same time, unaffected units can appear more distant from the shifted center. The experiment stops at 40%, so behavior at or above a degraded majority is not inferred.

G. CUE FIDELITY, FAULT COVERAGE, AND TEMPORAL EVIDENCE ACCUMULATION

Per-mechanism analysis distinguishes high aggregate performance from uniform fault coverage. Thermal drift showed near-perfect separation. Resistance, capacity, accelerated aging, sulfation, and sensor offset also produced substantial one-vs-normal AUPRC. Intermittent voltage drops, slow recovery, and voltage drift were less detectable from a single event (Fig. 6). This ordering is consistent with acute mechanisms altering phase-localized cues during the event, while gradual or intermittent mechanisms may produce only small departures within a single window.

The top-three cue set was compared with a mechanism–cue map fixed before held-out analysis. Mechanism-consistent cues occupied 58.8% of selected positions, versus 38.3% after label permutation. This corresponds to a 1.54-fold enrichment. At least one mechanism-consistent cue appeared in 93.6% of anomalous battery–event explanations. The cue enrichment supports the physical consistency of the ranking explanation. Because expected cue sets overlap across mechanisms, these cues explain the score rather than uniquely identify a degradation cause.

Temporal accumulation of the fixed fused score provided a separate test of weak longitudinal evidence. Across the dedicated longitudinal suite, global AUPRC increased in all 15 held-out evaluations. The mean gain was 0.024, with a descriptive 95% interval of [0.018, 0.030]. Across the targeted weak-fault condition evaluations, all 35 matched comparisons were positive and the mean gain was 0.063, with a descriptive 95% interval of [0.046, 0.080]. Hard-condition voltage drift reached an absolute AUPRC of 0.084 after accumulation, a

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2026.3753456

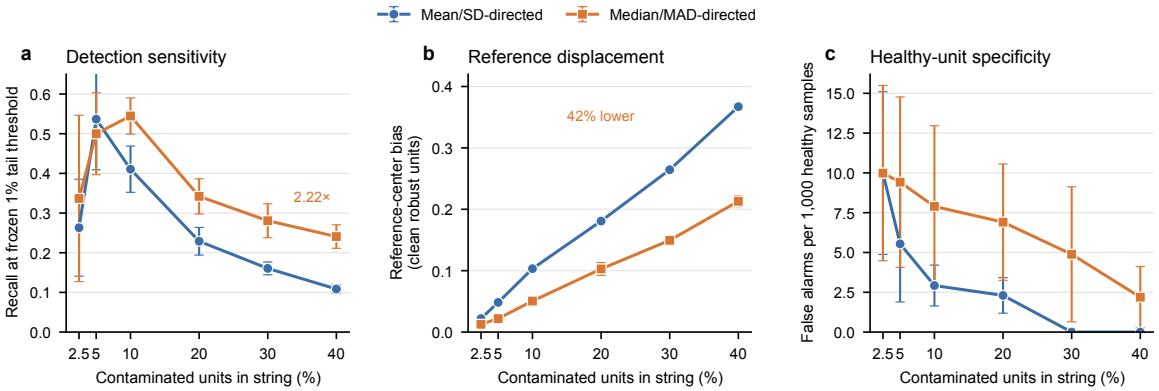


FIGURE 5. **Peer-reference behavior under controlled within-string contamination on five development realizations. (a) Recall under healthy-validation thresholds. (b) Reference-center displacement from the uncontaminated event center. (c) False-alarm rate per 1,000 unaffected samples. All conditions share identical directional constraints and top-three aggregation.**

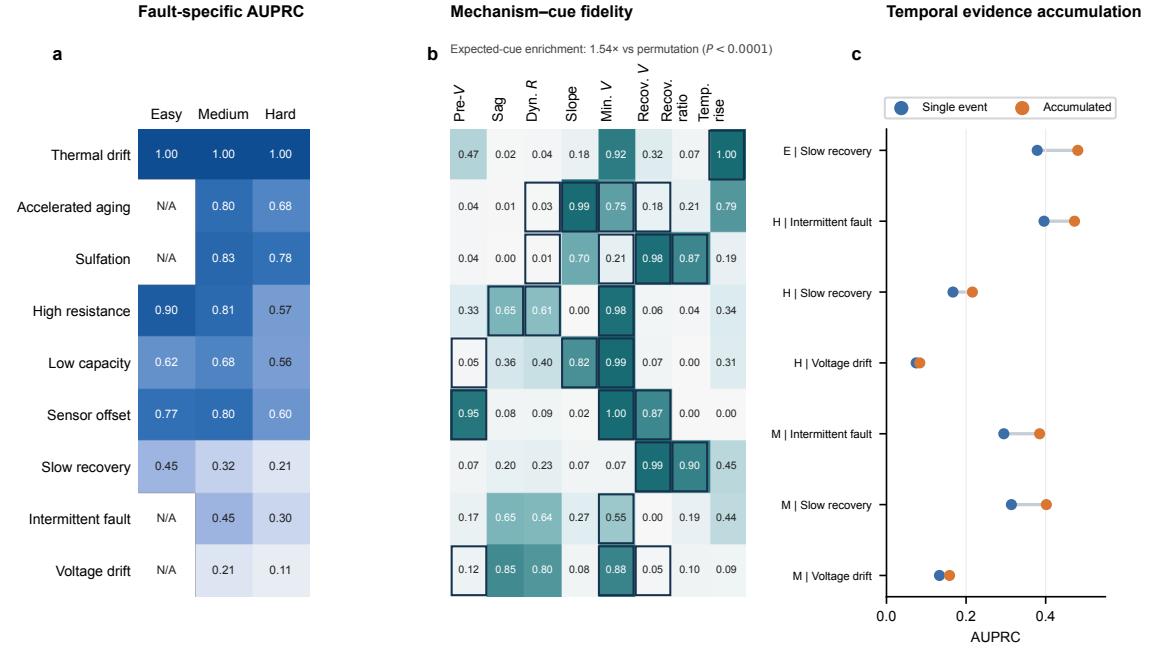


FIGURE 6. **Fault coverage, cue fidelity, and temporal evidence accumulation. Panels (a,b) use a separate suite of five realizations per difficulty. (a) Fixed-fusion one-vs-normal AUPRC by mechanism. (b) Top-three cue probability matrix; outlined cells mark expected mechanism–cue pairs. (c) Single-event versus accumulated fixed-fusion AUPRC on a dedicated longitudinal suite for gradual and phase-localized degradation modes.**

gain of 0.009. For hard slow recovery, the interval for the gain crossed zero.

Repeated small departures become more separable when their sign persists across events. Sparse or inconsistent departures can still be averaged away, and a finite event history cannot create a phase-localized cue that is absent. Temporal accumulation is therefore a useful extension for persistence, not a substitute for diagnostic excitation.

H. FIXED DIRECTIONS ACROSS PHYSICAL DOMAINS

For 14 of the 16 physical conditions, all required healthy and faulty response profiles were simulated successfully in each of the 10 independent realizations, allowing detection performance to be evaluated. Directed-score AUPRC ranged from 0.9930 to 1.0000; transferring the nominal threshold yielded 97.25–100% recall and 0.38–1.56% false positives.

Both scores reached 1.0000 in 12 domains and exceeded 0.9999 in the unconditioned domain (no prior charge removal), limiting discrimination between them. The mild combined domain showed the largest directional advantage: AUPRC was 0.9930 (95% t-interval 0.9892–0.9967), versus 0.9853 (0.9805–0.9902) for symmetric scoring. Mild pore blockage remained less separable, with one-vs-normal AUPRC of 0.6821 (0.5761–0.7881), versus 0.4704 (0.3354–0.6055). These native-grid, full-model results are distinct from the four-domain electrochemical benchmark.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2026.3753456

TABLE 9. Direction Agreement in Complete Physical Domains. Each Feature Has 1,680 Matched Fault–Healthy Profile Comparisons.

Feature	q_k	Aligned	Opposed	Unresolved
Pre-event voltage	-1	1500	0	180
Voltage sag	$+1$	1521	54	105
Current-step resistance	$+1$	1579	60	41
Discharge slope	-1	1680	0	0
Minimum voltage	-1	1680	0	0
Recovery voltage	-1	1308	0	372
Recovery ratio	-1	420	1260	0
Temperature rise	$+1$	1573	0	107

Classification uses the prespecified numerical deadbands. Counts describe correlated paired profiles, not 1,680 independent realizations. The 20 incomplete domain–realization libraries are excluded from this complete-domain table and reported separately.

TABLE 10. Stress-Test Design and Representative Fixed-Fusion Results.

Test	Factor levels	Healthy threshold	Representative result: mean [min,max]
Sampling	10, 2, 1, 0.5, 0.1 Hz; four phases each	99th percentile at 10 Hz/zero phase or matching condition	All 10–0.5 Hz conditions meet minimum raw support; all 0.1 Hz conditions fail it despite high AUPRC
Telemetry	Baseline plus 22 offset, gain, quantization, drift, timing, and loss conditions	Clean-score 99th percentile, fixed	10 mV quantization FPR 1.22% [0.00,5.32]; -20 mV offset 10.05% [4.40,13.04]
Mixed background	Nominal plus four backgrounds at 10/25/40%; healthy and fault sequences	Nominal or matching healthy 99th percentile	More-aged 40%: nominal FPR 40.70% [39.86,43.40]; matching FPR **1.24%** [0.00,2.78]
Replacement	0/4/10/16 positions; five backgrounds; three transitions and history policies; $\alpha = 0.80, 1$	Nominal healthy accumulated-score 99th percentile, separately for each α	Nominal weak-to-healthy, four positions, $\alpha = 0.80$: inherited FPR 9.61% [4.95,25.00]; both reset **3.54%** [0.00,18.75]

Bold indicates the lower mean FPR within each matched threshold or history-policy comparison. The false-positive rate (FPR) uses healthy positions as its denominator. Replacement FPR above is restricted to the formerly weak positions; mixed-background FPR uses the entire healthy string. Ranges are across 10 realizations with one fixed reconstruction model. The zero-replacement control does not create changed positions; undefined conditional metrics remain unreported rather than set to zero.

Individual signs were less transferable than aggregate detection (Table 9). Minimum voltage and discharge slope aligned in all 1,680 paired comparisons, whereas recovery ratio opposed the prescribed sign in 75%. Its direction depended on the fault mechanism because both recovery above the minimum and initial voltage sag determine the ratio. One-sided projection suppresses oppositely directed departures, allowing top-three aggregation to retain evidence from other descriptors. This provides a possible explanation for effective aggregate detection despite an uninformative cue; it does not establish universal validity of the prescribed signs.

The two severe combined domains lacked complete event observations. Of 260 planned profiles, 230 completed: 20 acid-loss profiles failed numerically near local electrolyte depletion, and 10 pore-blockage profiles terminated at a voltage event. Across all 16 domains, 2,050 of 2,080 profiles completed. Solver and mesh checks did not establish reliable completion under the specified conditions. Missing discharge and recovery observations prevented valid descriptor computation, so these two domains were excluded from complete-domain detection metrics.

I. SENSITIVITY TO SAMPLING, TELEMETRY, AND REPLACEMENT

Table 10 summarizes the operational stress tests. Unless stated otherwise, results describe the fixed fused score across 10 realizations; bracketed ranges are realization minima and maxima.

At 0.1 Hz, no sampling phase met the minimum raw-observation requirements, although all phases at 10–0.5 Hz did. Nevertheless, 0.1 Hz fused AUPRC remained 0.9954–0.9968 across phase means, while false-positive rates under the 10 Hz threshold ranged from 0.79% to 100%. Matching healthy calibration reduced false positives to 1.56–1.67%. The peer score was less phase-sensitive, with AUPRC of 1.0000 and nominal-threshold false-positive rates of 0.87–1.06%. High ranking accuracy and recalibration therefore did not compensate for missing native observations at the field record's median sampling rate.

Quantization preserved low false-positive rates, whereas voltage gain and timestamp errors could flag every affected healthy position. Quantization at 1 or 10 mV kept false-positive rates near the clean-baseline rate of 1.17%. Voltage offsets were asymmetric: -20 and $+20$ mV produced 10.05% and 0.81%, respectively. A -1% voltage gain error and $+2$ s timestamp shift produced 25.83% and 25.86% overall false positives. Unaffected positions remained near 1.1%, indicating disrupted comparability at the affected units.

Contiguous record loss reduced AUPRC more than random loss. At 10% random loss, AUPRC was 0.9339 [0.9093,0.9523]; contiguous loss reduced it to 0.4974 [0.4507,0.5492]. Minimum raw support remained in 45.28% of events under random loss but only 2 of 360 under contiguous loss. Corresponding false-positive rates were 5.52% and 13.07%.

False-positive rates varied with the type and proportion of the altered healthy background. Increasing the more-aged fraction from 10% to 25% to 40% raised whole-string false positives from 11.05% to 25.88% to 40.70%. Fresh and thicker-design backgrounds instead yielded 0.63–1.12%. For the peer score, more-aged and thinner subgroups were almost uniformly flagged, whereas fresh and thicker subgroups pro-

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2026.3753456

duced no alerts. At 40% more-aged units, matching calibration reduced false positives to 1.24% but lowered fault recall from 97.78% to 80.42%. The thinner-design counterpart changed from 24.80% to 1.01% false positives and from 97.36% to 87.64% recall. Recalibration therefore traded alert burden against sensitivity rather than restoring background invariance.

Replacement separated history effects from background mismatch. For four nominal-background weak-to-healthy replacements, resetting both histories reduced post-replacement false positives from 9.61% to 3.54% at $\alpha = 0.80$ (Table 10). With both histories reset and $\alpha = 1$, thinner-design and more-aged replacements still yielded 97.40% and 100% false positives. Healthy-to-healthy controls gave the same rates, showing that these alerts persisted without accumulator memory.

For healthy-to-weak replacements, condition-mean 24-event recall ranged from 93.36% to 100% at $\alpha = 0.80$, and from 89.84% to 100% at $\alpha = 1$. The nominal four-position condition achieved 100% recall throughout; high recall in heterogeneous backgrounds must be read alongside their high healthy-control alert rates.

VI. DISCUSSION

A. WHAT THE RESULTS ESTABLISH

GRAPE-UPS defines the reference distribution at the shared load event and restricts evidence to prespecified adverse directions. Peer referencing reduces shared operating variation, while one-sided projection excludes departures opposite to the direction map. Their combination distinguishes the method from magnitude-based inconsistency scoring.

Task separation preserves this physical comparison for maintenance ordering while admitting reconstruction evidence for alert screening. Stable ranking across aggregation widths supports a compact top-three explanation set. Reconstruction can then contribute information relevant to alert discrimination without changing the basis of maintenance priority.

The alerting benefit of validation selection varied across evaluation realizations. The easy-condition gain in the five-realization structural subset became a small loss in the 15-realization comparison, whereas medium/hard mean gains remained positive. Validation selection estimates which alert score will generalize from finite validation data, so its selected configuration and observed reconstruction benefit can vary across realizations. The physical ranking rule remains unchanged. These results support an average alerting benefit under the more difficult conditions, without guaranteeing a benefit from fusion on every evaluation sample.

Temperature or initial-charge differences between comparable lead–acid units can alter voltage, current-step resistance, and recovery responses, resembling or masking degradation cues. Temperature rise also depends on heat dissipation, so a small rise need not indicate good health. Aging mechanisms can change sag and recovery by different amounts, as the recovery-ratio reversals in Table 9 illustrate. One-sided projection sets a feature contribution to zero when its peer-relative deviation opposes the prescribed direction; other descriptors may still contribute. Aggregate detection does not establish the validity of every feature sign. Mechanisms with weak short-event contrast, including pore blockage and gradual drift, remain difficult to detect.

The physical-domain evaluation covers one lead–acid model family, with complete-domain detection metrics for 14 of the 16 designed conditions. Aging and charge state enter through parameter changes and prior charge removal. Other chemistries can have different voltage–charge and relaxation responses. Their feature directions, and those under untested temperature or charge conditions, would require comparison with independently assessed battery condition.

Common deterioration and background heterogeneity affect peer scoring in different ways. If units deteriorate similarly, within-event centering can remove the shared change and leave little relative evidence, even when absolute condition has worsened. At 40% contamination, median/MAD retained higher recall than mean/standard deviation, but recall was still only 0.241. Heterogeneity can instead make a healthy subgroup appear adverse relative to the dominant background, as the mixed-background and replacement tests show. The shared current controls excitation, not design, service age, temperature, or initial charge. Historical or comparable-string references could provide additional contrasts if their healthy condition and operating comparability were independently established; their effectiveness remains to be evaluated.

B. RELATION TO EXISTING BATTERY AND TIME-SERIES METHODS

Relative to statistical pack methods [3], [27], [28], GRAPE-UPS includes physical sign in the anomaly hypothesis. Relative to learned pack methods [17], [18], [19], [20], [21], [22], [29], it defines contemporaneous peers before representation learning and preserves a direct maintenance-order score. Its screening task is narrower than multi-perspective or fault-differentiation approaches [30], [31].

Existing approaches use multiple evidence sources for different diagnostic purposes. Wang et al. [33] combine reconstruction, density, and distributional evidence for anomaly alarms, while Zhang et al. [32] address labeled cross-chemistry classification and Luder et al. [26] generate simulated battery data. GRAPE-UPS assigns distinct roles to its evidence: directed peer departures determine maintenance order, while reconstruction provides an additional option for alerting.

C. OPERATIONAL INTERPRETATION

The maintenance ranking identifies units to inspect first, while the top-three cues identify measurements that warrant attention. A resistance-dominated score points to the current-step and voltage-sag records; a temperature-dominated score points to temperature measurements and local operating conditions. An alert calls for review of observation quality, peer comparability, and corroborating battery measurements. Repeated events provide evidence for accumulation across successive events, and persistence across well-supported events can help prioritize follow-up. Capacity or discharge assessment informs

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2026.3753456

replacement decisions. The reported cues explain contributions to the score rather than uniquely establish a fault cause.

The measured cost supports event-level screening on a CPU-based gateway once the event record is available. On a single thread of a consumer-grade CPU (Apple M5 Pro), preprocessing and scoring an in-memory 40-unit event took 32.14 ms for GRAPE-UPS, compared with 19.38 ms for PCA, 20.26 ms for Feature-AE, and 343.70 ms for zero-shot MOMENT-1-small. This is a modest increase over the compact baselines and substantially below the foundation-model baseline. Operational latency also includes collection of the discharge and recovery observations. Peak process memory was 1,045.4 MiB, compared with 974.9, 940.9, and 1,394.2 MiB, respectively; these desktop peaks include the software runtime, data handling, and caches.

For an embedded implementation using FP32 and sequential per-unit inference, GRAPE-UPS has an estimated core storage footprint of approximately 0.4 MiB for parameters, event inputs, features, explicit intermediate tensors, and retained history. This core footprint fits within 1 MiB-class RAM, with additional allocation required for acquisition buffers and runtime workspace. The estimate provides a memory budget for implementation; execution time on the target processor would need to be measured separately.

Recalibration data should reflect the target sensor configuration, loads, temperatures, and string composition. With independently assessed healthy data alone, the peer-score threshold and healthy alert rate can be assessed, but fault recall remains unmeasured. Where independently verified weak outcomes are also available, thresholds should be selected on calibration data and the frozen alert rule evaluated on held-out observations for both false-positive rate and recall. These healthy and weak labels also support full alert selection. In the 40% more-aged mixed background, matching calibration reduced the false-positive rate from 40.70% to 1.24% while recall fell from 97.78% to 80.42%. The threshold changes the operating point, not the underlying score ordering.

History resets address a different source of false alarms. A confirmed record linking a replacement battery to its position allows inherited descriptor history and accumulated evidence to be cleared. The replacement tests show that background-related alerts can persist after both histories are reset. Reassessing peer comparability, resetting identity-linked history, and recalibrating thresholds therefore address distinct problems.

Acquisition must resolve the phases used by the descriptors. Unit voltage, unit temperature, and shared current require a common time base and native observations before and after the current transition. Record coverage must include the discharge and recovery windows in Table 2, with temperature measurements through $\min(240, d + 60)$ s. The 160 s minimum-support endpoint does not ensure full temperature-window coverage for a 120 s discharge. The evaluated reconstruction inputs span $[-5, 240]$ s. Events whose recovery window extends beyond 240 s would require revised preprocessing before this reconstruction pathway could be applied. The required sampling rate therefore depends on the event windows and their alignment.

In the field record, none of the 40 positions had two native observations within ± 5 s of the inferred onset, and gaps reached 110 s. These irregular gaps differ from the simulated loss patterns, and interpolation does not restore missing observations. Poorly aligned or insufficiently supported events should be reported as diagnostically unevaluable. A deployment would also need an explicit, separately validated policy for updating descriptor history and accumulated scores.

D. LIMITATIONS AND NEXT VALIDATION STEP

The field record establishes the monitoring context but lacks independent health labels. Prospective validation should assess agreement with battery condition and the value of the inspection priorities.

A prospective study could link each physical battery to its string position, installation or replacement interval, and synchronized load-event records through a persistent identifier. Controlled capacity or discharge-performance testing under documented current, temperature, initial charge, and endpoint conditions would provide the reference assessment of usable performance. Resistance or conductance measurements would provide complementary evidence, rather than serve as substitutes for measured capacity. Health assessments would be interpreted independently of GRAPE-UPS outputs; rankings, alerts, and replacement decisions would not define the health labels.

The evaluation would include independently assessed units across the score range, including unflagged units. Assessment dates and intervening service changes would delimit the events to which each label applies; unassessed units would remain unlabeled and be excluded from health-performance metrics. The fractions of unassessed units and unevaluable events would be reported alongside performance. Maintenance ranking would be evaluated by event-wise NDCG against independently defined health grades and the number of confirmed weak units identified within a fixed inspection budget. Alerts would be evaluated separately using recall and precision against independently assessed health labels, with the false-positive rate measured among independently assessed healthy units. Calibration and evaluation would use separate physical strings, with each string's batteries and repeated events kept within the same partition.

VII. CONCLUSION

GRAPE-UPS formulates weak-battery screening as an event-conditioned, one-sided peer comparison. Robust within-event statistics reduce common-mode variation, and physical direction defines the adverse alternative. The directed score determines maintenance ranking. Validation-gated reconstruction can improve alert discrimination without changing that ranking.

Across 15 independent simulated realizations per difficulty, GRAPE-UPS achieved the highest mean AUPRC among 19 methods, with all 54 paired comparisons significant after Holm correction. Reference–direction controls identified

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2026.3753456

the contribution of combining event-conditioned peers with adverse-direction evidence. Separately generated lead–acid electrochemical simulations supported transfer beyond the primary generator under the evaluated operating conditions. These simulation results support task-separated maintenance prioritization and alert screening for battery strings with comparable units, a healthy majority, and adequately resolved load-event responses.

AUTHOR CONTRIBUTIONS

Shibo Luo: investigation, methodology, hardware, and original draft writing. Bin Shen: investigation, methodology, software, and original draft writing. Xiaoning Chen: methodology and original draft writing. Yiquan Fang: investigation, writing review and editing, and resources. Huifeng Wang: methodology and resources.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

DATA AND CODE AVAILABILITY

The simulation datasets and generators, de-identified field observations, GRAPE-UPS implementation, evaluation scripts, and an example trained model are available at https://github.com/mrbin0222/GRAPE-UPS (release v0.1.0). Code and model weights are licensed under Apache-2.0, and datasets under CC BY 4.0.

ACKNOWLEDGMENT

ChatGPT (OpenAI) was used for language editing and LaTeX preparation. The authors reviewed and approved all scientific claims, analyses, and the final text.

REFERENCES

[1] T. Tsujikawa and T. Matsushima, "Remote monitoring of VRLA batteries for telecommunications systems," *J. Power Sources*, vol. 168, no. 1, pp. 99–104, 2007, doi: 10.1016/j.jpowsour.2006.12.062.

[2] J. de la Peña Llerandi, C. Sancho de Mingo, and J. Carpio Ibáñez, "Continuous battery health diagnosis by on-line internal resistance measuring," *Energies*, vol. 12, no. 14, 2019, Art. no. 2836, doi: 10.3390/en12142836.

[3] K. Bhaskar, A. Kumar, J. Bunce, J. Pressman, N. Burkell, and C. D. Rahn, "Data-driven thermal anomaly detection in large battery packs," *Batteries*, vol. 9, no. 2, 2023, Art. no. 70, doi: 10.3390/batteries9020070.

[4] J. Schaeffer, E. Lenz, D. Gulla, M. Z. Bazant, R. D. Braatz, and R. Findeisen, "Gaussian process-based online health monitoring and fault analysis of lithium-ion battery systems from field data," *Cell Rep. Phys. Sci.*, vol. 5, no. 11, 2024, Art. no. 102258, doi: 10.1016/j.xcrp.2024.102258.

[5] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation forest," in *Proc. 8th IEEE Int. Conf. Data Mining*, 2008, pp. 413–422, doi: 10.1109/ICDM.2008.17.

[6] M. Sakurada and T. Yairi, "Anomaly detection using autoencoders with nonlinear dimensionality reduction," in *Proc. 2nd Workshop Mach. Learn. Sensory Data Anal.*, 2014, pp. 4–11, doi: 10.1145/2689746.2689747.

[7] J. Xu, H. Wu, J. Wang, and M. Long, "Anomaly Transformer: Time series anomaly detection with association discrepancy," in *Proc. Int. Conf. Learn. Represent.*, 2022. [Online]. Available: https://openreview.net/forum?id=LzQQ89U1qm_

[8] S. Tuli, G. Casale, and N. R. Jennings, "TranAD: Deep Transformer networks for anomaly detection in multivariate time series data," *Proc. VLDB Endow.*, vol. 15, no. 6, pp. 1201–1214, 2022, doi: 10.14778/3514061.3514067.

[9] Y. Yang, C. Zhang, T. Zhou, Q. Wen, and L. Sun, "DCdetector: Dual attention contrastive representation learning for time series anomaly detection," in *Proc. 29th ACM SIGKDD Conf. Knowl. Discovery Data Mining*, 2023, pp. 3033–3045, doi: 10.1145/3580305.3599295.

[10] X. Liu, X. Li, Y. Li, F. Tang, and M. Zhao, "RTdetector: Deep Transformer networks for time series anomaly detection based on reconstruction trend," in *Proc. 34th Int. Joint Conf. Artif. Intell.*, 2025, pp. 5788–5796, doi: 10.24963/ijcai.2025/644.

[11] W. Yue *et al.*, "Sub-Adjacent Transformer: Improving time series anomaly detection with reconstruction error from sub-adjacent neighborhoods," in *Proc. 33rd Int. Joint Conf. Artif. Intell.*, 2024, pp. 2524–2532, doi: 10.24963/ijcai.2024/279.

[12] X. Wu *et al.*, "CATCH: Channel-aware multivariate time series anomaly detection via frequency patching," in *Proc. Int. Conf. Learn. Represent.*, 2025, pp. 17017–17045. [Online]. Available: https://openreview.net/forum?id=m08aK3xxdJ

[13] S. Wang *et al.*, "TimeMixer++: A general time series pattern machine for universal predictive analysis," in *Proc. Int. Conf. Learn. Represent.*, 2025, pp. 16980–17016. [Online]. Available: https://openreview.net/forum?id=1CLzLXSFNn

[14] Q. Shentu *et al.*, "Towards a general time series anomaly detector with adaptive bottlenecks and dual adversarial decoders," in *Proc. Int. Conf. Learn. Represent.*, 2025, pp. 81358–81381. [Online]. Available: https://proceedings.iclr.cc/paper_files/paper/2025/hash/ca7998666c2e53cc1e882b7268414d8a-Abstract-Conference.html

[15] M. Goswami, K. Szafer, A. Choudhry, Y. Cai, S. Li, and A. Dubrawski, "MOMENT: A family of open time-series foundation models," in *Proc. 41st Int. Conf. Mach. Learn.*, PMLR, vol. 235, 2024, pp. 16115–16152. [Online]. Available: https://proceedings.mlr.press/v235/goswami24a.html

[16] Q. Liu and J. Paparrizos, "The elephant in the room: Towards a reliable time-series anomaly detection benchmark," in *Adv. Neural Inf. Process. Syst.*, vol. 37, 2024, pp. 108231–108261, doi: 10.52202/079017-3437.

[17] H. Li, Z. Liu, M. Bin Kaleem, L. Duan, S. Ruan, and W. Liu, "Fault detection for lithium-ion batteries of electric vehicles with spatio-temporal autoencoder," *Appl. Energy*, vol. 392, 2025, Art. no. 125933, doi: 10.1016/j.apenergy.2025.125933.

[18] W. Nie, Z. Deng, J. Li, K. Zhang, J. Zhou, and F. Xiang, "An early fault detection method of series battery packs based on multi-feature clustering and unsupervised scoring," *Energy*, vol. 323, 2025, Art. no. 135754, doi: 10.1016/j.energy.2025.135754.

[19] L. Zhang, Z. Wu, Y. Wang, C. Zhang, S. Li, and W. Zhang, "Two-tier anomaly detection for real-world data of energy storage battery pack with machine learning and sensitivity-aware feature analysis," *J. Energy Storage*, vol. 139, pt. B, 2025, Art. no. 118905, doi: 10.1016/j.est.2025.118905.

[20] C. Zhang *et al.*, "A multi-scale framework for unsupervised anomaly detection in lithium-ion battery packs," *J. Energy Storage*, vol. 151, 2026, Art. no. 120341, doi: 10.1016/j.est.2026.120341.

[21] W. Li, Y. Yang, and S. Pischinger, "Lithium-ion battery anomaly detection with variational autoencoder based on feature representation learning," *J. Energy Storage*, vol. 152, pt. A, 2026, Art. no. 120545, doi: 10.1016/j.est.2026.120545.

[22] J. Song, W. Liu, J. Li, K. Zhang, H. Yu, and X. Hu, "Unsupervised probabilistic fault detection and risk quantification for lithium-ion batteries in real-world electric vehicles," *Energy Storage Mater.*, vol. 90, 2026, Art. no. 105326, doi: 10.1016/j.ensm.2026.105326.

[23] T. Yin *et al.*, "ScatterAD: Temporal-topological scattering mechanism for time series anomaly detection," in *Adv. Neural Inf. Process. Syst.*, vol. 38, 2025, pp. 58187–58218, doi: 10.52202/085713-1742.

[24] D. Cho, J. Han, K. Kang, M. Kim, H. Ryu, and N. Jung, "Structured temporal causality for interpretable multivariate time series anomaly detection," in *Adv. Neural Inf. Process. Syst.*, vol. 38, 2025, pp. 66011–66048, doi: 10.52202/085713-1983.

[25] X. Sun, H. Zhou, and C. Li, "Multivariate time series anomaly detection with idempotent reconstruction," in *Adv. Neural Inf. Process. Syst.*, vol. 38, 2025, pp. 178665–178723, doi: 10.52202/085713-5374.

[26] D. Luder *et al.*, "Big data generation platform for battery faults under real-world variances," *Green Energy Intell. Transp.*, vol. 4, no. 3, 2025, Art. no. 100282, doi: 10.1016/j.geits.2025.100282.

[27] L. Wang, Z. Hu, A. Tian, C. Chang, and M. Wu, "An inconsistency fault diagnosis method for lithium-ion cells in the battery pack based on piecewise dimensionality reduction and outlier identification," *J. Electrochem. Energy Convers. Storage*, vol. 20, no. 1, 2023, Art. no. 011016, doi: 10.1115/1.4054734.

[28] J. Yu, Y. Guo, and W. Zhang, "Anomaly detection for charging voltage profiles in battery cells in an energy storage station based on robust principal component analysis," *Appl. Sci.*, vol. 14, no. 17, 2024, Art. no. 7552, doi: 10.3390/app14177552.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2026.3753456

[29] Z. Chen, W. Liu, D. Zhou, T. Xia, and E. Pan, "Inconsistency identification for lithium-ion battery energy storage systems using deep embedded clustering," *Appl. Energy*, vol. 388, 2025, Art. no. 125677, doi: 10.1016/j.apenergy.2025.125677.
[30] L. Ma, B. Duan, C. Zhang, Y. Kang, C. Li, and K. Liu, "Detection and differentiation of multiple types of minor anomalies in battery packs," *Energy*, vol. 322, 2025, Art. no. 135557, doi: 10.1016/j.energy.2025.135557.
[31] J. Qu *et al.*, "Diagnosing inconsistencies in battery energy storage systems: A framework integrating electrical, thermal, and aging perspectives," *Appl. Energy*, vol. 405, 2026, Art. no. 127203, doi: 10.1016/j.apenergy.2025.127203.
[32] X. Zhang, Z. Luo, L. Ma, and Y. Pan, "Dynamic transfer diagnosis of long-term service faults in battery packs fusing spatio-temporal features," *Green Energy Intell. Transp.*, early access, Jul. 20, 2026, Art. no. 100447, doi: 10.1016/j.geits.2026.100447.
[33] W. Wang *et al.*, "A multi-model fusion framework for battery system fault diagnosis in real vehicles," *Green Energy Intell. Transp.*, early access, May 15, 2026, Art. no. 100429, doi: 10.1016/j.geits.2026.100429.
[34] B. Hariprakash, S. K. Martha, A. Jaikumar, and A. K. Shukla, "Online monitoring of lead–acid batteries by galvanostatic non-destructive technique," *J. Power Sources*, vol. 137, no. 1, pp. 128–133, 2004, doi: 10.1016/j.jpowsour.2004.05.045.
[35] P. Křivík and P. Vanýsek, "Changes of temperature during pulse charging of lead acid battery cell in a flooded state," *J. Energy Storage*, vol. 14, pt. 3, pp. 364–371, 2017, doi: 10.1016/j.est.2017.03.018.
[36] V. Sulzer, S. G. Marquis, R. Timms, M. Robinson, and S. J. Chapman, "Python Battery Mathematical Modelling (PyBaMM)," *J. Open Res. Softw.*, vol. 9, no. 1, 2021, Art. no. 14, doi: 10.5334/jors.309.
[37] V. Sulzer, S. J. Chapman, C. P. Please, D. A. Howey, and C. W. Monroe, "Faster lead-acid battery simulations from porous-electrode theory: Part I. Physical model," *J. Electrochem. Soc.*, vol. 166, no. 12, pp. A2363–A2371, 2019, doi: 10.1149/2.0301910jes.
[38] V. Sulzer, S. J. Chapman, C. P. Please, D. A. Howey, and C. W. Monroe, "Faster lead-acid battery simulations from porous-electrode theory: Part II. Asymptotic analysis," *J. Electrochem. Soc.*, vol. 166, no. 12, pp. A2372–A2382, 2019, doi: 10.1149/2.0441908jes.



SHIBO LUO received the bachelor's degree in computer science and engineering from East China University of Science and Technology (ECUST), Shanghai, China. He is currently an Engineer with the Informatization and Data Management Center, ECUST. His research interests include intelligent perception and fault detection.



BIN SHEN (Member, IEEE) received the M.Eng. degree in control engineering from East China University of Science and Technology (ECUST), Shanghai, China. He is currently an Engineer with the Informatization and Data Management Center and the School of Information Science and Engineering, ECUST. His research interests include deep learning and artificial intelligence.



XIAONING CHEN received the M.Eng. degree in computer engineering from East China University of Science and Technology (ECUST), Shanghai, China. He is currently a Senior Engineer with the Informatization and Data Management Center, ECUST. His research interests include artificial intelligence and information security.



YIQUAN FANG received the M.Eng. degree in computer science and engineering from East China University of Science and Technology (ECUST), Shanghai, China. She is currently a Professor-Level Senior Engineer with the Informatization and Data Management Center, ECUST. Her research interest is information security.



HUIFENG WANG received the Ph.D. degree in control science and engineering from East China University of Science and Technology (ECUST), Shanghai, China. She is currently a Professor with the School of Information Science and Engineering, ECUST. Her research interest is intelligent perception.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/