CULPRIT: LOCALIZING MECHANISM SHIFTS FROM SHARED SENSORS UNDER VARYING CONDITIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Root cause analysis (RCA) for industrial systems must name the faulty component from sensor data. Because sensors typically sit between components, ranking them as root causes mistakes the site of measurement for the site of failure. Our key insight is that a fault is a mechanism shift on a component-level latent variable. Under this view, a system's sensor-to-component association and its operating condition together can localize that shift from sensor data. We introduce CULPRIT, an unsupervised approach to pinpoint the root cause in industrial systems under varying conditions. It matches each fault sample to healthy samples at its operating point and projects the matched residuals onto one latent per component through the association. It then separates root causes from downstream effects along the component edges. We evaluate CULPRIT against 15 baselines on six benchmarks, where it holds the best average Top-1 accuracy of 0.52. We further show that operating-condition adjustment and the sensor-to-component association are each necessary. Matching at the operating point lifts Top-1 accuracy from 0.35 to 0.68 on N-CMAPSS, and one latent per component, instead of a sensor maximum, lifts the average Top-1 from 0.42 to 0.52. Code is available at https://anonymous.4open.science/r/culprit-DC73.

1 INTRODUCTION

Root cause analysis (RCA) answers one question after a failure, namely which component of the system caused it (Pham et al., 2024b). For a gas turbine the answer is a high-pressure turbine or a compressor (Arias Chao et al., 2021), and for a building air handler it is a damper, a cooling coil, or a fan (Granderson et al., 2023). Sensors rarely measure one component alone, because many are installed in between two components, and their measurements reflect the state of both. For example, the gas temperature sensor at the exit of the high-pressure turbine also measures the inlet temperature of the low-pressure turbine (Simon, 2010). When such a sensor deviates, its value alone cannot say which of the two components changed. The operating condition adds a second challenge (Song et al., 2007; Wang et al., 2008). An aircraft climbing to cruise moves all gas-path sensors without any component degrading (Arias Chao et al., 2021). An RCA method must therefore take the sensor data and the operating conditions together and derive from them the set of root cause components.

Prior work addresses parts of these challenges, but does not combine them in a component-level diagnosis. Deviation-based and neural localization methods rank sensors from their deviations or learned attributions (Shan et al., 2019; Pham et al., 2024a; Shimillas et al., 2026). Because faults propagate, the highest-scoring sensor may record a downstream effect instead of the component whose mechanism shifted. Causal methods use learned or supplied structure (Ikram et al., 2022; Blöbaum et al., 2024; Orchard et al., 2025; Lee et al., 2026), but their diagnoses remain tied to graph vertices at the granularity of that structure. They do not recover component state through a sensor-to-component association in which one sensor may straddle several components. A third line accounts for changes in system behaviour. CondAttr (Mishra et al., 2026) retrieves similar healthy windows in a learned sensor representation, StaR (Huang et al., 2026) retains recent behaviour, and StableRCA (Lin et al., 2026) tests local mechanism changes after accounting for neighbouring variables. These methods derive their comparison from the sensors and return sensor-level diagnoses.

In this paper, we introduce CULPRIT, a training-free method that localizes mechanism shifts on component-level latent variables from sensor data (Figure 1). Our key observation is that an indus-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

trial system already supplies both inputs this task requires. The sensor-to-component association indicates which sensors a component's latent may load on, the sparsity pattern of the mixing that sparsity-based identifiability results rest on (Zheng et al., 2022; Yao et al., 2024). The operating condition is where a mechanism shift must be read, since a fault changes how a component responds at a given operating point. CULPRIT realizes this observation in three stages. Stage 1 handles the operating condition by matching each fault sample to the healthy samples at its operating point, which bounds the condition confound linearly in the matching radius when the healthy response is Lipschitz (Proposition 3.1). Stage 2 handles sensor sharing by projecting each component's matched residuals onto one latent under a linear mixing model on the association. The projection direction is the first principal direction of the component's healthy residuals, found by principal component analysis (PCA). Stage 3 scores each latent's displacement over the fault window, separates root causes from downstream effects, and ranks the components with one evidence line each.

We evaluate CULPRIT extensively against 15 RCA baselines on six benchmarks, spanning turbofans, buildings, a wind tunnel, a water plant, and microservices, with a total of 1,261 failure cases. CULPRIT holds the best Top-1 on all six benchmarks. The ablations isolate the contributions of operating-condition adjustment and the system model. The results show that Stage 1 matching raises Top-1 accuracy from 0.35 to 0.68 on N-CMAPSS, and separation along the component edges raises overall Top-1 accuracy from 0.56 to 0.77 on the LBNL FDD benchmark.

Our contributions are as follows:

- We formalize component-level RCA from shared sensors as shift localization on latent variables (Section 2). We prove when the separated displacement and the velocity score rank the root causes first, for faults that move a component's state or its rate of change (Theorem 3.2, Proposition C.12).
- We present CULPRIT, to the best of our knowledge the first RCA method to combine pointwise condition matching, latent recovery through a sensor-to-component association and structural separation along component edges, all in closed form without training (Section 3).
- We benchmark CULPRIT against 15 RCA methods on six benchmarks, where it reaches the best Top-1 on each benchmark and the best average Top-1 of 0.52, and show that CULPRIT needs both operating-condition adjustment and the sensor-to-component association (Sections 4.2 and 4.3).

2 Component-Level RCA as Latent Shift Localization

We now define the objects this paper works with and state the diagnosis problem formally. Throughout, components are indexed by c, sensors by s, time by t, and environments by e.

System model. A system consists of m components, collected in $\mathcal{C} = \{1, \dots, m\}$, and n sensors. A *system model* supplies the component list $\mathcal{C}$, a sensor-to-component association $A \in \{0,1\}^{n \times m}$, and optionally a set of edges $\mathcal{E} \subseteq \mathcal{C} \times \mathcal{C}$ stating which components feed which. An entry $A_{s,c} = 1$ indicates that sensor s measures a quantity that component c shapes. The association fixes only which entries of the mixing matrix W of Section 3.2 are nonzero. The magnitudes are estimated from data. A *local* sensor has exactly one associated component, while a *shared* sensor straddles several components and mixes their states. One system can contain both kinds.

Case. We diagnose one case at a time. A case supplies sensor time series $x_{1:T} \in \mathbb{R}^{T \times n}$, optional condition channels $u_{1:T} \in \mathbb{R}^{T \times q}$, a fault onset $t_0 \in \{1, \dots, T\}$, and the system model above. The condition channels describe the system's operating point (e.g., an aircraft's altitude, a microservice's offered load). The onset splits the series into a healthy window $t < t_0$ and a fault window $t \geq t_0$.

Generative view. At time t, each component c has one scalar *latent* $z_t^{(c)} \in \mathbb{R}$. It summarizes the component-specific state variables that jointly influence the measurements of its sensors. The full latent state concatenates these latents, $z_t = (z_t^{(1)}, \dots, z_t^{(m)})$. The healthy and fault windows provide finite samples from two *environments* (Chen et al., 2023; 2024), corresponding respectively to the system before the fault ($e = 0$) and after it ($e = 1$). The method computes statistics over the observed windows, whereas mechanisms and expectations are indexed by environment. Within an environment, the latent of component c is generated by its own *mechanism* (Schölkopf et al., 2021; Budhathoki et al., 2021). This mechanism maps the states of c's parents, the operating condition, and an exogenous process to that latent,

$$z_t^{(c)} = f_e^{(c)}\big(z_t^{\mathrm{Pa}(c)},\, u_t,\, \zeta_t^{(c)}\big). \tag{1}$$

Here $\mathrm{Pa}(c)$ denotes the parents of c, u_t is the operating condition, and $\zeta^{(c)}$ is an exogenous process. The exogenous processes are independent across components but may be autocorrelated. This independence expresses causal sufficiency conditional on u_t (Spirtes et al., 2001; Perry et al., 2022), meaning that every common cause of two latents is either u_t or a component in $\mathcal{C}$. We treat u_t as exogenous because quantities such as altitude and outdoor temperature are set outside the system and unaffected by faults. The declared edges may form loops, such as an air handler's return-air path. The reading of sensor s depends on u_t and only on the latents allowed by its association row,

$$x_{t,s} = g_s\big(z_t^{(A_s)},\, u_t\big) + \varepsilon_{t,s}, \qquad A_s = \{c \in \mathcal{C} : A_{s,c} = 1\}, \tag{2}$$

where g_s is an unknown mixing function and $\varepsilon_{t,s}$ is sensor noise. The operating condition can act on a sensor directly and through the latents, e.g., the fan-inlet pressure on N-CMAPSS, which altitude and Mach number alone set. Stage 1 compares each reading with healthy readings at its operating condition, which removes the healthy mean effect of both routes. Neither the latent state, the mechanisms, nor the mixing is observed. The association and edges come with the system, e.g., gas-path station numbers, equipment schemas, instrument tags or the service name in a metric name.

Definition 2.1 (Fault and root cause set)**.** A *fault* is a mechanism shift (Schölkopf et al., 2021; Budhathoki et al., 2021) on one or more components. For each such c, the shift changes the map $f^{(c)}$ or the law of $\zeta^{(c)}$ in Equation 1. The *root cause set* of a case is the set of shifted components,

$$S^\star = \big\{\, c \in \mathcal{C} : f_1^{(c)} \neq f_0^{(c)} \;\text{ or }\; P_1\big(\zeta_{1:T}^{(c)}\big) \neq P_0\big(\zeta_{1:T}^{(c)}\big) \,\big\},$$

where the maps are compared on the operating conditions both windows visit.

A physical component can have several state variables, e.g., the efficiency and the flow capacity of a turbine module (Li, 2002). Prognostics summarizes a component's condition by one health index (Wang et al., 2008; Chao et al., 2022), because degradation moves those variables together and a diagnosis names the component as a whole. For instance, N-CMAPSS simulates most failure modes as a joint loss of a module's efficiency and flow capacity (Arias Chao et al., 2021). One latent per component follows that practice, and Appendix G.4 also evaluates two latents per component.

When the edges form no loop and the parents are non-degenerate, $S^\star$ is the set of components whose law given their parents and the operating condition differs between the fault environment and the healthy one, the shifted-node estimand of Chen et al. (2024). On a loop, a shift of one member moves the conditional law of each member, and $S^\star$ names the shifted member alone. A component outside $S^\star$ can still deviate strongly, because a shifted ancestor feeds it through Equation 1 and the deviation propagates along the edges. We call such a component a *downstream effect*. Given a case and its system model, the component-level RCA problem is to return a ranking over $\mathcal{C}$ that places $S^\star$ above all other components, the downstream effects included.

Two properties that make the problem hard. First, $S^\star$ is defined on latent variables that cannot be observed, and a shared sensor mixes several components' states. Hence, per-sensor deviation scores cannot express a component-level answer. Second, the operating condition u_t acts on all sensors through Equations 1 and 2 even when all mechanisms stay fixed, which makes a change in demand and a change in health indistinguishable in raw deviations. A solution must therefore recover component-level state from mixed sensors and must contrast the two environments at matched operating conditions. The method of Section 3 is built on these two requirements.

3 CULPRIT: From Matched Residuals to Root-Cause Rankings

We propose CULPRIT, a three-stage RCA method that meets the two requirements of Section 2. Figure 1 runs the three stages on the turbofan, our running example. All proofs are in the appendix.

3.1 Stage 1: Condition Matching

CULPRIT contrasts each case against a healthy reference restricted to the same operating conditions.

Building the reference. Condition channels $u_{1:T}$ are the channels a case marks as inputs. They describe the demand on the system instead of its response and lie outside the sensor set $x_{1:T}$. Let $\mathcal{T}_0 = \{t : t < t_0\}$ be the healthy samples, $\mathcal{T}_1 = \{t : t \geq t_0\}$ the fault window, and $\tilde{u}_t = (u_t - \mu_0)/\sigma_0$ the condition channels standardized by the healthy moments. The *condition support*

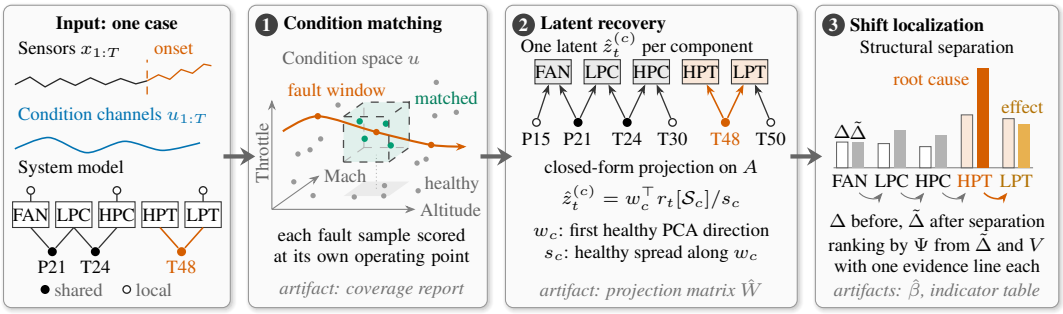


Figure 1: **CULPRIT localizes the mechanism shift behind shared sensors.** On the N-CMAPSS turbofan, orange marks the fault and the shared sensor T48, and blue marks the condition channels $u_{1:T}$. Stage 1 matches each fault sample to healthy samples in a Chebyshev ball around its operating point (green). Stage 2 projects each component's residuals onto one latent, which T48 displaces for HPT and LPT alike. Their displacement scores Δ (hollow) nearly tie until the structural separation of Stage 3 lifts HPT, the root cause, above its downstream effect LPT (filled, $\tilde{\Delta}$).

$\mathcal{U}_1 = \{\tilde{u}_t : t \in \mathcal{T}_1\}$ is the region the system moves through while the fault is present, and $\mathcal{U}_0$ is the healthy support. The condition-matched reference of a sample t collects the healthy samples inside a Chebyshev ball of radius τ around its operating point,

$$\mathcal{M}_\tau(t) = \{\, t' \in \mathcal{T}_0 \;:\; \|\tilde{u}_{t'} - \tilde{u}_t\|_\infty \le \tau \,\}. \tag{3}$$

The ℓ_∞ ball is a conjunction of per-channel tolerances. A healthy sample therefore qualifies only when it agrees with the fault sample on all condition dimensions, on altitude and throttle alike.

Pointwise statistic. A per-sample reference tracks the case's trajectory through condition space. Let $\hat{\mu}_s(t)$ and $\hat{\sigma}_s(t)$ be the mean and standard deviation of sensor s over $\mathcal{M}_\tau(t)$. The adjusted residual and the resulting sensor score are

$$r_{t,s} = \frac{x_{t,s} - \hat{\mu}_s(t)}{\hat{\sigma}_s(t)}, \qquad d_s = \max_{t \in \mathcal{T}_s} |r_{t,s}|, \tag{4}$$

where $\mathcal{T}_s \subseteq \mathcal{T}_1$ holds the fault samples on which sensor s is *scorable*. Sensor s is scorable at sample t when s is recorded there and $\mathcal{M}_\tau(t)$ holds at least $n_{\min}$ healthy samples recording s. The reference spread $\hat{\sigma}_s(t)$ must also be bounded away from zero. Applied to each healthy sample against its ball, Equation 4 yields the healthy-window residuals Stage 2 needs.

The radius τ trades condition mismatch against reference size (Proposition 3.1).

Proposition 3.1 (Matching error)**.** *Fix a sensor s and let $h_s(u) = \mathbb{E}[x_{t,s} \mid \tilde{u}_t = u, e = 0]$ be its healthy response. Assume that h_s is quasi-static, $\mathbb{E}[x_{t',s} \mid \tilde{u}_{1:T}, e = 0] = h_s(\tilde{u}_{t'})$, and L_s-Lipschitz in $\|\cdot\|_\infty$. Assume further that the healthy noise $x_{t',s} - h_s(\tilde{u}_{t'})$ has, given $\tilde{u}_{1:T}$, variance at most σ_s^2 and absolute autocorrelations summing to at most χ_s. Then each fault sample t with $N = |\mathcal{M}_\tau(t)| \ge n_{\min}$ satisfies, with probability at least $1 - \delta$,*

$$\left|\hat{\mu}_s(t) - h_s(\tilde{u}_t)\right| \;\le\; L_s\tau + \sigma_s\sqrt{(1 + 2\chi_s)/(\delta N)}.$$

The condition-induced term of the residual numerator is therefore at most $L_s\tau$ plus a sampling term that shrinks as the matched set grows. One reference over the whole healthy window instead leaves a condition-induced term that can reach the order of $\mathrm{osc}_s = \sup_{u,u' \in \mathcal{U}_0 \cup \mathcal{U}_1} |h_s(u) - h_s(u')|$, the full swing of the healthy response across both windows. On a flight, it is the gap between sea level and cruise altitude and hides any degradation signature. Both references share the mechanism-induced term. Section 4.4 measures the shrinkage and evaluates a fault-free rule for different τ values.

Once τ reaches the Chebyshev diameter of $\mathcal{U}_0 \cup \mathcal{U}_1$, each matched set equals $\mathcal{T}_0$ and d_s becomes the unadjusted maximum z-score (Corollary B.1). The whole-window reference is therefore the wide-radius end of a single continuum, and a case without condition channels sits at that end.

3.2 Stage 2: Latent Recovery through the Sensor-to-Component Association

Stage 2 recovers one state per component in closed form through the association A. Its input is the residuals $r_{t,s}$ of Equation 4 over $\mathcal{T}_0$ and $\mathcal{T}_1$. We write $r_t \in \mathbb{R}^n$ for the residuals at sample t, $r_t[\mathcal{S}]$

for their restriction to a sensor set $\mathcal{S}$, and $W[\mathcal{S}, c]$ for the rows $\mathcal{S}$ of column c of W. An unscorable entry takes the value of its healthy reference, zero (Appendix E).

Linear mixing at the operating point. Stage 2 replaces the mixing of Equation 2 on these residuals by the linear mixing model,

$$r_t = W z_t + \nu_t, \qquad W \in \mathbb{R}^{n \times m}, \qquad W_{s,c} = 0 \text{ whenever } A_{s,c} = 0, \tag{5}$$

where ν_t is the residual noise, independent of z_t. Because each residual is taken against healthy samples at the same operating point, Equation 5 is the first-order expansion of g_s around the healthy state there, with z_t the deviation from it. Gas-path analysis (Simon et al., 2022), multistage assembly diagnosis (Chung et al., 2024) and structural damage localization (Allahdadian et al., 2019) rest on the same expansion, and recent root-cause discovery models a fault as a mean shift that propagates through a linear structural equation model (Li et al., 2026). Linear models of multivariate sensor data also match or outperform deep detectors (Audibert et al., 2022; Sarfraz et al., 2024).

Let $\mathcal{S}_c = \{s : A_{s,c} = 1\}$ be the sensors assigned to component c and a_c the entries of column c of W on those sensors. Restricted to $\mathcal{S}_c$, Equation 5 reads $r_t[\mathcal{S}_c] = a_c z_t^{(c)} + \eta_t^{(c)}$, where $\eta_t^{(c)}$ collects the sensor noise and the contributions of the other components sharing those sensors. This is a one-factor model on the sensors of one component with loading vector a_c.

One latent per component. Stage 2 reduces only the sensors in $\mathcal{S}_c$ to the latent of c, which ties it to component c instead of to a direction mixing all components. The sensors of a component move together when its state changes. CULPRIT fits the direction of that common movement on healthy data alone, and the fault window therefore cannot shape the weights. Let Σ_c be the covariance of $r_t[\mathcal{S}_c]$ over $\mathcal{T}_0$. The projection direction w_c is the first principal direction of PCA on the healthy residuals of $\mathcal{S}_c$, i.e., the direction along which they vary most. The latent of c is the standardized projection on it,

$$\hat{z}_t^{(c)} = \frac{w_c^\top r_t[\mathcal{S}_c]}{s_c}, \qquad w_c = \arg\max_{\|w\|=1} w^\top \Sigma_c w, \qquad s_c = \sqrt{\lambda_1(\Sigma_c)}, \tag{6}$$

with the sign of w_c fixed so that $\mathbf{1}^\top w_c \geq 0$, and $s_c = 1$ for a component with no healthy spread. Dividing by s_c, the healthy standard deviation of the projection, gives $\hat{z}_t^{(c)}$ unit variance on $\mathcal{T}_0$ and puts all components in healthy-sigma units. These common units let Stage 3 rank the components against one another. Stacking the m directions into $\hat{W}$ makes the whole latent state one matrix product per sample, $\hat{z}_t = D^{-1}\hat{W}^\top r_t$ with $D = \mathrm{diag}(s_1, \ldots, s_m)$.

Why the first principal direction? When all sensors of c are local, $\eta_t^{(c)}$ is sensor noise alone, and $w^\top r_t[\mathcal{S}_c] = (w^\top a_c)\, z_t^{(c)} + w^\top \eta_t^{(c)}$ for a unit-norm w. Any w with $w^\top a_c \neq 0$ thus recovers $z_t^{(c)}$ up to scale, and w sets only how much sensor noise enters. Under isotropic residual noise, the first principal direction is parallel to a_c (Proposition C.10). Unlike equal weights, it can give a negative weight to a sensor that moves against the other sensors of c.

Shared sensors. A shared sensor lies in the sensor set of each component it is assigned to. Then $\eta_t^{(c)}$ also carries the latent of each component c' sharing a sensor with c, weighted by $w_c^\top W[\mathcal{S}_c, c']$, the *smearing* of c' into c. On N-CMAPSS, HPT has one sensor, T48, which straddles HPT and LPT (Figure 1). Hence $w_c = 1$, and a fault on HPT displaces the latents of both components. Theorem 3.2 states when the separated displacement survives this smearing (Li, 2002).

3.3 Stage 3: Shift Localization and Ranking

At diagnosis time, CULPRIT contrasts each latent over the fault window against its healthy distribution and turns the contrast into one score per component, into which three quantities enter.

Displacement score. The displacement score of component c is the mean absolute displacement of its latent from the healthy centre over the fault window, $\Delta(c) = |\mathcal{T}_1|^{-1} \sum_{t \in \mathcal{T}_1} |\hat{z}_t^{(c)}|$. A fault is a sustained change of a component's mechanism, hence the signature is how far the latent stays displaced over the window. A maximum over the window cannot tell that from a single spike.

Structural separation. A high displacement on a component whose parents are also high is often a downstream effect instead of a root cause. CULPRIT separates the two on the latents, before scores

are computed. For component c with parents $\mathrm{Pa}(c)$, CULPRIT regresses its latent on its parents' latents over the healthy window. It then replaces its latent by the standardized structural residual

$$\tilde{z}_t^{(c)} = \frac{1}{\tilde{s}_c}\Big(\hat{z}_t^{(c)} - \sum_{c' \in \mathrm{Pa}(c)} \hat{\beta}_{cc'}\, \hat{z}_t^{(c')}\Big), \tag{7}$$

where $\hat{\beta}_{cc'}$ is the ridge coefficient of parent c' in that regression. The spread $\tilde{s}_c$ is the healthy standard deviation of that structural residual, which measures component c's state given its parents' states under the healthy law. With no smearing and exact coefficients, it equals $\zeta^{(c)}$ plus noise, which a shift on an ancestor leaves untouched (Proposition C.6). A component with no parent keeps its latent, and the separated displacement $\tilde{\Delta}(c)$ is the same window mean of $|\tilde{z}_t^{(c)}|$. A *displacement fault* translates the exogenous process of each root cause by a constant or a time-varying amount (Assumption C.3). Under a constant one, $\tilde{z}_t^{(c)}$ in the fault window is a constant shift ξ_c plus a null part of mean zero (Lemma C.5). The components are *equally noisy* when their null parts share one symmetric law with zero in its support.

Theorem 3.2 (Ranking under bounded smearing)**.** *Assume linear mixing, linear mechanisms on the declared edges and a constant displacement fault (Assumptions C.1 to C.3). (i) If the components are equally noisy, the population separated displacement ranks $S^\star$ first iff* $\min_{r\in S^\star}|\xi_r| > \max_{c\notin S^\star}|\xi_c|$. *(ii) If the null parts have variance at most $\bar{\sigma}^2$, it does so if* $\min_{r\in S^\star}|\xi_r| > \max_{c\notin S^\star}(\bar{\sigma}^2 + \xi_c^2)^{1/2}$.

The shift ξ_c is c's own term, zero outside $S^\star$, plus the smearing from components sharing its sensors and from the error of the fitted coefficients (Proposition C.7), and without smearing the condition of (i) always holds. Appendix C extends (i) to growing shifts such as wear, justifies keeping the shared sensors of a component, which carry its state, and shows that a projection on local sensors reads no other component (Corollary C.8).

Velocity score. Some faults leave a component's displacement normal but change how fast its sensors move from sample to sample. This change in the increment law of $\zeta^{(c)}$ is a mechanism shift under Definition 2.1. We call such a fault a *velocity fault*. A stuck damper, whose sensors stop moving, is the typical example. The separated displacement of Theorem 3.2 reads displacement faults, and Proposition C.12 gives when the velocity score ranks the component of a velocity fault first. We build the velocity score from two absolute log ratios of first-difference variance,

$$V^\circ(c) = \max\Big(\rho\big(w_c^\top \tilde{x}[\mathcal{S}_c]\big),\ \max_{s\in\mathcal{S}_c}\frac{\rho(\tilde{x}_s)}{|A_s|}\Big), \qquad \rho(y) = \Big|\log\frac{\mathrm{Var}_{t\in\mathcal{T}_1}\big(y_t - y_{t-1}\big)}{\mathrm{Var}_{t\in\mathcal{T}_0}\big(y_t - y_{t-1}\big)}\Big|, \tag{8}$$

where $\tilde{x}_{t,s}$ is sensor s's signal standardized by its healthy moments. Flooring each variance at 10^{-12} makes ρ large but finite on a frozen signal. The ratio ρ is invariant to the scale s_c. It uses the standardized signals instead of the matched residuals because a stuck device freezes its sensors, while the condition-matched reference would put the condition's movement back into the frozen signal. The healthy and fault windows see similar condition dynamics, hence ρ needs no correction for them (Appendix E). The latent form responds to a change in the whole component's dynamics, and the per-sensor form to a single sensor whose movement collapses or explodes, which the latent form averages away. Dividing by $|A_s|$, the number of components sensor s is assigned to, counts a local sensor in full and splits a shared sensor's evidence equally among its components. Sampling noise moves a ratio of sample variances away from one even under an unchanged increment law. Each term of Equation 8 therefore counts only above that noise. For a term $\rho(y)$ over n_0 healthy and n_1 fault-window increments $d_t = y_t - y_{t-1}$, the delta method gives the null standard deviation of its log ratio

$$\hat{\sigma}_\rho(y) = \Big(\big(\tfrac{1}{n_0} + \tfrac{1}{n_1}\big)\frac{\mathrm{LRV}\big((d-\bar{d})^2\big)}{\mathrm{Var}(d)^2}\Big)^{1/2}, \tag{9}$$

where LRV is the Newey–West estimate of the long-run variance of the squared centred healthy increments (Newey & West, 1987), with $\bar{d}$ their healthy mean. The velocity score $V(c)$ is the largest term of Equation 8 with $\rho(y) > \Phi^{-1}(1-\alpha/2J)\,\hat{\sigma}_\rho(y)$, and zero when no term passes. Here Φ is the standard Gaussian distribution function, J counts the terms of all components, and $\alpha = 0.01$.

Combined score and ranking. Displacement and velocity are divided by their median over components to read as how far a component stands above a typical one, and joined by the maximum,

$$\Psi(c) = \max\Big(\frac{\tilde{\Delta}(c)}{\mathrm{med}^+_{c'}\tilde{\Delta}(c')},\ \frac{V(c)}{\mathrm{med}^+_{c'}V^\circ(c')}\Big). \tag{10}$$

Here med^+ is the median over the positive values, and one when no value is positive, which makes both denominators positive. The velocity median runs over the ungated scores. The gate therefore removes a component's velocity score without rescaling the others. The maximum takes whichever of the two sees the fault at full strength. A rank average would instead halve a signature the other misses. When the fault adds a constant to the standardized signals, the gate lets the maximum keep the ranking of Theorem 3.2 with probability at least $1 - \alpha - o(1)$ (Proposition 3.3).

Proposition 3.3 (The gate keeps the displacement ranking)**.** *Let the fault window's standardized signals be, jointly in time, a process with the healthy law plus a constant. Let their increments be stationary and weakly dependent (full conditions in Appendix C). As $n_0, n_1 \to \infty$, all velocity scores are zero with probability at least $1 - \alpha - o(1)$, and Ψ then orders the components as $\tilde{\Delta}$ does. If instead a term's increment variance moves from $v_0 > 0$ to a fixed $v_1 \neq v_0$, zero included as on a stuck device, that term passes the gate with probability tending to one.*

4 Evaluation

4.1 Experimental Setup

Benchmarks. We evaluate CULPRIT on six benchmarks with 1,261 failure cases in total. RCAEval (Pham et al., 2025) has 735 microservice cases, OpenRCA (Xu et al., 2025) has 330 production cases from three systems (Bank, Telecom and Market), and SWaT (Goh et al., 2016) has 20 attacks on a water plant. Causal Chamber (Gamella et al., 2025b) has 28 wind-tunnel interventions, N-CMAPSS (Arias Chao et al., 2021) has 99 turbofan engines, and LBNL FDD has 49 cases from two building-HVAC datasets, SDAHU (12) and DDAHU (37). N-CMAPSS, LBNL FDD and OpenRCA record the operating condition. Appendix D describes each benchmark.

Baselines. We compare CULPRIT against 15 RCA methods: CIRCA (Li et al., 2022), DoWhy-GCM (Budhathoki et al., 2022; Blöbaum et al., 2024), Traversal (Chen et al., 2014), Smooth Traversal (Orchard et al., 2025), RCD (Ikram et al., 2022), LatentScope (Xie et al., 2024), BARO (Pham et al., 2024a), AERCA (Han et al., 2025), StaR (Huang et al., 2026), RUN (Lin et al., 2024), Cholesky (Li et al., 2026), Score Ordering (Orchard et al., 2025), CondAttr (Mishra et al., 2026), ALoRa (Shimillas et al., 2026), BRCD (Lee et al., 2026). Details are in Appendix F.

Metric. We adopt the standard Top-k and Avg@k with one change for tied rankings, i.e., two ranked components has the same score. Therefore, we report *Expected Top-k*, the expectation over a uniformly random ordering of equally scored components, following (McSherry & Najork, 2008),

$$\text{Top-}k = \frac{1}{|\mathcal{F}|} \sum\nolimits_{j \in \mathcal{F}} \mathbb{E}_\pi \left[\frac{\sum_{i \leq k} \mathbb{1}\left\{ R_\pi^j[i] \in S_j^\star \right\}}{\min(k, |S_j^\star|)} \right], \tag{11}$$

where $\mathcal{F}$ is the set of cases, $S_j^\star$ the true root cause set of case j, and R_π^j its ranking under tie-break π. Avg@k is the mean of Top-1 through Top-k. Other implementation details are in Appendix E.

4.2 Effectiveness

Table 1 presents the overall results, with the benchmarks grouped by whether they record an operating condition and declared edges. Appendix G.1 gives the detailed results.

Overall Performance. CULPRIT holds the best Top-1 accuracy on all six benchmarks. On the two N-CMAPSS and LBNL FDD benchmarks, it scores 0.68 and 0.77 Top-1, against 0.58 (CIRCA) and 0.71 (BRCD) for the strongest baseline. On the RCAEval and SWaT benchmarks, which do not provide operating condition or declared edges, CULPRIT shares the best Top-1 accuracy on both, at 0.71 with LatentScope and at 0.25 with Cholesky, CIRCA and RCD.

Performance under Sensor Sharing. On the Causal Chamber benchmark, ten of its eleven sensors mix the influence of several components, i.e., actuators. CULPRIT reaches 0.39 Top-1 accuracy on its 28 failure cases while the baselines sit between 0.01 and 0.24. The baselines take a maximum over a component's sensors, which lifts all actuators sharing a sensor equally.

Performance at Different Granularity. OpenRCA labels root causes at three granularities, i.e., nodes, pods and services. Components at different levels are assigned different numbers of sensors, and a maximum over sensors can land on the component assigned the most. Stage 2 removes this bias by scoring each latent in the units of its healthy spread (Appendix G.7). Using the sensor maximum instead drops CULPRIT's performance on OpenRCA from 0.34 to 0.22 Top-1 (Table 13).

Table 1: **CULPRIT holds the best Top-1 on six benchmarks.** Best in bold, second best underlined.

	With operating condition and declared edges						Without operating condition or declared edges					
	N-CMAPSS		LBNL FDD		OpenRCA		Causal Chamber		RCAEval		SWaT	
Method	Top-1	Avg@3	Top-1	Avg@3	Top-1	Avg@3	Top-1	Avg@3	Top-1	Avg@3	Top-1	Avg@3
AERCA	0.43	0.52	0.10	0.20	0.05	0.09	0.10	0.20	0.16	0.24	0.13	0.19
ALoRa	0.43	0.53	0.09	0.17	0.04	0.08	0.08	0.16	0.07	0.14	0.08	0.09
BARO	0.50	0.62	0.07	0.28	0.14	0.23	0.15	0.30	0.23	0.52	0.15	0.27
BRCD	0.57	0.71	0.71	0.83	0.12	0.25	0.24	0.49	0.30	0.48	0.15	0.18
Cholesky	0.47	0.61	0.52	0.66	0.16	0.21	0.13	0.29	0.06	0.11	**0.25**	0.37
CIRCA	0.58	0.70	0.23	0.35	0.19	0.27	0.15	0.31	0.51	0.70	**0.25**	0.30
CondAttr	0.57	0.71	0.49	0.68	0.33	**0.47**	0.18	0.33	0.69	**0.81**	0.23	**0.42**
DoWhy-GCM	0.38	0.48	0.05	0.19	0.17	0.22	0.01	0.07	0.50	0.53	0.10	0.16
LatentScope	0.44	0.62	0.33	0.33	0.24	0.37	0.15	0.30	**0.71**	**0.81**	0.10	0.27
RCD	0.46	0.57	0.05	0.11	0.08	0.13	0.05	0.10	0.33	0.38	**0.25**	0.28
RUN	0.48	0.57	0.03	0.07	0.02	0.05	0.11	0.21	0.09	0.17	0.02	0.05
Score Ordering	0.54	0.59	0.46	0.59	0.03	0.07	0.09	0.18	0.06	0.11	0.19	0.26
Smooth Traversal	0.35	0.53	0.16	0.20	0.14	0.21	0.01	0.03	0.05	0.09	0.10	0.17
StaR	0.48	0.54	0.03	0.07	0.04	0.07	0.08	0.17	0.10	0.17	0.02	0.04
Traversal	0.36	0.52	0.09	0.16	0.15	0.21	0.02	0.05	0.04	0.09	0.09	0.17
CULPRIT	**0.68**	**0.75**	**0.77**	**0.86**	**0.34**	0.45	**0.39**	**0.57**	**0.71**	0.80	**0.25**	0.33

Table 2: CULPRIT outperforms the strongest baseline with and without Stage 1 adjustment.

	Strongest baseline, raw		Strongest baseline, adjusted		
Benchmark	Top-1	Method	Top-1	Method	CULPRIT
N-CMAPSS	0.58	CIRCA	0.63	DoWhy-GCM	**0.68**
LBNL FDD	0.52	Cholesky	0.37	Smooth Traversal	**0.77**
OpenRCA	0.24	LatentScope	0.28	DoWhy-GCM	**0.34**

4.3 Ablation Study

In this section, we ablate CULPRIT to verify that operating-condition adjustment, the sensor-to-component association and other design decisions are contributing positively.

Operating-Condition Adjustment. We compare three references for the healthy behaviour, i.e., no adjustment, case-level matching, and pointwise matching (Appendix G.3). On N-CMAPSS, pointwise matching achieves 0.68 Top-1, against 0.35 without adjustment and 0.36 with case-level matching. Because case-level matching draws a single reference from the healthy samples inside the fault window's condition support, the improvement comes from the granularity of the reference.

Sensor-to-Component Association. We randomly assign a fraction of the sensors to a wrong component and measure the performance of CULPRIT (Appendix G.8). Errors about other components matter little, since the effect concentrates on the cases whose own root cause is misdocumented. On N-CMAPSS, misdocumenting up to 20% of the sensors changes Top-1 by at most 0.03. On the DDAHU benchmark, misdocumenting one of its sensors lowers Top-1 accuracy from 0.72 to 0.69.

Granting the Ingredients to Baselines. Some baselines already receive the association, however, none receives the condition channels. We therefore run each baseline on Stage 1's condition-matched residuals in place of the raw measurements, keeping its own scoring, graph and aggregation unchanged (Table 2). With adjustment, DoWhy-GCM rises from 0.38 to 0.63 Top-1 on N-CMAPSS, 0.05 to 0.35 on LBNL FDD and 0.17 to 0.28 on OpenRCA. The adjustment claim thus holds for nine different methods. Nevertheless, CULPRIT leads both baseline columns on all three benchmarks.

Design Decisions. We pair each design decision of CULPRIT with one alternative and report Top-1 averaged over the six benchmarks (Table 3, Appendix G.4). Replacing one latent per component by a sensor maximum lowers the overall Top-1 from 0.52 to 0.42, and removing the velocity score lowers it to 0.44. The two velocity forms complement each other, because the latent form alone drops LBNL FDD to 0.43 and the per-sensor form alone drops the Causal Chamber from 0.39 to 0.23 (Appendix G.4). The lower rows change the displacement score Δ, i.e., Stage 3 before structural separation and without the velocity score, at 0.41 average Top-1. Raw standardized signals in place of Stage 1's matched residuals lower it to 0.31. The maximum latent deviation in place of its integral

Table 3: **Each design decision raises the average Top-1 of CULPRIT.**

Design decision	Replaced by	Kept	Replaced	Δ
Changed in CULPRIT				
One latent per component	sensor maximum	0.52	0.42	−0.10
Velocity score	none	0.52	0.44	−0.08
Both velocity forms	latent form only	0.52	0.47	−0.05
Both velocity forms	per-sensor form only	0.52	0.49	−0.03
Changed in the displacement score				
Matched residuals	raw signals	0.41	0.31	−0.10
Integral over the window	maximum	0.41	0.32	−0.09
Structural separation	none	0.44	0.41	−0.03
Structural separation	post-hoc penalty	0.44	0.28	−0.16

over the fault window lowers it to 0.32, because the maximum discards how long a latent stays displaced, which is the signature of a sustained shift. Structural separation lifts Top-1 on LBNL FDD from 0.22 to 0.45, because a damper fault also displaces the cooling coil and fan downstream, and the structural residual removes the part their declared parents explain.

4.4 Efficiency, Robustness, and Parameter Sensitivity

Efficiency. We measure the run time of CULPRIT and the 15 baselines over all 1,261 cases (Figure 2). CULPRIT diagnoses a case in 21ms at the median, about 500 times faster than the graph-search methods CIRCA (10 s) and RCD (16 s).

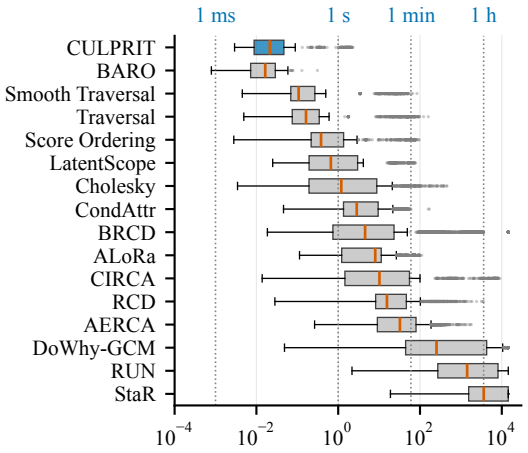


Figure 2: Running time per case (log scale).

Robustness. On RCAEval, we add Gaussian noise up to twice a sensor's healthy standard deviation, and randomly remove sensors (Appendix G.8). CULPRIT keeps at least 0.59 Top-1 at all noise strengths, and 0.52 with half the sensors removed, against 0.30 for BARO. We test the linear mixing model with a tanh per sensor that compresses deviations beyond $1/\gamma$ standard deviations. The saturation strength γ ranges from 0 to 2 (Table 21). CULPRIT stays above BARO and Cholesky throughout, while BARO falls from 0.44 to 0.15. CULPRIT's Top-1 is 0.68 to 0.70 on N-CMAPSS, 0.72 to 0.77 on DDAHU, and 0.39 on the Causal Chamber. Stage 1 standardizes each sensor by its matched spread. A monotone sensor map rescales that spread by the same first-order multiplier, which keeps these curves flat.

Parameter Sensitivity. We vary the condition-match tolerance, the ridge of the structural fit and the matching rule (Appendix G.8). Up to a tolerance of 0.50, Top-1 stays within 0.03 of the reported value on N-CMAPSS, SDAHU and DDAHU. Varying the ridge over two orders of magnitude moves the average over the same three datasets by 0.01. We also replace the ball by seven alternative matching rules, among them k-nearest neighbours with and without bias correction, kernel weighting and regression-based matching. None of them improves Top-1 significantly on any dataset.

5 Conclusion

We presented CULPRIT, a training-free RCA method that treats a fault as a mechanism shift on a component-level latent variable and pinpoints it from shared sensors under varying conditions. Against 15 baselines, it holds the best Top-1 on all six benchmarks and the best average Top-1 of 0.52. The ablations show that CULPRIT needs both operating-condition adjustment and the sensor-to-component association. Theorem 3.2 and Proposition C.12 state when the separated displacement and the velocity score place the root causes first, for displacement and velocity faults respectively. The lesson is that the components a ranking can localize are constrained by the system model an engineer already holds, instead of by environment diversity the engineer must create. We hope CULPRIT serves as a foundation for shift localization across physical and software systems.

AI USE STATEMENT

We used generative AI tools (Claude Fable 5.1 and Claude Opus 5.5 by Anthropic) to help implement the source code and to verify the theoretical results and proofs. We have not used generative AI tools to formulate mathematical claims, write proofs, propose hypotheses, design experiments, or interpret results, and the remaining required-disclosure tasks are not applicable to this work. We also used Claude and ChatGPT to edit the manuscript for grammar, wording, and readability. We have reviewed all AI-assisted work. The authors drafted all theoretical results and proofs before the models checked them, Claude checked the code for correctness and for agreement with the paper, and we verified all references by hand against the original sources. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Alberto Abadie and Guido W Imbens. Large sample properties of matching estimators for average treatment effects. *econometrica*, 74(1):235–267, 2006.

Saeid Allahdadian, Michael Döhler, Carlos Ventura, and Laurent Mevel. Towards robust statistical damage localization via model-based sensitivity clustering. *Mechanical Systems and Signal Processing*, 134:106341, 2019.

Manuel Arias Chao, Chetan Kulkarni, Kai Goebel, and Olga Fink. Aircraft engine run-to-failure dataset under real flight conditions for prognostics and diagnostics. *Data*, 6(1):5, 2021.

Julien Audibert, Pietro Michiardi, Frédéric Guyard, Sébastien Marti, and Maria A Zuluaga. Do deep neural networks contribute to multivariate time series anomaly detection? *Pattern Recognition*, 132:108945, 2022.

Piero Baraldi, Francesco Di Maio, Pietro Turati, and Enrico Zio. Robust signal reconstruction for condition monitoring of industrial components via a modified auto associative kernel regression method. *Mechanical Systems and Signal Processing*, 60:29–44, 2015.

Ali Basiri, Niosha Behnam, Ruud de Rooij, Lorin Hochstein, Luke Kosewski, Justin Reynolds, and Casey Rosenthal. Chaos engineering. *IEEE Software*, 33(3):35–41, 2016. doi: 10.1109/MS.2016.60.

Stephen Bates, Emmanuel Candès, Lihua Lei, Yaniv Romano, and Matteo Sesia. Testing for outliers with conformal p-values. *The Annals of Statistics*, 51(1):149–178, 2023.

Luca Bindini, Lorenzo Perini, Stefano Nistri, Jesse Davis, and Paolo Frasconi. Dealing with uncertainty in contextual anomaly detection. *Transactions on Machine Learning Research*, 2026. ISSN 2835-8856. URL https://openreview.net/forum?id=yLoXQDNwwa.

Patrick Blöbaum, Peter Götz, Kailash Budhathoki, Atalanti A Mastakouri, and Dominik Janzing. Dowhy-gcm: An extension of dowhy for causal inference in graphical causal models. *Journal of Machine Learning Research*, 25(147):1–7, 2024.

Philippe Brouillard, Sébastien Lachapelle, Julia Kaltenborn, Yaniv Gurwicz, Dhanya Sridhar, Alexandre Drouin, Peer Nowack, Jakob Runge, and David Rolnick. Causal representation learning in temporal data via single-parent decoding. *arXiv preprint arXiv:2410.07013*, 2024.

Kailash Budhathoki, Dominik Janzing, Patrick Blöbaum, and Hoiyi Ng. Why did the distribution change? In Arindam Banerjee and Kenji Fukumizu (eds.), *The 24th International Conference on Artificial Intelligence and Statistics, AISTATS 2021, April 13-15, 2021, Virtual Event*, volume 130 of *Proceedings of Machine Learning Research*, pp. 1666–1674. PMLR, 2021. URL http://proceedings.mlr.press/v130/budhathoki21a.html.

Kailash Budhathoki, Lenon Minorics, Patrick Blöbaum, and Dominik Janzing. Causal structure-based root cause analysis of outliers. In *International conference on machine learning*, pp. 2357–2369. PMLR, 2022.

Manuel Arias Chao, Chetan Kulkarni, Kai Goebel, and Olga Fink. Fusing physics-based and deep learning models for prognostics. *Reliability Engineering & System Safety*, 217:107961, 2022.

Jeffryes W Chapman, Thomas M Lavelle, Ryan D May, Jonathan S Litt, and Ten-Huei Guo. Toolbox for the modeling and analysis of thermodynamic systems (t-mats) user's guide. Technical report, 2014.

Pengfei Chen, Yong Qi, Pengfei Zheng, and Di Hou. Causeinfer: Automatic and distributed performance diagnosis with hierarchical causality graph in large distributed systems. In *IEEE INFOCOM 2014-IEEE Conference on Computer Communications*, pp. 1887–1895. IEEE, 2014.

Tianyu Chen, Kevin Bello, Bryon Aragam, and Pradeep Ravikumar. iscan: Identifying causal mechanism shifts among nonlinear additive noise models. *Advances in Neural Information Processing Systems*, 36:44671–44706, 2023.

Tianyu Chen, Kevin Bello, Francesco Locatello, Bryon Aragam, and Pradeep Ravikumar. Identifying general mechanism shifts in linear causal representations. *Advances in Neural Information Processing Systems*, 37:42405–42429, 2024.

Jihoon Chung, Bo Shen, and Zhenyu Kong. A novel sparse bayesian learning and its application to fault diagnosis for multistation assembly systems. *IISE Transactions*, 56(1):84–97, 2024.

Ricardo Dunia and S Joe Qin. Subspace approach to multidimensional fault identification and reconstruction. *AICHE journal*, 44(8):1813–1831, 1998.

Shicheng Fan, Nour Elhendawy, Jianle Sun, Ke Fang, Kun Zhang, Yihang Wang, and Lu Cheng. Mosaic: Module discovery via sparse additive identifiable causal learning for scientific time series. *arXiv preprint arXiv:2605.05524*, 2026.

Amare Desalegn Fentaye, Valentina Zaccaria, and Konstantinos Kyprianidis. Aircraft engine performance monitoring and diagnostics based on deep convolutional neural networks. *Machines*, 9 (12):337, 2021.

Juan L. Gamella, Simon Bing, and Jakob Runge. Sanity checking causal representation learning on a simple real-world system. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 18143–18169. PMLR, 13–19 Jul 2025a. URL https://proceedings.mlr.press/v267/gamella25a.html.

Juan L Gamella, Jonas Peters, and Peter Bühlmann. Causal chambers as a real-world physical testbed for ai methodology. *Nature Machine Intelligence*, 7(1):107–118, 2025b.

Jonathan Goh, Sridhar Adepu, Khurum Nazir Junejo, and Aditya Mathur. A dataset to support research in the design of secure water treatment systems. In *International conference on critical information infrastructures security*, pp. 88–99. Springer, 2016.

Jessica Granderson, Guanjing Lin, Yimin Chen, Armando Casillas, Jin Wen, Zhelun Chen, Piljae Im, Sen Huang, and Jiazhen Ling. A labeled dataset for building hvac systems operating in faulted and fault-free states. *Scientific data*, 10(1):342, 2023.

Xiao Han, Saima Absar, Lu Zhang, and Shuhan Yuan. Root cause analysis of anomalies in multivariate time series through granger causal discovery. In *International Conference on Learning Representations*, volume 2025, pp. 45137–45153, 2025.

Eric S Hendricks and Justin S Gray. pycycle: A tool for efficient optimization of gas turbine engine cycles. *Aerospace*, 6(8):87, 2019.

Norbert Herrndorf. A functional central limit theorem for weakly dependent sequences of random variables. *The Annals of Probability*, 12(1):141–153, 1984.

Rohan Hore and Aaditya Ramdas. Distribution-free root cause analysis. *arXiv preprint arXiv:2605.21627*, 2026.

Haiyu Huang, Man Tik Ng, Jiewei Lyu, Yujie Huang, Guangba Yu, Yilun Wang, and Michael R Lyu. Star: Stateful dynamic-graph root cause analysis through memory-enhanced causality discovery. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pp. 1828–1839, 2026.

Azam Ikram, Sarthak Chakraborty, Subrata Mitra, Shiv Saini, Saurabh Bagchi, and Murat Kocaoglu. Root cause analysis of failures in microservices through causal discovery. *Advances in Neural Information Processing Systems*, 35:31158–31170, 2022.

Azam Ikram, Kenneth Lee, Shubham Agarwal, Shiv Kumar Saini, Saurabh Bagchi, and Murat Kocaoglu. Root cause analysis of failures from partial causal structures. In *The 41st Conference on Uncertainty in Artificial Intelligence*, 2025.

Jun Inoue, Yoriyuki Yamagata, Yuqi Chen, Christopher M Poskitt, and Jun Sun. Anomaly detection for a water treatment system using unsupervised machine learning. In *2017 IEEE international conference on data mining workshops (ICDMW)*, pp. 1058–1065. IEEE, 2017.

S Joe Qin. Statistical process monitoring: basics and beyond. *Journal of Chemometrics: A Journal of the Chemometrics Society*, 17(8-9):480–502, 2003.

Kenneth Lee, Zihan Zhou, and Murat Kocaoglu. Root cause analysis of failures in microservices via bayesian root cause discovery. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=EOSV5tlpqg.

Jinzhou Li, Benjamin B Chu, Ines F Scheller, Julien Gagneur, and Marloes H Maathuis. Root cause discovery via permutations and cholesky decomposition. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 88(3):779–798, 2026.

Mingjie Li, Zeyan Li, Kanglin Yin, Xiaohui Nie, Wenchi Zhang, Kaixin Sui, and Dan Pei. Causal inference-based root cause analysis for online service systems with intervention recognition. In *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining*, pp. 3230–3240, 2022.

Yi-Guang Li. Performance-analysis-based gas turbine diagnostics: A review. *Proceedings of the Institution of Mechanical Engineers, Part A: Journal of Power and Energy*, 216(5):363–377, 2002.

Zijian Li, Yifan Shen, Kaitao Zheng, Ruichu Cai, Xiangchen Song, Mingming Gong, Guangyi Chen, and Kun Zhang. On the identification of temporal causal representation with instantaneous dependence. In *International Conference on Learning Representations*, volume 2025, pp. 35203–35242, 2025.

Wendong Liang, Armin Kekić, Julius von Kügelgen, Simon Buchholz, Michel Besserve, Luigi Gresele, and Bernhard Schölkopf. Causal component analysis. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Cheng-Ming Lin, Ching Chang, Wei-Yao Wang, Kuang-Da Wang, and Wen-Chih Peng. Root cause analysis in microservice using neural granger causal discovery. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 206–213, 2024.

Xiaoyu Lin, Nicholas Tagliapietra, Kehan Li, Lavdim Halilaj, and Juergen Luettin. Stablerca: Robust graph-agnostic mechanism-level root cause analysis. *arXiv preprint arXiv:2606.05636*, 2026.

John F MacGregor and Theodora Kourti. Statistical process control of multivariate processes. *Control engineering practice*, 3(3):403–414, 1995.

Frank McSherry and Marc Najork. Computing information retrieval performance measures efficiently in the presence of tied scores. In *European conference on information retrieval*, pp. 414–421. Springer, 2008.

Shashank Mishra, Karan Patil, Cedric Schockaert, Didier Stricker, and Jason Rambach. Conditional attribution for root cause analysis in time-series anomaly detection. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 165–181. Springer, 2026.

Hiroshi Morioka and Aapo Hyvärinen. Causal representation learning made identifiable by grouping of observational variables. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024.

Whitney K Newey and Kenneth D West. A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica*, 55(3):703–708, 1987.

William Roy Orchard, Nastaran Okati, Sergio Garrido Mejia, Patrick Blöbaum, and Dominik Janzing. Root cause analysis of outliers with missing structural knowledge. In *Advances in Neural Information Processing Systems*, pp. 161634–161671, 2025. doi: 10.52202/085713-5399.

Ravi Kumar Pandit and David Infield. Scada-based wind turbine anomaly detection using gaussian process models for wind turbine condition monitoring purposes. *IET Renewable Power Generation*, 12(11):1249–1255, 2018.

Juan Luis Pérez-Ruiz, Yu Tang, and Igor Loboda. Aircraft engine gas-path monitoring and diagnostics framework based on a hybrid fault recognition approach. *Aerospace*, 8(8):232, 2021.

Ronan Perry, Julius Von Kügelgen, and Bernhard Schölkopf. Causal discovery in heterogeneous environments under the sparse mechanism shift hypothesis. *Advances in Neural Information Processing Systems*, 35:10904–10917, 2022.

Luan Pham, Huong Ha, and Hongyu Zhang. Baro: Robust root cause analysis for microservices via multivariate bayesian online change point detection. *Proceedings of the ACM on Software Engineering*, 1(FSE):2214–2237, 2024a.

Luan Pham, Huong Ha, and Hongyu Zhang. Root cause analysis for microservice system based on causal inference: How far are we? In *Proceedings of the 39th IEEE/ACM International Conference on Automated Software Engineering*, pp. 706–715, 2024b.

Luan Pham, Hongyu Zhang, Huong Ha, Flora Salim, and Xiuzhen Zhang. Rcaeval: A benchmark for root cause analysis of microservice systems with telemetry data. In *Companion Proceedings of the ACM on Web Conference 2025*, pp. 777–780, 2025.

Donald B Rubin. Matching to remove bias in observational studies. *Biometrics*, pp. 159–183, 1973.

M. Saquib Sarfraz, Mei-Yen Chen, Lukas Layer, Kunyu Peng, and Marios Koulakis. Position: Quo vadis, unsupervised time series anomaly detection? In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=W7Vqx1Jvc2.

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.

Huasong Shan, Yuan Chen, Haifeng Liu, Yunpeng Zhang, Xiao Xiao, Xiaofeng He, Min Li, and Wei Ding. ϵ-diagnosis: Unsupervised and real-time diagnosis of small-window long-tail latency in large-scale microservice platforms. In *The World Wide Web Conference*, pp. 3215–3222, 2019.

Charalampos Shimillas, Kleanthis Malialis, Konstantinos Fokianos, and Marios Polycarpou. Low rank transformer for multivariate time series anomaly detection and localization. In *International Conference on Learning Representations*, volume 2026, pp. 59972–60005, 2026.

Donald L. Simon. Propulsion diagnostic method evaluation strategy (ProDiMES) user's guide. Technical Report NASA/TM-2010-215840, NASA Glenn Research Center, 2010.

Donald L Simon, Sébastien Borguet, Olivier Léonard, and Xiaodong Zhang. Aircraft engine gas path diagnostic methods: Public benchmarking results. Technical Report 4, 2014.

Donald L Simon, Randy Thomas, and Kyle M Dunlap. Considerations for the extension of gas path analysis to electrified aircraft propulsion systems. *Journal of Engineering for Gas Turbines and Power*, 144(3):031004, 2022.

Xiangchen Song, Weiran Yao, Yewen Fan, Xinshuai Dong, Guangyi Chen, Juan Carlos Niebles, Eric Xing, and Kun Zhang. Temporally disentangled representation learning under unknown nonstationarity. *Advances in Neural Information Processing Systems*, 36:8092–8113, 2023.

Xiuyao Song, Mingxi Wu, Christopher Jermaine, and Sanjay Ranka. Conditional anomaly detection. *IEEE Transactions on Knowledge and Data Engineering*, 19(5):631–645, 2007.

Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, prediction, and search*. The MIT press, 2001.

Chandler Squires, Anna Seigal, Salil S Bhate, and Caroline Uhler. Linear causal disentanglement via interventions. In *International conference on machine learning*, pp. 32540–32560. PMLR, 2023.

Henrik Sebastian Steude, Alexander Diedrich, Ingo Pill, Lukas Moddemann, Daniel Vranješ, and Oliver Niggemann. Data-driven diagnosis for large cyber-physical-systems with minimal prior information. pp. 325–332, 2026.

Nils Sturma, Chandler Squires, Mathias Drton, and Caroline Uhler. Unpaired multi-domain causal representation learning. volume 36, pp. 34465–34492, 2023.

Julius von Kügelgen, Michel Besserve, Liang Wendong, Luigi Gresele, Armin Kekić, Elias Bareinboim, David Blei, and Bernhard Schölkopf. Nonparametric identifiability of causal representations from unknown interventions. *Advances in Neural Information Processing Systems*, 36: 48603–48638, 2023.

Tianyi Wang, Jianbo Yu, David Siegel, and Jay Lee. A similarity-based prognostics approach for remaining useful life estimation of engineered systems. In *Proceedings of the IEEE International Conference on Prognostics and Health Management (PHM)*, pp. 1–6. IEEE, 2008.

Stephan W Wegerich. Similarity based modeling of time synchronous averaged vibration signals for machinery health monitoring. In *2004 IEEE Aerospace Conference Proceedings (IEEE Cat. No. 04TH8720)*, volume 6, pp. 3654–3662. IEEE, 2004.

Zhe Xie, Shenglin Zhang, Yitong Geng, Yao Zhang, Minghua Ma, Xiaohui Nie, Zhenhe Yao, Longlong Xu, Yongqian Sun, Wentao Li, et al. Microservice root cause analysis with limited observability through intervention recognition in the latent space. In *Proceedings of the 30th ACM SIGKDD conference on knowledge discovery and data mining*, pp. 6049–6060, 2024.

Junjielong Xu, Qinan Zhang, Zhiqing Zhong, Shilin He, Chaoyun Zhang, Qingwei Lin, Dan Pei, Pinjia He, Dongmei Zhang, and Qi Zhang. OpenRCA: Can large language models locate the root cause of software failures? In *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025. URL https://openreview.net/forum?id=M4qNIzQYpd.

Cansu Yalim, Resit Unal, and Holly AH Handley. A three-stage causal root-cause diagnostic protocol for nonstationary industrial time series data. *International Journal of Intelligent Systems*, 2026(1):3436744, 2026.

Dingling Yao, Danru Xu, Sébastien Lachapelle, Sara Magliacane, Perouz Taslakian, Georg Martius, Julius von Kügelgen, and Francesco Locatello. Multi-view causal representation learning with partial observability. In *International Conference on Learning Representations*, volume 2024, pp. 34817–34848, 2024.

Weiran Yao, Yuewen Sun, Alex Ho, Changyin Sun, and Kun Zhang. Learning temporally causal latent processes from general temporal data. In *International Conference on Learning Representations (ICLR)*, 2022.

Yi Yu, Tengyao Wang, and Richard J Samworth. A useful variant of the davis–kahan theorem for statisticians. *Biometrika*, 102(2):315–323, 2015.

Yujia Zheng, Ignavier Ng, and Kun Zhang. On the identifiability of nonlinear ica: Sparsity and beyond. *Advances in neural information processing systems*, 35:16411–16422, 2022.

Yujia Zheng, Biwei Huang, Wei Chen, Joseph Ramsey, Mingming Gong, Ruichu Cai, Shohei Shimizu, Peter Spirtes, and Kun Zhang. Causal-learn: Causal discovery in python. *Journal of Machine Learning Research*, 25(60):1–8, 2024.

A Related Work

This appendix positions CULPRIT within the three literatures the main text draws on. CULPRIT competes with root cause analysis methods, takes its identifiability language from causal representation learning, and inherits its matching step from prognostics.

Root cause analysis for operational systems. The RCA methods this appendix reviews divide into seven families. Localization on a given or pre-learned graph fits a model on healthy data and scores the post-onset deviation against the graph. The score comes from outlier attribution (DoWhy-GCM (Budhathoki et al., 2022; Blöbaum et al., 2024)), ancestor traversal (Chen et al., 2014; Orchard et al., 2025), or intervention recognition on a supplied call graph (Li et al., 2022). Causal graph search recovers the structure from the case itself before scoring it (Ikram et al., 2022; Lin et al., 2024). A related line asks what remains identifiable when the structure is only partly known (Ikram et al., 2025). BRCD learns a completed partially directed acyclic graph (CPDAG) on the healthy prefix. It places a posterior over the sensors a failure indicator acts on, marginalized over the orientations the CPDAG leaves open. It comes with generic identifiability and posterior consistency under interventional faithfulness and causal sufficiency (Lee et al., 2026). Deviation ranking scores each component from its own departure from a healthy reference, by a robust univariate statistic (Pham et al., 2024a) or by a similarity test on small windows (Shan et al., 2019). Deep learning methods train a model on healthy data and read the root cause off it. AERCA and StaR rank the exogenous z-scores of an autoencoder or a graph memory (Han et al., 2025; Huang et al., 2026). ALoRa scores a sensor by the reconstruction errors of all sensors, weighted by its contribution to each reconstruction inside a low-rank transformer (Shimillas et al., 2026). CondAttr credits a sensor with the drop in a trained detector's score when its trajectory is swapped for those of the nearest healthy windows. It retrieves these windows in a learned embedding of the other sensors (Mishra et al., 2026). A single-sample line diagnoses the first post-onset sample alone, and neither of its two methods needs the graph. One whitens the sample against the healthy covariance under a linear structural causal model (Li et al., 2026). The other orders marginal anomaly scores with a guarantee under a bounded in-degree (Orchard et al., 2025). Two recent methods relax the observability the rest assume. LatentScope scores components as latents above the sensor layer (Xie et al., 2024). The other, a statistical line, builds finite-sample confidence sets for the earliest-changing stream from conformal p-values (Hore & Ramdas, 2026) (Appendix G.6). Surveys and benchmarks of the microservice setting report that no family dominates and that evaluation protocol drives much of the spread between published numbers (Pham et al., 2024b; 2025). On the same setting, the best language-model agent solves 11.34% of the cases (Xu et al., 2025). Except for LatentScope, these families as published assign each sensor to one component, and none of them takes the recorded operating condition as an input of its own and compares the fault window with healthy data at that condition. LatentScope links a component to several sensors and ranks it by the largest residual among them after a pairwise-regression correction (Xie et al., 2024). CULPRIT instead estimates a latent from the joint movement of the component's sensors and separates it from its parents' latents before scoring, which distinguishes the component's own mechanism shift from downstream effects. Appendix G.7 shows that LatentScope's maximum over sensors inherits the sensor hierarchy of the instrumentation. The regression-based methods score each sensor given its graph parents. This reaches the demand only where the graph places a demand signal among them, as CIRCA's graph does with each service's traffic (Li et al., 2022). StaR fits its causal coefficients given a memory that the monitored sensors update (Huang et al., 2026). CondAttr retrieves healthy windows that resemble the fault window on the other monitored sensors. A time-varying dynamic Bayesian network learns one graph per regime and places the regime boundaries at change points in the monitored sensors (Yalim et al., 2026). However, those sensors record the system's response, which the fault itself moves, while the condition channels record its demand. CULPRIT lifts both restrictions, with the sensor-to-component association for shared sensors and pointwise matching for the operating condition.

Causal representation learning and mechanism shifts. Temporal causal representation learning recovers latent processes from observed sequences under assumptions on the transition dynamics, nonstationarity, or the sparsity of the latent structure (Yao et al., 2022; Song et al., 2023; Li et al., 2025; Brouillard et al., 2024; Morioka & Hyvärinen, 2024). MOSAIC applies this line to named process sensors (Fan et al., 2026). A sparse temporal variational autoencoder with an additive decoder recovers a sparse loading over the sensors for each latent factor, and MOSAIC ranks the factors by their association with the regime to return a module partition. On the Tennessee Eastman

process, its regime-associated factor settles on the product-composition cluster that the authors describe as the downstream response to all four faults. That response is the propagation that Stage 3 removes, and Stages 2 and 3 are closed form, whereas MOSAIC trains a variational autoencoder. A second line establishes what interventions buy, from nonparametric identifiability under unknown interventions (von Kügelgen et al., 2023) to causal component analysis (Liang et al., 2023) and partial observability (Yao et al., 2024). Unpaired multi-domain causal representation learning observes linear latents through overlapping groups of variables and identifies the sparsity pattern and the latent graph when each latent has two partial pure children, i.e., observed variables whose only latent parent is that latent (Sturma et al., 2023). In that vocabulary, the association A is the bipartite sparsity pattern and an local sensor is a partial pure child. This paper takes the framing of that programme (Schölkopf et al., 2021). The system model supplies the association these methods set out to recover, and with it one local sensor suffices to resolve a component (Corollary C.8). Those works place the mechanism on lagged transitions, whereas this paper reads it on the edges in $\mathcal{E}$ at the same time step. It keeps the dynamics in the exogenous process, the form in which the shifted-node estimand is stated (Chen et al., 2024). The estimand it does share is the mechanism shift, studied as a comparison between environments (Chen et al., 2023; 2024) and under the sparse mechanism shift hypothesis for discovery (Perry et al., 2022). Chen et al. (2024) identify the latents whose coefficients or noise scale shift between two environments of a linear model. Their result assumes a noiseless mixing and non-Gaussian exogenous noise, and it holds up to the names of the latents. StableRCA localizes the shift for root cause analysis without a global graph, as a change in a variable's conditional distribution given its estimated local Markov boundary, and recovers the shifted variable with probability converging exponentially in sample size (Lin et al., 2026). Each of its candidates is an observed variable. CULPRIT names the shifted component through the association and allows shared sensors and noisy mixing under varying conditions.

Operating-point matching in prognostics. The condition-monitoring literature has long compared a system against healthy behaviour at the same operating point. Some methods retrieve healthy exemplars close to the current operating point (Wegerich, 2004; Baraldi et al., 2015). Others match a degrading system against run-to-failure trajectories partitioned by regime (Wang et al., 2008) or adjust the anomaly score for environmental attributes (Song et al., 2007; Pandit & Infield, 2018). Contextual anomaly detection now makes this adjustment with heteroscedastic Gaussian process regression (Bindini et al., 2026). These methods answer whether a system is degrading and leave the component to the engineer. Gas-path diagnostics carries the deviation to a component through an engine-specific thermodynamic model (Li, 2002) or a library of simulated fault signatures (Fentaye et al., 2021; Pérez-Ruiz et al., 2021; Simon et al., 2014). The isolation is therefore per engine type and per fault, and the sensor-to-component association replaces it. Stage 1 is the matching estimator of observational studies, which compares a treated unit against controls balanced on the covariates that drive the outcome (Rubin, 1973; Abadie & Imbens, 2006). Here the operating condition is the covariate and the fault window the treated unit.

B Proofs for Stage 1

We prove Proposition 3.1 and Corollary B.1 and treat direct condition effects on sensors.

Fix a sensor s and let $h_s(u) = \mathbb{E}\big[x_{t,s} \mid \tilde{u}_t = u,\, e = 0\big]$ be its healthy response to the operating condition, with healthy noise $\varepsilon_{t',s} = x_{t',s} - h_s(\tilde{u}_{t'})$ at each healthy sample t'. The healthy response is quasi-static when $\mathbb{E}[x_{t',s} \mid \tilde{u}_{1:T}, e = 0] = h_s(\tilde{u}_{t'})$. At a fault sample t, the residual numerator of Equation 4 subtracts a reference mean from $x_{t,s}$. The matched reference uses $\hat{\mu}_s(t)$, and the whole-window reference uses the healthy mean $\bar{\mu}_s$ over $\mathcal{T}_0$. Each numerator splits into a condition-induced term, $h_s(\tilde{u}_t)$ minus the reference mean, and the remainder $x_{t,s} - h_s(\tilde{u}_t)$. The remainder holds the mechanism-induced term and the noise of the fault sample and does not depend on the reference. Let $\mathrm{osc}_s = \sup_{u,u' \in \mathcal{U}_0 \cup \mathcal{U}_1} |h_s(u) - h_s(u')|$ be the swing of the healthy response across both windows.

Proposition 3.1 (Matching error), full statement. *Let the healthy response h_s be quasi-static and L_s-Lipschitz with respect to $\|\cdot\|_\infty$, and let the healthy noise have, given $\tilde{u}_{1:T}$, variance at most σ_s^2 and absolute autocorrelations summing to at most χ_s. Then each fault sample t satisfies (i) $\big|\hat{\mu}_s(t) - h_s(\tilde{u}_t)\big| \le L_s\tau + \sigma_s\sqrt{(1+2\chi_s)/(\delta N)}$ with probability at least $1-\delta$ whenever $N = |\mathcal{M}_\tau(t)| \ge n_{\min}$, and (ii) $\big|\bar{\mu}_s - h_s(\tilde{u}_t)\big| \le \mathrm{osc}_s + \big||\mathcal{T}_0|^{-1}\sum_{t' \in \mathcal{T}_0} \varepsilon_{t',s}\big|$.*

By (i), the condition-induced term of the matched reference is at most $L_s\tau$ plus a sampling term that shrinks as the matched set grows. By (ii), the condition-induced term of the whole-window reference is at most osc_s plus its own sampling term. The two references share the mechanism-induced term.

Proof of Proposition 3.1. Condition on the condition channels $\tilde{u}_{1:T}$, which fixes the matched set, and decompose each healthy measurement as $x_{t',s} = h_s(\tilde{u}_{t'}) + \varepsilon_{t',s}$, with $\mathbb{E}[\varepsilon_{t',s} \mid \tilde{u}_{1:T}] = 0$ by the quasi-static premise. The matched mean splits into a bias term and a noise term,

$$\hat{\mu}_s(t) - h_s(\tilde{u}_t) = \frac{1}{N} \sum_{t' \in \mathcal{M}_\tau(t)} \big(h_s(\tilde{u}_{t'}) - h_s(\tilde{u}_t)\big) \; + \; \frac{1}{N} \sum_{t' \in \mathcal{M}_\tau(t)} \varepsilon_{t',s}.$$

Each $t' \in \mathcal{M}_\tau(t)$ obeys $\|\tilde{u}_{t'} - \tilde{u}_t\|_\infty \le \tau$ by Equation 3, hence the Lipschitz property bounds each summand of the first term by $L_s\tau$, and the bound survives the average. The second term has mean zero and variance $N^{-2} \sum_{t',t''} \mathrm{Cov}(\varepsilon_{t',s}, \varepsilon_{t'',s} \mid \tilde{u}_{1:T}) \le \sigma_s^2(1 + 2\chi_s)/N$, since for a fixed t' each time lag occurs at most twice among the distinct t'' of the matched set. Chebyshev's inequality, applied given $\tilde{u}_{1:T}$, gives the first claim.

For the second claim, write the fault sample as $x_{t,s} = h_{s,1}(\tilde{u}_t) + \varepsilon_{t,s}$ under the fault environment and expand the numerator of Equation 4,

$$x_{t,s} - \hat{\mu}_s(t) = \underbrace{\big(h_{s,1}(\tilde{u}_t) - h_s(\tilde{u}_t)\big)}_{\text{mechanism-induced}} \; + \; \underbrace{\big(h_s(\tilde{u}_t) - \hat{\mu}_s(t)\big)}_{\text{condition-induced}} \; + \varepsilon_{t,s}.$$

The first claim bounds the condition-induced term by $L_s\tau$ up to the sampling term. Under a whole-window reference over the healthy support $\mathcal{U}_0$ the matched mean becomes the healthy average $\bar{\mu}_s = |\mathcal{T}_0|^{-1} \sum_{t' \in \mathcal{T}_0} x_{t',s}$, with $\bar{\sigma}_s$ the standard deviation over it. The condition-induced term of this reference is $h_s(\tilde{u}_t) - |\mathcal{T}_0|^{-1} \sum_{t'} h_s(\tilde{u}_{t'})$ up to its own sampling term. That difference is an average of increments of h_s between points of $\mathcal{U}_0 \cup \mathcal{U}_1$, hence bounded by osc_s. It attains that order whenever the fault window sits at one extreme of the conditions both windows cover, the ordinary situation for a flight or a load cycle. Both references share the mechanism-induced term. □

Corollary B.1 (Whole-window reference as the wide-ball limit)**.** *Once τ reaches the Chebyshev diameter of $\mathcal{U}_0 \cup \mathcal{U}_1$, each matched set equals $\mathcal{T}_0$, the matched moments become the whole-window healthy moments, and d_s equals the unadjusted maximum z-score.*

Proof of Corollary B.1. Let $\tau \ge \mathrm{diam}_\infty(\mathcal{U}_0 \cup \mathcal{U}_1)$. Then $\|\tilde{u}_{t'} - \tilde{u}_t\|_\infty \le \tau$ holds for all $t' \in \mathcal{T}_0$ and $t \in \mathcal{T}_1$, hence Equation 3 gives $\mathcal{M}_\tau(t) = \mathcal{T}_0$. The matched moments $\hat{\mu}_s(t)$ and $\hat{\sigma}_s(t)$ are the whole-window healthy mean and standard deviation and no longer depend on t. Equation 4 then becomes the maximum over the fault window of $|x_{t,s} - \bar{\mu}_s|/\bar{\sigma}_s$. This maximum is the unadjusted maximum z-score the ablation of Section 4.3 computes. Both coverage numbers of Equation 15 therefore equal one, and the statistic is tolerance-invariant beyond this radius. □

Direct condition effects on sensors. The healthy response $h_s(u)$ is the healthy mean of sensor s at condition u. Proposition 3.1 therefore holds when u_t acts through the latents of Equation 1, directly through g_s in Equation 2, or through both. On N-CMAPSS the flight condition alone sets the fan-inlet pressure P2. The standard-atmosphere ram pressure at the recorded altitude and Mach number explains at least 99.99% of its healthy variance on each of the 99 engines (five-fold cross-validated). The fault-window mean of P2 stays within 0.4 healthy standard deviations of that fit, and the system model assigns it to no component. An inlet or outdoor temperature is itself a condition channel on both condition-labelled plants, the fan-inlet temperature on N-CMAPSS and the outdoor air temperature on LBNL FDD. A condition-dependent gain of g_s leaves no condition-induced term and scales the mechanism-induced term with u_t. The division by $\hat{\sigma}_s(t)$ cancels that gain where the healthy spread at u_t carries the same gain.

Recording an ambient quantity that acts on the sensors as a condition channel moves it into u_t. Matching then removes its effect wherever the healthy window covers its fault-window values. DDAHU treats the outdoor air humidity this way.

C THEORY FOR STAGES 2 AND 3

Definition 2.1 defines the root cause set on latent variables, and Stage 3 ranks the latents Stage 2 computes. This appendix states when the ranking recovers the root cause set, in four steps. First, Assumptions C.1 to C.3 fix the model, and Definition C.4 names the structural classes an engineer can read off the association. Second, Lemma C.5 and Propositions C.6 and C.7 describe the structural residuals under a fault. Third, Theorem 3.2 and Corollary C.8 state when the ranking places $S^\star$ first, and Proposition C.9 treats an unresolved source component. Fourth, Proposition 3.3 extends the ranking to the combined score, and Proposition C.12 states when the velocity score ranks the component of a velocity fault first. The proofs close the appendix in dependency order. CULPRIT takes the sparsity pattern of the mixing as input instead of estimating it. Causal representation learning methods that estimate the mixing from data fail to recover the causal factors of a simple physical system (Gamella et al., 2025a). The column labels of A fix which latent carries which name. Condition channels enter in two places. The first is the estimand, since Definition 2.1 compares the mechanisms at the conditions both windows visit. Stage 1 realizes it by matching. The second is the single W of Assumption C.1 across operating conditions.

C.1 ASSUMPTIONS AND STRUCTURAL CLASSES

Throughout this appendix z_t is the deviation of the latent state from its healthy conditional mean at the operating point u_t, the reading of Equation 5. The statements concern the samples at which all residuals are scorable. They are population statements. The fits w_c, s_c, $\hat{B}$ and $\tilde{s}_c$ are held fixed, with the scales s_c and $\tilde{s}_c$ at the healthy standard deviations they estimate. The law of a quantity in either window is its law on a fresh window under these fixed transforms.

Assumption C.1 (Linear mixing on the association A)**.** In both windows the residuals of Equation 4 satisfy $r_t = Wz_t + \nu_t$, with $W_{s,c} \neq 0$ exactly when $A_{s,c} = 1$ and one mixing matrix W across $\mathcal{U}_0 \cup \mathcal{U}_1$. The residual noise ν_t has mean zero and is independent of the latent process.

The first-order expansion of Section 3.2 gives one Jacobian of g_s per operating point, and Stage 1 divides it by the healthy spread at that point. One W asks that this ratio stay put across the visited operating points. The nonzero pattern is what the drawing records, since a sensor assigned to a component sits where that component's state reaches it. The residual noise collects the standardized sensor noise, the linearization remainder and the sampling part of the matching error of Proposition 3.1. Independence is natural for the sensor noise.

Assumption C.2 (Linear mechanisms on the edges $\mathcal{E}$)**.** In both windows the latent state follows $z_t = Bz_t + \zeta_t$, where B has spectral radius below one and $B_{c\ell} = \beta_{c\ell}$ is nonzero only for $\ell \in \mathrm{Pa}(c)$. The coefficients do not vary across $\mathcal{U}_0 \cup \mathcal{U}_1$, and the exogenous processes $\zeta^{(c)}$ are mutually independent with mean zero in the healthy window. The edge set $\mathcal{E}$ is the true edge set.

Assumption C.2 is the linear case of Equation 1, written on the deviations of Equation 5, where Stage 1 has removed the healthy effect of the operating condition. On an acyclic $\mathcal{E}$ the spectral radius of B is zero, since B is nilpotent in a topological order. An air handler closes a loop, with air returning from the zones through the return fan and the mixing damper to the coil. The LBNL FDD schemas record that loop. A spectral radius below one makes $I - B$ invertible, and $z_t = (I-B)^{-1}\zeta_t$ is then the limit of the recurrence $z \mapsto Bz + \zeta_t$, the state the loop settles to.

Assumption C.3 (Displacement faults)**.** Write $\zeta_{t,0}$ and $\zeta_{t,1}$ for the exogenous process in the healthy and the fault environment. Then $\zeta_{t,1} \overset{d}{=} \zeta_{t,0} + \vartheta_t$ jointly in time, where $\vartheta_{t,c}$ is nonzero for some t exactly when $c \in S^\star$, and the coefficients B do not change.

A shift that grows over the window, as wear does after a fault onset (Arias Chao et al., 2021), is a time-varying ϑ_t. A shift that inflates the variance of $\zeta^{(c)}$ also places $S^\star$ first for equally noisy components under Gaussian laws and without smearing (after the proof of Theorem 3.2). A velocity fault, e.g., a damper stuck at its position, is the target of the velocity score of Equation 8 (Proposition C.12).

What the projection returns. Under Assumption C.1, the latent of Equation 6 expands as

$$s_c\, \hat{z}_t^{(c)} = (w_c^\top a_c)\, z_t^{(c)} + \sum_{c' \neq c} \left(w_c^\top W[\mathcal{S}_c, c']\right) z_t^{(c')} + w_c^\top \nu_t[\mathcal{S}_c], \tag{12}$$

where the coefficient of c' is zero whenever c' shares no sensor with c. In matrix form $\hat{z}_t = Mz_t + D^{-1}\hat{W}^\top\nu_t$ with $M = D^{-1}\hat{W}^\top W = D_\alpha + H$. Here $D_\alpha = \mathrm{diag}(\alpha_c)$ holds the gains $\alpha_c = w_c^\top a_c / s_c$. The matrix H, zero on its diagonal, holds the *smearing* $H_{cc'} = w_c^\top W[\mathcal{S}_c, c']/s_c$ of component c' into the latent of c. We take $w_c^\top a_c \neq 0$ for each c, i.e., each latent reads its own component, hence D_α is invertible. We also take the top eigenvalue of each Σ_c to be positive and simple, hence w_c is unique up to the sign of Section 3.2 and $s_c > 0$. Proposition C.10 analyzes the direction for a component whose sensors are all local.

Structural classes. Whether the sensors of c can tell c apart from the other components sharing them is a property of the association, which we name at two levels.

Definition C.4 (Resolved and anchored component)**.** Component c is *resolved* by the association A when

$$\bigcap_{s \in \mathcal{S}_c} A_s = \{c\}, \tag{13}$$

i.e., when no other component is assigned all of c's sensors. Component c is *anchored* when it has at least one local sensor, $A_s = \{c\}$ for some $s \in \mathcal{S}_c$. An anchored component is resolved.

Resolution matches the structural sparsity condition of Zheng et al. (2022) read for one component. A projection on the local sensors of an anchored component reads no other component (Corollary C.8). Proposition C.9 flags as unresolved a source component that is its only child's only parent and whose sensors all belong to that child. Both conditions are graphical, i.e., they depend only on the association and the edges, hence they are known before a diagnosis. All 10 components are resolved and anchored on SDAHU. On DDAHU 24 of 26 are resolved, each water pump sharing all sensors with its coil, and 20 are anchored. On N-CMAPSS 4 of 5 are resolved and 3 are anchored. The Causal Chamber is the stress case, with 2 of 21 components resolved, as ten of its eleven sensors respond to several actuators. There the projection lifts Top-1 from 0.27 to 0.39 over the sensor maximum (Table 13).

C.2 Separation and ranking

Write $\hat{B}$ for the coefficients $\hat{\beta}_{cc'}$ Equation 7 uses, in latent units with row c supported on $\mathrm{Pa}(c)$. Write also $\tilde{D} = \mathrm{diag}(\tilde{s}_c)$ with $\tilde{s}_c = 1$ for a component with no parent, and $B^\star = D_\alpha B D_\alpha^{-1}$ for the healthy coefficients in latent units. The structural residuals are then $\tilde{z}_t = \tilde{D}^{-1}(I - \hat{B})\hat{z}_t$, with the rows of a parentless component reading its latent.

Lemma C.5 (The separated residual under a fault)**.** *Under Assumptions C.1 to C.3, with w_c, s_c, $\hat{B}$ and $\tilde{s}_c$ fitted on the healthy window, the structural residuals in the fault window are*

$$\tilde{z}_{t,1} = \epsilon_t + \xi_t, \qquad \xi_t = \tilde{D}^{-1}K\vartheta_t, \qquad K = (I - \hat{B})\,M\,(I - B)^{-1} = D_\alpha + \tilde{H},$$
$$\tilde{H} = \big[(I - \hat{B})H - (\hat{B} - B^\star)D_\alpha\big](I - B)^{-1},$$

where the null part ϵ_t has mean zero. If ν has one law in both windows, ϵ_t has the law of $\tilde{z}_t$ on a fresh healthy window.

The null part is the map that gives $\tilde{z}_t$ from (ζ_t, ν_t) in the healthy window, applied to $(\zeta_{t,1} - \vartheta_t, \nu_{t,1})$. The shift ξ_t is the root causes' own terms $\alpha_c\vartheta_{t,c}/\tilde{s}_c$ plus the *smearing* $\tilde{D}^{-1}\tilde{H}\vartheta_t$, which collects the latents of the components sharing each component's sensors and the error of the fitted coefficients.

Proposition C.6 (Structural separation)**.** *Under Assumptions C.1 to C.3, if $H = 0$ and $\hat{B} = B^\star$, then $\tilde{H} = 0$. The structural residual of each component outside $S^\star$ is its null part whatever its ancestors do, and that of each $c \in S^\star$ is translated by $\alpha_c\vartheta_{t,c}/\tilde{s}_c$.*

A marginal deviation score alone cannot take the place of the structural residual. Squared z-scores do not identify the root cause of a linear structural equation model in general (Li et al., 2026).

Proposition C.7 gives the coefficients Equation 7 fits when the parents' latents carry noise. For components whose sensors are all local, write Pa for $\mathrm{Pa}(c)$ in subscripts and superscripts, and R_{Pa} for the healthy correlation matrix of the latents of the parents. Its part carried by the parents' states is $R^z_{\mathrm{Pa}} = D_{\alpha,\mathrm{Pa}}\mathrm{Cov}_0(z^{\mathrm{Pa}})D_{\alpha,\mathrm{Pa}}$, and its part carried by noise is $R^\nu_{\mathrm{Pa}} = R_{\mathrm{Pa}} - R^z_{\mathrm{Pa}}$. Write also $b_c^\star$ for row c of $B^\star$. The population ridge fit of Equation 7 is its ridge solution under the healthy covariances.

Proposition C.7 (Fitted coefficients)**.** *Under Assumptions C.1 and C.2, let c lie on no loop of $\mathcal{E}$, let $H = 0$ on the rows of c and $\mathrm{Pa}(c)$, and let the noise on the sensors of c be uncorrelated with that on the sensors of $\mathrm{Pa}(c)$. (i) The population ridge fit with penalty λ is $\hat{b}_c = (R_{\mathrm{Pa}} + \lambda I)^{-1} R^z_{\mathrm{Pa}}\, b^\star_c$. (ii) A shift μ_{Pa} of z^{Pa} leaves $b^{\star\top}_c (R^\nu_{\mathrm{Pa}} + \lambda I)(R_{\mathrm{Pa}} + \lambda I)^{-1} D_{\alpha,\mathrm{Pa}}\, \mu_{\mathrm{Pa}}$ in the numerator of $\tilde{z}^{(c)}$. (iii) If $\Gamma_k = \mathrm{Cov}_0(\hat{z}^{\mathrm{Pa}}_{t-k}, \hat{z}^{\mathrm{Pa}}_t)$ is nonsingular and the residual noise is uncorrelated at lag k, then $\Gamma_k^{-1}\mathrm{Cov}_0(\hat{z}^{\mathrm{Pa}}_{t-k}, \hat{z}^{(c)}_t) = b^\star_c$.*

For one parent c', the shift left in (ii) is the fraction $1 - \mathrm{rel}_{c'}/(1+\lambda)$ of the child's propagated shift. Here $\mathrm{rel}_{c'} = \alpha^2_{c'} \mathrm{Var}_0(z^{(c')})$ is the share of the parent's latent that its state carries. Under the hypotheses of (iii), a nonsingular Γ_k means that the parents' states are autocorrelated at lag k.

Theorem 3.2 holds for any fixed $\hat{B}$, since $\tilde{H}$ carries its error, and hence covers the LBNL FDD loop.

Why a ridge fit on the parents at the same sample? Air and water reach the next component of an air handler within minutes, well inside one sample of LBNL FDD, which records once per hour. The regression therefore reads the parents at the same sample. Reading them one sample earlier lowers the average Top-1 from 0.52 to 0.47 ($p = 0.007$), most of it on LBNL FDD. The ridge penalty keeps the fit stable when two parents move together. Coefficients from the instrument of (iii) at $k = 1$, used where the parents are autocorrelated, reach the same average Top-1 of 0.52. CULPRIT uses the ridge fit, which needs no autocorrelation of the parents.

For Theorem 3.2, write $\bar{\Delta}(c) = \mathbb{E}_1|\tilde{z}^{(c)}_t|$ for the population separated displacement and ξ_t for the shift of Lemma C.5. The components are *equally noisy* when their null parts $\epsilon^{(c)}_t$ share one symmetric law with zero in its support across components and samples. By Lemma C.5, the standard Gaussian is such a law when the healthy structural residuals are Gaussian and ν keeps its law across the windows.

Theorem 3.2 (Ranking under bounded smearing), full statement. *Let Assumptions C.1 to C.3 hold. (i) Let the components be equally noisy. For a constant shift $\vartheta_t = \vartheta$, $\bar{\Delta}$ ranks $S^\star$ first iff $\min_{r\in S^\star}|\xi_r| > \max_{c\notin S^\star}|\xi_c|$. For a varying shift, it does so whenever $|\xi_{t,r}| \geq |\xi_{t,c}|$ for each $r \in S^\star$, $c \notin S^\star$ and t, strictly at some t. (ii) Let the null parts have variance at most $\bar{\sigma}^2$. For a constant shift, $\bar{\Delta}$ ranks $S^\star$ first whenever $\min_{r\in S^\star}|\xi_r| > \max_{c\notin S^\star}(\bar{\sigma}^2 + \xi_c^2)^{1/2}$. (iii) Both conditions of (i) hold when $\tilde{H} = 0$, and the constant-shift condition holds whenever*

$$\min_{r\in S^\star} \frac{|\alpha_r \vartheta_r| - |(\tilde{H}\vartheta)_r|}{\tilde{s}_r} > \max_{c\notin S^\star} \frac{|(\tilde{H}\vartheta)_c|}{\tilde{s}_c}.$$

Dividing all $\bar{\Delta}(c)$ by the one positive number $\mathrm{med}^+_{c'}\bar{\Delta}(c')$ keeps their order, hence (i) and (ii) also hold for the ranking by $\bar{\Delta}(c)/\mathrm{med}^+_{c'}\bar{\Delta}(c')$. Part (ii) covers components that are noisier in the fault window than in the healthy window.

Corollary C.8 (Anchored components)**.** *If the projection direction of each anchored component is supported on its local sensors, the rows of H of the anchored components vanish. If all components are anchored, then $H = 0$. In that case, for equally noisy components and under the conditions of Proposition C.7(iii) with its coefficients, the ranking of Theorem 3.2 places $S^\star$ first whatever the shared sensors are.*

Why all sensors of a component? A shared sensor still carries its component's own signal, e.g., the supply air temperature on an air handler reads the coil and the supply fan at once. Restricting the directions of the anchored components to their local sensors drops that signal, and the average Top-1 falls from 0.52 to 0.47, mostly on SDAHU and DDAHU. CULPRIT therefore projects all sensors of each component, and Theorem 3.2 accounts for the components sharing them through $\tilde{H}$.

Proposition C.9 (A source component inside its child's sensors)**.** *Let a component h with no parent be the only parent of its only child l, with $\mathcal{S}_h \subseteq \mathcal{S}_l$. For any laws of the exogenous processes and the noise and for generic loadings, each model of Assumptions C.1 to C.3 with $S^\star = \{l\}$ gives the same sensor series in both windows as some model of these assumptions with $S^\star = \{h\}$.*

Such an h is unresolved, as l is assigned all of its sensors. No method reading the sensors, the association and the edges can place $S^\star$ first in both models. The two models exchange the exogenous

sources of h and l, the neighbour-mixing ambiguity of linear causal disentanglement (Squires et al., 2023), here inside a known sparsity pattern.

For a component c whose sensors are all local, write $\Sigma_c = \mathrm{Var}_0(z^{(c)})\, a_c a_c^\top + \Sigma_c^\nu$ with Σ_c^ν the healthy covariance of $\nu_t[\mathcal{S}_c]$, and write $v = a_c/\|a_c\|$.

Proposition C.10 (Projection direction)**.** *Let all sensors of c be local. (i) Any unit w with $w^\top a_c \neq 0$ gives a latent to which Proposition C.6 and Theorem 3.2 apply. (ii) If Σ_c^ν is positive definite, the share of the latent's variance carried by $z^{(c)}$ is largest at $w \propto (\Sigma_c^\nu)^{-1} a_c$. (iii) If $\mathrm{Var}_0(z^{(c)}) > 0$, then $\mathrm{Var}_0(z^{(c)})\|a_c\|^2 \sin\angle(w_c, v) \leq 2\|\Sigma_c^\nu - \bar{\psi} I\|$ for the first principal direction w_c and any $\bar{\psi}$.*

When $\Sigma_c^\nu = \bar{\psi} I$, the bound of (iii) is zero and $w_c = \pm v$.

Why the first principal direction? The average Top-1 drops from 0.52 to 0.47 with the noise-weighted direction of (ii), estimated with a one-factor fit on components with three or more sensors. Appendix G.9 measures how much of a real fault the projection keeps.

Common causes. A common cause of two components that a channel records is part of u_t, and Stage 1 removes it wherever the healthy window covers its fault-window values, as it removes a recorded ambient quantity (Appendix B).

C.3 Velocity score and combined score

Write $v_e(y)$ for the variance of the increments $d_t = y_t - y_{t-1}$ of a term y of Equation 8 in window e, and $\hat{v}_e(y)$ for its sample value. The absolute log ratio of term y is then $\rho(y) = |\log(\hat{v}_1(y)/\hat{v}_0(y))|$. Write $e_t = (d_t - \mathbb{E}d_t)^2$ for the centred squared increments, $\bar{e}_0$ and $\bar{e}_1$ for their window means, and $\mathrm{LRV}(e) = \sum_{k\in\mathbb{Z}} \mathrm{Cov}(e_0, e_k)$ for their long-run variance. For a latent term, $\hat{C}_e$ is the sample increment covariance of its component's sensors in window e.

Assumption C.11 (Weakly dependent increments)**.** For each term y, the increments d_t are strictly stationary, d_t and e_t have absolutely summable autocovariances, and $\mathrm{LRV}(e) > 0$. The window means obey $\big(n_0^{1/2}(\bar{e}_0 - \mathbb{E}e_t),\, n_1^{1/2}(\bar{e}_1 - \mathbb{E}e_t)\big) \to \mathcal{N}\big(0, \mathrm{LRV}(e)\, I_2\big)$, and the Newey–West estimate of Equation 9, computed on the sample-centred squares, satisfies $\widehat{\mathrm{LRV}}/\mathrm{LRV}(e) \to 1$ in probability. For each latent term, w_c converges in probability and $\hat{C}_e - C = O_P(n_e^{-1/2})$ in each window, with C the common increment covariance. The number of terms J is fixed, and $n_0, n_1 \to \infty$.

Strictly stationary increments that are strongly mixing across the onset satisfy the conditions of Assumption C.11 on d_t, e_t, their window means and $\widehat{\mathrm{LRV}}$ whenever $\mathrm{LRV}(e) > 0$, provided d_t has finite moments of order above eight and the mixing coefficients decay at the polynomial rate that Theorem 2 of Newey & West (1987) pairs with those moments. The covariance inequality and the central limit theorem of Herrndorf (1984, Lemma 2.1 and Corollary 1) give the summable autocovariances and the Gaussian limit of each window mean. Strong mixing across the onset makes the two window means asymptotically independent. The same Theorem 2 gives $\widehat{\mathrm{LRV}}/\mathrm{LRV}(e) \to 1$ in probability, because the bandwidth of Appendix E grows as $n_0^{2/9}$, slower than the $n_0^{1/4}$ rate the theorem allows. A combination rule is a map g with $g(\cdot, 0)$ strictly increasing on $[0, \infty)$ that scores component c by $g\big(\tilde{\Delta}(c)/\mathrm{med}_{c'}^{+}\tilde{\Delta}(c'),\, V(c)/\mathrm{med}_{c'}^{+}V^{\circ}(c')\big)$. The maximum of Equation 10 and the sum are two such rules.

Proposition 3.3 (The gate keeps the displacement ranking), full statement. *(i) Under Assumption C.11, let the fault window's standardized signals be, jointly in time, a process with the healthy law plus a constant vector. Then with probability at least $1-\alpha-o(1)$ all velocity scores are zero and each combination rule ranks the components as $\tilde{\Delta}$ does. (ii) If instead a term's increment variance moves from $v_0 > 0$ to a fixed $v_1 \neq v_0$, under Assumption C.11 in each window, or from $v_0 > 10^{-12}$ to $v_1 = 0$, as on a stuck device, under Assumption C.11 in the healthy window, then that term passes the gate with probability tending to one.*

When all velocity scores are zero, $\Psi(c) = \tilde{\Delta}(c)/\mathrm{med}_{c'}^{+}\tilde{\Delta}(c')$ for each c, since the second argument of Equation 10 is zero and the first is nonnegative. Under the premise of (i), Theorem 3.2 gives when the separated displacement ranks $S^\star$ first, and (i) transfers that ranking to Ψ at the price α. The premise holds for a constant displacement fault of Assumption C.3 on sensors whose spread

does not change with the operating condition, when the condition channels move alike in both windows. On the three physical systems their increment variance changes by a median of at most 7% (Appendix E).

A fault may instead change how fast a component moves. Let the standardized signals follow $\tilde{x}_t = h(u_t) + D_x(Wz_t + \nu_t)$ with a fixed diagonal D_x, the healthy response h, and z_t as in Assumption C.2. The model has *uncorrelated increments* when the increments of $h(u_t)$, of each exogenous process and of ν_t are mutually uncorrelated. A *velocity fault on* r moves the increment variance of $\zeta^{(r)}$ from v_0 to $v_1 \neq v_0$ and keeps the increment variances of $h(u_t)$, of the other exogenous processes and of ν_t the same in both windows. For a term $y = f^\top \tilde{x}$ of Equation 8 write $b_y = (f^\top D_x W (I - B)^{-1} e_r)^2$, and read the ratio a_y/b_y below as ∞ when $b_y = 0$.

Proposition C.12 (A velocity fault weakens downstream)**.** *In this model with uncorrelated increments, consider a velocity fault on r. (i) Each term y has $v_e(y) = a_y + b_y v_e$ with $a_y \geq 0$ fixed across the windows. (ii) The population log ratio $|\log(v_1(y)/v_0(y))|$ is zero when $b_y = 0$ and strictly decreasing in a_y/b_y otherwise. (iii) The population V° ranks r first if the latent term of r, or the term of a local sensor of r, has a smaller a_y/b_y than each term of the other components.*

The ratio a_y/b_y is the increment variance of a term that does not come from r, per unit of the part that does. A component downstream of r receives r's increments through the edges together with its own exogenous increments and noise. The condition therefore asks for a term that tracks r's state more cleanly than any term of another component, e.g., a local sensor of r. A stuck device, $v_1 = 0$, read by a term with $a_y = 0$ makes the log ratio infinite, which the variance clip of Appendix E caps.

C.4 Proofs

Proof of Lemma C.5. By Assumption C.2, $z_t = (I - B)^{-1}\zeta_t$. By Equation 12, $\tilde{z}_t = \tilde{D}^{-1}(I - \hat{B})\big(M(I - B)^{-1}\zeta_t + D^{-1}\hat{W}^\top \nu_t\big)$, a linear map $\mathcal{L}$ of (ζ_t, ν_t) that the healthy window fixes. In the fault window write $\zeta_{t,1} = \bar{\zeta}_t + \vartheta_t$, where $(\bar{\zeta}_t)_t \overset{d}{=} (\zeta_{t,0})_t$ by Assumption C.3. Linearity gives $\tilde{z}_{t,1} = \mathcal{L}(\bar{\zeta}_t, \nu_{t,1}) + \tilde{D}^{-1}(I - \hat{B})M(I - B)^{-1}\vartheta_t$, and the first term has mean zero since $\bar{\zeta}$ and ν do. When ν has one law in both windows, independence of ν from the latents gives $(\bar{\zeta}_t, \nu_{t,1}) \overset{d}{=} (\zeta_{t,0}, \nu_{t,0})$ jointly, hence ϵ_t has the law of $\tilde{z}_t$ on a fresh healthy window. For $\tilde{H}$, write $\hat{B} = B^\star + (\hat{B} - B^\star)$ and $M = D_\alpha + H$, then $(I - \hat{B})(D_\alpha + H) = D_\alpha(I - B) + (I - \hat{B})H - (\hat{B} - B^\star)D_\alpha$, since $B^\star D_\alpha = D_\alpha B$, and multiply by $(I - B)^{-1}$. □

With ν uncorrelated with ζ, the mean and covariance of $\tilde{z}$ still follow, settling the Gaussian case.

Proof of Proposition C.6. With $H = 0$ and $\hat{B} = B^\star$ the bracket defining $\tilde{H}$ vanishes, hence $\xi_{t,c} = \alpha_c \vartheta_{t,c}/\tilde{s}_c$. This shift is zero for all t exactly outside $S^\star$, and Lemma C.5 gives the translation. □

Proof of Proposition C.7. For such components $\hat{z}^{(c)} = \alpha_c z^{(c)} + n_c$ and $\hat{z}^{\mathrm{Pa}} = D_{\alpha,\mathrm{Pa}} z^{\mathrm{Pa}} + n_{\mathrm{Pa}}$ with noise terms n independent of the latents. Here $z^{(c)} = \beta_c^\top z^{\mathrm{Pa}} + \zeta^{(c)}$, and $\zeta^{(c)}$ is independent of z^{Pa} since c lies on no loop and is therefore no ancestor of its parents. Hence the healthy covariance of the parents' and the child's latents is $\mathrm{Cov}_0(\hat{z}^{\mathrm{Pa}}, \hat{z}^{(c)}) = D_{\alpha,\mathrm{Pa}}\mathrm{Cov}_0(z^{\mathrm{Pa}})\beta_c\alpha_c = R^z_{\mathrm{Pa}} b_c^\star$. The population ridge solution of Equation 7, on latents of unit healthy variance without an intercept, is $(R_{\mathrm{Pa}} + \lambda I)^{-1}$ times it, which is (i). For (ii), the numerator of $\tilde{z}^{(c)}$ moves by $\alpha_c\beta_c^\top \mu_{\mathrm{Pa}} - \hat{b}_c^\top D_{\alpha,\mathrm{Pa}}\mu_{\mathrm{Pa}} = (b_c^\star - \hat{b}_c)^\top D_{\alpha,\mathrm{Pa}}\mu_{\mathrm{Pa}}$, and

$$b_c^\star - \hat{b}_c = (R_{\mathrm{Pa}} + \lambda I)^{-1}(R^\nu_{\mathrm{Pa}} + \lambda I)\, b_c^\star$$

because $R_{\mathrm{Pa}} = R^z_{\mathrm{Pa}} + R^\nu_{\mathrm{Pa}}$. One parent c' has $R_{\mathrm{Pa}} = 1$ and $R^z_{\mathrm{Pa}} = \mathrm{rel}_{c'}$. For (iii),

$$\hat{z}_t^{(c)} = b_c^{\star\top}\hat{z}_t^{\mathrm{Pa}} + \alpha_c\zeta_t^{(c)} + n_{c,t} - b_c^{\star\top} n_{\mathrm{Pa},t}.$$

The instrument $\hat{z}^{\mathrm{Pa}}_{t-k}$ is uncorrelated with the last three terms for three reasons. First, z^{Pa}_{t-k} is a function of the processes of the ancestors of $\mathrm{Pa}(c)$, which exclude c. Second, the latents are independent of the noise, and third, the noise is uncorrelated at lag k. Hence $\mathrm{Cov}_0(\hat{z}^{\mathrm{Pa}}_{t-k}, \hat{z}_t^{(c)}) = \Gamma_k b_c^\star$. □

Proof of Theorem 3.2. Let X follow the common law of the null parts and $\phi(x) = \mathbb{E}|X + x|$. By Lemma C.5, $\bar{\Delta}(c)$ is the window average of $\phi(\xi_{t,c})$. The symmetry of X makes ϕ even. Moreover, ϕ is convex with right derivative $P(X \geq -x) - P(X < -x) = P(-x \leq X \leq x)$ for $x > 0$, which is positive since zero lies in the support of X. Hence ϕ is strictly increasing in $|x|$. For a constant shift $\bar{\Delta}(c) = \phi(\xi_c)$, and the order of $\bar{\Delta}$ is the order of $|\xi_c|$, which gives the equivalence. For a varying shift the condition gives $\phi(\xi_{t,r}) \geq \phi(\xi_{t,c})$ at each t with strict inequality at some t, and the window average keeps the strict inequality. For (iii), when $\tilde{H} = 0$, $\xi_{t,c} = \alpha_c \vartheta_{t,c}/\tilde{s}_c$ is zero for $c \notin S^\star$ and, since $\alpha_r \neq 0$, nonzero for $r \in S^\star$ at some t, which gives both conditions. In general $|\xi_r| \geq (|\alpha_r \vartheta_r| - |(\tilde{H}\vartheta)_r|)/\tilde{s}_r$ and $\xi_c = (\tilde{H}\vartheta)_c/\tilde{s}_c$ for $c \notin S^\star$, where $\vartheta_c = 0$. For (ii), Jensen's inequality gives $|\xi_c| \leq \mathbb{E}|\epsilon_t^{(c)} + \xi_c| \leq (\mathrm{Var}(\epsilon_t^{(c)}) + \xi_c^2)^{1/2}$, hence $\bar{\Delta}(r) \geq |\xi_r| > (\bar{\sigma}^2 + \xi_c^2)^{1/2} \geq \bar{\Delta}(c)$. Dividing all $\bar{\Delta}$ by one positive median keeps their order. □

Under a variance inflation and Gaussian laws, $\mathbb{E}|X|$ increases with the standard deviation of a Gaussian X at a fixed mean. With $H = 0$ and $\hat{B} = B^\star$ the exogenous term of c enters only $\tilde{z}^{(c)}$, hence the same ranking follows.

Proof of Corollary C.8. A local sensor of c lies in no $\mathcal{S}_{c'}$ with $c' \neq c$. Hence a direction supported on such sensors has $w_c^\top W[\mathcal{S}_c, c'] = 0$, and row c of H vanishes. With $H = 0$ Proposition C.7(iii) applies to all components and gives $\hat{B} = B^\star$, hence $\tilde{H} = 0$ and Theorem 3.2 applies. □

Proof of Proposition C.9. Write $z_t^{(l)} = \beta z_t^{(h)} + \zeta_t^{(l)}$ with $\beta \neq 0$ and $z_t^{(h)} = \zeta_t^{(h)}$, and let the first model shift $\zeta^{(l)}$ by $\vartheta_{t,l}$, hence $S^\star = \{l\}$. Define the second model by $z_t'^{(h)} = z_t^{(h)} - z_t^{(l)}/\beta = -\zeta_t^{(l)}/\beta$ and $z_t'^{(l)} = z_t^{(l)}$, which gives $z_t'^{(l)} = -\beta z_t'^{(h)} + \zeta_t'^{(l)}$ with $\zeta_t'^{(l)} = \beta\zeta_t^{(h)}$. Its exogenous processes $\zeta'^{(h)} = -\zeta^{(l)}/\beta$ and $\zeta'^{(l)} = \beta\zeta^{(h)}$ are independent processes of any law, and the edge $h \to l$ keeps the coefficient $-\beta$. The other components are unchanged, since only l among h and l has children outside the pair and $z'^{(l)} = z^{(l)}$. The mixing columns $a_h' = a_h$ and $a_l' = a_l + a_h/\beta$ give $W' z_t' = W z_t$ at each t, hence identical sensor series with the same noise. Moreover, a_l' is supported on $\mathcal{S}_h \cup \mathcal{S}_l = \mathcal{S}_l$ and nonzero there for generic loadings, which is the pattern of A. The fault moves $\zeta'^{(h)}$ by $-\vartheta_{t,l}/\beta$ and leaves $\zeta'^{(l)}$ unchanged, hence $S^\star = \{h\}$ in the second model. □

Proof of Proposition 3.3. (i) Fix a term y at the limit direction. Its increments in the fault window follow the healthy law, since the constant cancels in the difference. The standardization by healthy moments rescales both windows alike, which ρ ignores. Write $v = \mathbb{E}e_t$ for the common increment variance, which is positive because $\mathrm{LRV}(e) > 0$ excludes $e_t \equiv 0$. The sample variance of window e is $\hat{v}_e = \bar{e}_e - (\bar{d}_e - \mathbb{E}d_t)^2$, and the summable autocovariances of d_t make the last term $O_P(n_e^{-1})$. The joint central limit theorem and the delta method then give $\big(n_0^{1/2}(\log \hat{v}_0 - \log v),\ n_1^{1/2}(\log \hat{v}_1 - \log v)\big) \to \mathcal{N}\big(0, (\mathrm{LRV}(e)/v^2)\, I_2\big)$. Hence $\log(\hat{v}_1/\hat{v}_0)/\sigma_\rho(y) \to \mathcal{N}(0,1)$ with $\sigma_\rho(y) = \big((1/n_0 + 1/n_1)\,\mathrm{LRV}(e)/v^2\big)^{1/2} > 0$. The diagonal limit covariance is consistent with the summable autocovariances. The covariance of the two windows' sums of e_t is $\sum_k N_k \,\mathrm{Cov}(e_0, e_k)$, where $N_k \leq \min(k, n_0, n_1)$ counts the pairs at lag k across the onset, and dominated convergence makes it $o\big((n_0 n_1)^{1/2}\big)$. The relative consistency of $\widehat{\mathrm{LRV}}$ and $\hat{v}_0 \to v$ make $\hat{\sigma}_\rho(y)/\sigma_\rho(y)$ tend to one in probability. For a latent term, let $f(x) = \log(x^\top \hat{C}_1 x / x^\top \hat{C}_0 x)$ and let w be the limit direction, with $w^\top C w = v > 0$. Near w the gradient $2\hat{C}_1 x / x^\top \hat{C}_1 x - 2\hat{C}_0 x / x^\top \hat{C}_0 x$ is $O_P\big((1/n_0 + 1/n_1)^{1/2}\big)$, since $\hat{C}_1 - \hat{C}_0$ is of that order. The mean value theorem and $\hat{w}_c - w = o_P(1)$ then give $f(\hat{w}_c) - f(w) = o_P\big((1/n_0 + 1/n_1)^{1/2}\big)$, and Slutsky's lemma keeps the limit. Hence $P(\rho(y) > \Phi^{-1}(1 - \alpha/2J)\,\hat{\sigma}_\rho(y)) \to \alpha/J$, and a union bound over the J terms bounds the probability that any term passes by $\alpha + o(1)$. When none passes, $V \equiv 0$ and the second argument of the maximum in Equation 10 is zero, since med^+ is positive. The first argument is nonnegative, hence Ψ is $\tilde{\Delta}$ divided by one positive number. For a rule g, the score is $g(\tilde{\Delta}(c)/m, 0)$ with one positive m, and a strictly increasing $g(\cdot, 0)$ keeps the order of $\tilde{\Delta}$. (ii) $\log(\hat{v}_1/\hat{v}_0) \to \log(v_1/v_0) \neq 0$ in probability, while $\hat{\sigma}_\rho(y) \to 0$, as the Newey–West estimate converges and $1/n_0 + 1/n_1 \to 0$. Hence $\rho(y)$ passes the threshold with probability $1 - o(1)$. For $v_1 = 0$ the fault-window increments

vanish and the floor of Equation 8 sets $\hat{v}_1 = 10^{-12}$, hence $\rho(y) \to |\log(10^{-12}/v_0)| > 0$, while $\hat{\sigma}_\rho(y)$ reads only the healthy increments and still tends to zero. □

Proof of Proposition C.12. By Assumption C.2, $z_t = (I - B)^{-1}\zeta_t$, hence $y_t - y_{t-1} = f^\top(h(u_t) - h(u_{t-1})) + f^\top D_x W(I - B)^{-1}(\zeta_t - \zeta_{t-1}) + f^\top D_x(\nu_t - \nu_{t-1})$. The three parts are uncorrelated and the exogenous increments are mutually uncorrelated, hence $v_e(y) = a_y + b_y v_e$. Here a_y collects the parts other than $\zeta^{(r)}$ and takes one value in both windows. With $b_y > 0$ the log ratio is $|\log(1 + (v_1 - v_0)/(a_y/b_y + v_0))|$, which decreases strictly in a_y/b_y because $v_1 \neq v_0$, and with $b_y = 0$ it is zero. A term of r with the smallest a_y/b_y thus exceeds each term of another component, and dividing a shared sensor's term by $|A_s| \geq 1$ only lowers it. □

Proof of Proposition C.10. (i) is Equation 12 with no smearing term. (ii) The share is increasing in $(w^\top a_c)^2/w^\top \Sigma_c^\nu w$, a generalized Rayleigh quotient, which a positive definite Σ_c^ν maximizes at $(\Sigma_c^\nu)^{-1}a_c$. (iii) Write $\Sigma_c = (\mathrm{Var}_0(z^{(c)})\, a_c a_c^\top + \bar{\psi} I) + (\Sigma_c^\nu - \bar{\psi} I)$. The first term has top eigenvector v and eigengap $\mathrm{Var}_0(z^{(c)})\|a_c\|^2$, and the Davis–Kahan theorem of Yu et al. (2015) bounds $\sin \angle(w_c, v)$ by $2\|\Sigma_c^\nu - \bar{\psi} I\|$ divided by this eigengap. □

D Datasets and Preprocessing

Section 4.1 lists the six benchmarks with their case counts and operating conditions, and this appendix adds the details that list leaves out.

Case counts. N-CMAPSS ships 128 engines across ten subset files. The DS08d-010 file is truncated inside NASA's archive and does not open, which removes its 29 engines and leaves the 99 evaluated here. SWaT documents 36 attacks with valid start and end times once the five with no physical effect are omitted (Inoue et al., 2017). We consume the curated attack list bundled with AERCA and StaR, which keeps the 25 attacks whose target lies among the 51 recorded sensors. Five of those carry no attack-marked telemetry and produce an empty fault window, which leaves the 20 cases StaR also reports. The Causal Chamber count has no dataset-paper total to compare against, because the chamber paper defines no root-cause task on the wind tunnel (Gamella et al., 2025b). The 28 cases are the 28 experiments of the released `wt_validate_v1` dataset. Each experiment flips a single actuator between a low and a high value while all other actuators are held fixed. The flipped actuator is the root cause. The low-value samples make up the healthy window and the high-value samples the fault window. The 28 experiments span 21 distinct actuators. LBNL FDD labels faults of components, of sensors and of the control sequence (Granderson et al., 2023). We evaluate the faults of components, 12 on SDAHU and 37 on DDAHU, since a biased sensor or a mistuned control loop names no component of the system model.

Synthetic chain. The saturation sweep of Appendix G.8 adds a linear chain generator with five components in a line. Each component has one AR(1) latent driven by a slowly varying condition channel and by its upstream neighbour at coupling 0.6. Each component also carries two local sensors and one sensor shared with the next component. The sensor noise is Gaussian at the latent scale. A case shifts the mean and variance of one component's mechanism from sample 1500 of 2000, and 20 seeds over 5 root causes give 100 cases. The generator is linear by construction, hence the sweep's nonlinearity is the perturbation alone.

Declared edges. LBNL FDD and OpenRCA come with declared edges, and on N-CMAPSS the edges follow the published gas-path station order. The Causal Chamber ships a validated graph that points actuators at sensors and holds no actuator-to-actuator edge, and RCAEval and SWaT ship none. Hence, on all three benchmarks, the three graph-consuming baselines (DoWhy-GCM and the two traversals) run on a component DAG learned by PC from the healthy prefix, described below. On RCAEval each method, CULPRIT included, reads no declared edges. DoWhy-GCM requires a DAG, and LBNL FDD's airflow graph closes a recirculation loop. Hence for this method we drop the return-fan to damper edge, the one edge that closes it, and the traversals keep the loop.

Telemetry preparation. Each case is cut into a healthy prefix and a fault window at the benchmark's own onset label. A case whose healthy window holds fewer than 8 samples is dropped. OpenRCA needs the most preparation, since its metric CSVs arrive per service and per level. They are pivoted to a one-minute grid and forward-filled over gaps of at most five minutes. Constant sensors are

dropped, leaving 830 to 1821 live sensors per case. Five of its 335 records sit within 7 minutes of the start of their telemetry day and leave fewer than 8 healthy samples, hence 330 cases remain.

E Implementation Details and Compute

CULPRIT is training-free, and the settings below serve all benchmarks without per-dataset tuning.

Stage 1. Stage 1 uses tolerance $\tau = 0.10$ and requires at least $n_{\min} = 4$ matched healthy samples per scorable fault sample. It scores at most 2048 evenly spaced samples per window, each against the whole healthy window. The maximum in Equation 4 is taken over an evenly spaced subsample of at most 512 fault samples. This bounds the matching at $O(|\mathcal{T}_0| \cdot 512 \cdot q)$ distance evaluations for q condition channels. Section 3.1 defines when a sensor is scorable. Its spread requirement asks the matched standard deviation to exceed a small fraction of the sensor's healthy spread, which discards a sensor whose matched values are constant to floating-point precision. When no fault sample of a case is scorable, Stage 1 falls back to the whole-window reference over the healthy support $\mathcal{U}_0$.

Stages 2 and 3. Stage 2 needs at least 8 healthy samples per case. Structural separation regresses a component's latent on its parents' latents with ridge parameter 0.01. The velocity score needs at least 3 fault samples. It uses the 2048-sample cap in its latent form and the full windows in its per-sensor form. The gate of Equation 9 estimates the long-run variance with the Bartlett kernel and bandwidth $\lfloor 4(n_0/100)^{2/9} \rfloor$ and counts all terms of a case at level $\alpha = 0.01$. The optional set cut of Appendix G.6 returns each component whose combined score Ψ of Equation 10 exceeds five.

Healthy residuals. The direction w_c, the scale s_c, the ridge coefficients and $\tilde{s}_c$ are fitted on up to 2048 evenly spaced healthy samples. Each of them is scored against all healthy samples in its own ball, its time neighbours included, since those are the samples at its operating point. We cut placebo windows from the healthy data, with its even samples as the healthy window and its odd samples as the fault window. On these windows, the separated displacement of a component shows no consistent trend with its number of sensors. Its rank correlations with that number lie between -0.25 and 0.30 on the six datasets whose components differ in their number of sensors.

Velocity score. The velocity score compares two windows that see similar condition dynamics. The median absolute log ratio of the condition channels' increment variance between the windows is 0.07 on N-CMAPSS, 0.01 on SDAHU and 0.00 on DDAHU. This similarity of condition dynamics is the premise of Proposition 3.3. Without the gate, dividing ρ by the ratio the healthy window predicts at the fault window's condition dynamics gives an average Top-1 of 0.51 ($p = 0.11$). Computing the velocity score on one-step increments of the matched residuals gives 0.42, against 0.52 for the ungated score. The synthetic displacement faults along the loading of Appendix G.9 change no increment law. On them the ungated score ranks the root cause first in 0.85 and 0.92 of the cases with three and six sensors, and the gate lifts both fractions to 0.99.

Unscorable entries. Unscorable entries enter the projection as zero, i.e., as no evidence of a shift, and they arise in three forms. A fault sample whose matched set holds fewer than $n_{\min}$ samples has no scorable sensor. Such a sample sets all latents to zero, and with them the structural residuals, since the regression of Equation 7 has no intercept. Hence such a sample scales each component's displacement by the same multiplier, which the division by the median in Equation 10 cancels exactly. A sensor whose matched values are constant at that operating condition is unscorable while the other sensors of its component stay scorable. Examples include a fan held off or at a fixed speed, or a damper or a valve at its end stop. A sensor whose value is missing at that sample or recorded in fewer than $n_{\min}$ of its matched samples is unscorable as well. The direction w_c and the scale s_c are fitted on healthy residuals that carry the same zeros. The fraction of $\|w_c\|^2$ left on the scorable sensors averages 0.97 in both windows on SDAHU and 0.84 in both on DDAHU. Three variants confirm that the ranking is robust to the zeros. The first two renormalize the direction over the scorable sensors of each sample, to unit norm or by least squares. Both give an average Top-1 of 0.52 ($p = 0.56$ and $p = 0.71$ over all 1,261 cases). The third averages each component's displacement over the fault samples where it keeps a scorable sensor and also gives 0.52 ($p = 0.32$). Each of the three changes the Top-1 of at most seven cases, and no dataset moves significantly.

Compute and seeds. All runs execute on CPU-only machines with 4 cores and 16 GB of memory. The deep learning baselines use seeds 0, 1 and 2.

F BASELINE CONFIGURATION

Each baseline except CondAttr runs the authors' released code under the evaluation protocol of Section 4.1, and the deep learning baselines train per case on the healthy prefix. The paragraphs below give the adapter choices that change a baseline's numbers.

AERCA. AERCA (Han et al., 2025) uses the better of the authors' two published configurations on our data. It scores deviations two-sided, since our faults can push a sensor either way while the authors' benchmark injects only positive shifts. The change stays within the seed spread. The port reproduces the authors' published accuracy on their own linear benchmark. Five Market cases per seed fail with a non-finite divergence term.

StaR. StaR (Huang et al., 2026) extends AERCA with a temporal-graph-network memory and runs under the same protocol. Its coefficient networks scale cubically in the sensor count. On OpenRCA we therefore train it on one series per component, the mean of the component's z-scored channels, which is the input DoWhy-GCM receives on all benchmarks. This reduces 830 to 1,809 channels to 15 to 70 series, and the method, its hyperparameters and its scoring are unchanged. StaR still fails 18 of the 48 Telecom cases and 2 of the 134 Bank cases per seed with a non-finite divergence term. The released code bypasses part of the message-passing step. The paper-faithful fix changes N-CMAPSS Top-1 by -0.05 (seed 0 of both, not significant). We therefore report the method as released.

RUN. RUN (Lin et al., 2024) trains one gated DLinear forecaster per sensor, prunes the discovered graph to a DAG, and ranks components by personalised PageRank, with the top post-onset deviation as its trigger. One forecaster per sensor costs a median of 32 minutes per N-CMAPSS engine and 49 minutes per RCAEval case (Table 18). We report the released graph orientation, since the paper's reversal is weaker on each dataset (0.42 against 0.51 Top-1 on N-CMAPSS). On OpenRCA, RUN trains on the same one series per component as StaR. On SDAHU the released cycle pruning keeps 1 to 7 of the 89 to 108 discovered edges and isolates the true component in all 12 cases. RUN therefore scores 0.00 on SDAHU.

CondAttr. CondAttr (Mishra et al., 2026) has no released code, and we implement the paper's method in full. Its VAE anomaly detector uses window 50, or a quarter of the prefix on OpenRCA. The three nearest healthy windows are then retrieved in the VAE latent space. A sensor's attribution is the drop in detector score when its trajectory is swapped for the neighbour's. The port reaches 0.57 Top-3 on SWaT, against the paper's 0.54 over its 36 attacks. The three seeds move the nine-dataset mean by at most 0.02 and SWaT by 0.15 over its 20 cases.

ALoRa. ALoRa (Shimillas et al., 2026) runs the authors' released transformer with their window of 20. Its localization score is the per-series reconstruction error weighted by the contribution matrix read off the trained weights. We read the matrix in the orientation of the released code, which reproduces the paper's SMD table on the authors' artefacts. We report the full contribution matrix, the paper's choice for SWaT and MSDS. The matrix limits the method on a single-root-cause ranking. After row normalisation its diagonal holds 6 to 8 percent of a row on N-CMAPSS. The weighted score therefore averages the error over all series. The paper's reconstruction-only variant (ALoRa-Loc*) reaches 0.65 Top-1 on RCAEval, below CULPRIT at 0.71. Across the three seeds the mean varies by 0.03, and Top-1 on the 12 SDAHU cases spans 0.08 to 0.50.

BRCD. In BRCD (Lee et al., 2026), BOSS learns a CPDAG from at most 4,000 evenly spaced samples of the healthy prefix. Each channel's posterior as the intervened node is the linear-Gaussian marginal likelihood summed over the compatible orientations. The procedure is deterministic. BOSS takes a median of 7 seconds per case and 5 to 7 minutes on the wide RCAEval systems. On the raw OpenRCA channels BOSS needs 38 minutes to over 4 hours per case. BRCD therefore trains on the same one series per component as StaR and RUN. On this input one Market case fails on a singular covariance.

LatentScope. LatentScope (Xie et al., 2024) runs from the authors' released implementation (commit `fbe86c7`) with a deterministic tie-break. It uses each case's healthy-prefix length instead of the fixed 111-sample normal window. The benchmark's component edges are lifted to sensor pairs and filtered at Pearson correlation 0.4 as the authors do, with at most five parents per sensor. The parent cap makes the lifted graph scorable. Without it, a Market frontend sensor carries 46 to 58 regression

parents on 50 training samples, and Market Top-1 falls to 0.03. The reported row is Algorithm 1 as published, the stronger of the two variants (0.33 against 0.27 average Top-1).

The traversal baselines' target rule. Traversal and Smooth Traversal (Orchard et al., 2025) score only the ancestors of a designated target component, hence the choice of target bounds their accuracy before either method runs. Selecting the component with the largest post-onset deviation is the rule DoWhy-GCM uses. On DDAHU this rule picks a chilled-water pump whose ancestor set contains only itself, which leaves the true root cause unreachable in 30 of 37 cases. We therefore select the most downstream component, the one with the largest ancestor set, breaking ties by deviation. This rule reads graph structure alone and makes the true root cause reachable in all 37 DDAHU cases. It roughly triples Smooth Traversal's DDAHU accuracy relative to the deviation rule.

Learned graphs on the benchmarks without one. The PC algorithm (Spirtes et al., 2001) learns these graphs from the healthy prefix of the same one-series-per-component input the baselines read elsewhere. We use the causal-learn implementation (Zheng et al., 2024) with the Fisher-z test at $\alpha = 0.05$ and conditioning sets of at most three variables. The cap lets PC terminate on the microservice benchmarks. There the component means are so mutually dependent that the uncapped search runs past twelve minutes on a 15-node Sock Shop case. It does not finish on the 64-node Train Ticket cases. Every capped run completes within 25 seconds. Components whose series coincide up to sign and scale make the correlation matrix singular. Examples are the pump and UV unit P402 and UV401 on SWaT and seven actuator pairs of the Chamber that average the same sensors. Hence one of each pair enters PC and the other inherits its edges. The undirected edges of the CPDAG are oriented by column order and any cycle is broken as for LBNL FDD. PC fails on four Chamber cases whose 21 series have rank 11.

DoWhy-GCM. DoWhy-GCM (Blöbaum et al., 2024) fits the benchmark's component graph on the healthy prefix and attributes the post-onset samples by the Shapley values of Budhathoki et al. (2022). Its attribution cost is linear in the number of post-onset samples. We therefore fit on at most 4000 evenly spaced healthy samples, the prefix cap the deep learning baselines use, and attribute at most 500 evenly spaced post-onset samples.

G Additional Results

G.1 Per-Dataset RCA Effectiveness

Tables 4 to 8 give the per-dataset results behind the summary of Section 4.2, in the order the benchmarks appear there. They follow the convention of Table 1, i.e., best in bold, second best underlined. AERCA, StaR and RUN show the mean over three seeds, except on the Causal Chamber, where they show seed 0 with the other seeds within 0.03 Top-1. CondAttr, ALoRa and BRCD show seed 0 here, where Table 1 averages their three seeds (spread in Appendix F).

G.2 OpenRCA

OpenRCA (Xu et al., 2025) ships the telemetry of three AIOps Challenge systems, and each fault carries a component, a failure reason and an onset time. We score the component element alone, since a ranking method emits neither the failure reason nor an onset inside the benchmark's one-minute window. OpenRCA combines the three elements into seven task types. Our numbers are therefore not comparable to the accuracies reported for language-model agents on this benchmark. Table 9 reports the mean over the three systems, as Table 1 does.

G.3 Adjustment and Matching Ablations

Three tables isolate the two design decisions Stage 1 makes, the adjustment variant and the rule that selects the condition reference. Table 10 crosses condition matching (none, case-level, pointwise) with structural separation (off, on) on N-CMAPSS, and Section 4.3 discusses the comparisons that change one of the two. Table 11 runs the same adjustment ladder on the 49 LBNL FDD cases, where separation carries the gain. Table 12 replaces the tolerance ball with seven alternative matching rules, with everything else held at the reported configuration (Section 4.4).

Table 4: RCA performance on N-CMAPSS (99 run-to-failure engines).

Method	Top-1	Top-3	Avg@5
ALoRa	0.45	0.64	0.68
BARO	0.50	0.74	0.74
BRCD	0.57	**0.87**	0.81
Cholesky	0.47	0.76	0.74
CIRCA	0.58	0.81	0.80
CondAttr	0.58	0.86	0.82
DoWhy-GCM	0.38	0.61	0.64
LatentScope	0.44	0.80	0.75
RCD	0.46	0.70	0.71
Score Ordering	0.54	0.67	0.72
Smooth Traversal	0.35	0.70	0.69
Traversal	0.36	0.68	0.68
AERCA	0.43	0.64	0.68
StaR	0.48	0.61	0.68
RUN	0.48	0.68	0.71
CULPRIT	**0.68**	0.86	**0.84**

Table 5: RCA performance on the Causal Chamber wind tunnel (28 cases).

Method	Top-1	Top-3	Avg@5
AERCA	0.10	0.31	0.29
ALoRa	0.09	0.28	0.26
BARO	0.15	0.46	0.42
BRCD	0.24	**0.73**	**0.68**
Cholesky	0.13	0.44	0.41
CIRCA	0.15	0.46	0.42
CondAttr	0.18	0.48	0.45
DoWhy-GCM	0.01	0.12	0.12
LatentScope	0.15	0.46	0.42
RCD	0.05	0.14	0.14
RUN	0.12	0.32	0.30
Score Ordering	0.09	0.27	0.27
Smooth Traversal	0.01	0.06	0.05
StaR	0.08	0.26	0.25
Traversal	0.02	0.09	0.10
CULPRIT	**0.39**	0.71	**0.68**

Table 6: RCA performance on LBNL FDD (49 cases on two air handlers).

Method	LBNL FDD (49 cases)			
	Top-1	Top-3	Top-5	Avg@5
ALoRa	0.04	0.28	0.37	0.23
BARO	0.07	0.45	0.65	0.41
BRCD	0.71	**0.92**	0.94	0.87
Cholesky	0.52	0.73	**1.00**	0.78
CIRCA	0.23	0.42	0.54	0.40
CondAttr	0.50	0.82	0.91	0.76
DoWhy-GCM	0.05	0.27	0.33	0.24
LatentScope	0.33	0.35	0.41	0.36
RCD	0.05	0.16	0.27	0.16
Score Ordering	0.46	0.71	0.82	0.68
Smooth Traversal	0.16	0.22	0.24	0.22
Traversal	0.09	0.22	0.32	0.21
AERCA	0.10	0.28	0.38	0.26
RUN	0.03	0.12	0.26	0.13
StaR	0.03	0.11	0.21	0.12
CULPRIT	**0.77**	**0.92**	0.92	**0.89**

Table 7: RCA performance on SWaT (20 attack cases, 51 sensors).

Method	Top-1	Top-3	Avg@5
ALoRa	0.10	0.10	0.12
BARO	0.15	0.40	0.36
BRCD	0.15	0.23	0.26
Cholesky	**0.25**	0.42	0.39
CIRCA	**0.25**	0.38	0.39
CondAttr	0.15	**0.57**	**0.53**
DoWhy-GCM	0.10	0.23	0.19
LatentScope	0.10	0.38	0.40
RCD	**0.25**	0.32	0.30
Score Ordering	0.19	0.35	0.33
Smooth Traversal	0.10	0.25	0.20
Traversal	0.09	0.25	0.22
AERCA	0.13	0.23	0.23
StaR	0.02	0.06	0.06
RUN	0.02	0.07	0.07
CULPRIT	**0.25**	0.42	0.39

G.4 The Design-Decision Ablation on All Benchmarks

Table 13 reports the configurations of Table 3 on each of the six benchmarks, together with one further change of the displacement score. Replacing the projection by the sensor maximum lowers LBNL FDD from 0.77 to 0.45, OpenRCA from 0.34 to 0.22 and RCAEval from 0.71 to 0.66, each at $p < 0.05$. The SWaT column is the same in the reported and the sensor maximum rows, since each SWaT component has a single local sensor and the projection is then the identity up to the scale s_c. The two-latent-dims row keeps the first two principal directions of each component and divides both by the healthy standard deviation of the first. It then scores the window mean of the norm of the two projections, which is the displacement score on a component with one sensor. Against displacement only, it drops N-CMAPSS from 0.69 to 0.56 and RCAEval from 0.73 to 0.69, both at $p < 0.05$. Dividing the second direction by its own, smaller spread instead drops SDAHU from 0.83 to 0.17. Under this scaling, a displacement along that direction counts more than the same displacement along the first.

Table 8: Overall RCA performance on RCAEval (735 failure cases).

Method	Top-1	Top-3	Top-5	Avg@5
AERCA	0.15	0.29	0.40	0.28
ALoRa	0.06	0.19	0.31	0.19
BARO	0.23	0.77	0.89	0.66
BRCD	0.30	0.65	0.78	0.59
Cholesky	0.06	0.17	0.30	0.17
CIRCA	0.51	0.84	0.94	0.78
CondAttr	0.69	**0.90**	**0.96**	0.86
DoWhy-GCM	0.50	0.55	0.59	0.55
LatentScope	**0.71**	**0.90**	**0.96**	**0.87**
RCD	0.33	0.43	0.50	0.42
RUN	0.08	0.24	0.38	0.24
Score Ordering	0.06	0.16	0.30	0.17
Smooth Traversal	0.05	0.13	0.21	0.13
Traversal	0.04	0.14	0.25	0.15
CULPRIT	**0.71**	0.87	0.92	0.84

Table 9: RCA performance on OpenRCA, mean over its three systems (Bank 134 cases, Market 148, Telecom 48).

Method	Top-1	Top-3	Avg@5
AERCA	0.06	0.14	0.15
ALoRa	0.05	0.14	0.14
BARO	0.14	0.31	0.32
BRCD	0.12	0.38	0.36
Cholesky	0.16	0.27	0.26
CIRCA	0.19	0.35	0.33
CondAttr	0.33	**0.60**	**0.57**
DoWhy-GCM	0.17	0.26	0.25
LatentScope	0.24	0.47	0.45
RCD	0.08	0.18	0.17
RUN	0.03	0.09	0.09
Score Ordering	0.03	0.10	0.10
Smooth Traversal	0.14	0.27	0.25
StaR	0.03	0.09	0.09
Traversal	0.15	0.25	0.24
CULPRIT	**0.34**	0.55	0.52

Table 10: The condition matching x separation ablation on N-CMAPSS (99 cases), Top-1 / Avg@5 per cell. Each cell is the reported configuration with adjustment and separation set as labelled.

Adjustment	Separation off	Separation on
None	0.40 / 0.66	0.35 / 0.67
Case-level	0.39 / 0.67	0.36 / 0.67
Pointwise (reported)	0.72 / 0.84	0.68 / 0.84

G.5 Granting the Baselines the Ingredients

Section 4.3 asks whether the gain of CULPRIT comes from its two ingredients, adjustment and the association, or from how it consumes them. The experiments below produce Table 2.

Adjustment. A wrapper rewrites each case into Stage 1's per-operating-point residuals, the healthy prefix and the fault window alike, and hands the rewritten case to the baseline unchanged. The healthy residual leaves each sample out of its own matched set, as a fault-window sample is left out of its own. We run two rules for the reference. The tolerance ball is CULPRIT's own matcher at $\tau = 0.10$. The regime rule partitions the standardised healthy conditions into eight k-means regimes and standardises each sample by its regime. This is the operating-regime normalisation of the prognostics literature (Wang et al., 2008). Where a case carries no condition, or no fault-window sample has a matched reference, the wrapper is the identity. The regime rule adjusts 96 to 99% of the OpenRCA cases. Table 14 reports nine baselines under both rules. AERCA, StaR and RUN are excluded, because residuals remove the lag structure their scoring rests on.

Table 11: Adjustment ablation on LBNL FDD, Top-1 over all 49 cases. The first three rows vary only the adjustment with separation on, and the last row turns separation off.

Adjustment	Top-1	p vs. pointwise
No adjustment	0.745	7.4×10^{-1}
Case-level adjustment	0.765	1.0×10^{0}
Pointwise adjustment (reported)	0.765	–
Pointwise, separation off	0.561	1.6×10^{-3}

Table 12: Condition-matcher ablation, Top-1. Each row replaces only the rule selecting the matched healthy reference; pointwise matching and separation stay at the reported configuration. † marks a difference from the tolerance ball at $p < 0.05$.

Matching rule	N-CMAPSS	SDAHU	DDAHU
Tolerance ball (reported)	0.68	0.92	0.72
Regime clustering	0.45†	0.67	0.61
k-nearest neighbours	0.69	0.83	0.47†
k-NN, bias-corrected	0.68	0.92	0.30†
Kernel weighting	0.69	0.75	0.73
CV-tuned tolerance	0.51†	0.67	0.73
Regression-based matching	0.71	0.67	0.77
Regression, heteroscedastic	0.68	1.00	0.73

Table 13: The design-decision ablation on each of the six benchmarks, Top-1, aggregated as in Table 1, with the average over the six benchmarks. Upper rows change one decision of CULPRIT, lower rows change one decision of the displacement score.

Configuration	N-CMAPSS	LBNL FDD	OpenRCA	Causal Chamber	RCAEval	SWaT	Avg
CULPRIT	0.68	0.77	0.34	0.39	0.71	0.25	**0.52**
Changed in CULPRIT							
sensor maximum	0.69	0.45	0.22	0.27	0.66	0.25	0.42
no velocity	0.62	0.45	0.35	0.17	0.73	0.30	0.44
latent form only	0.68	0.43	0.33	0.39	0.71	0.25	0.47
per-sensor form only	0.62	0.81	0.34	0.23	0.71	0.20	0.49
Changed in the displacement score							
displacement only	0.69	0.22	0.38	0.17	0.73	0.30	0.41
no adjustment	0.43	0.18	0.38	0.17	0.72	0.00	0.31
max latent deviation	0.57	0.12	0.34	0.10	0.67	0.12	0.32
post-hoc penalty	0.37	0.22	0.21	0.17	0.72	0.00	0.28
two latent dims	0.56	0.33	0.39	0.13	0.69	0.30	0.40

The association. The baselines read the association through the aggregation of sensor scores to components, and the graph baselines receive the same edges as well. The association is therefore granted already, and the question is how much the correct one buys. Table 16 applies the derangement draws of Table 15 to five baselines. Each method that has accuracy to lose loses it, e.g., Cholesky on DDAHU drops from 0.64 to 0.01. These losses show that the baselines do consume the association. On N-CMAPSS the derangement costs them 0.03 to 0.09 Top-1 against 0.30 for CULPRIT. Their loss is small because a maximum over a component's sensors only relabels the highest-scoring sensor. On the gas path that sensor's label is nearly as often wrong as right. On the two air handlers ten draws replace three, since a single permutation can move the mean by a third. With twelve cases, this happens when the permutation lands the highest-scoring sensor on the majority root cause. A full derangement sets the accuracy floor. Under the same Expected Top-k a random ranking scores 0.05 on Causal Chamber, 0.05 on LBNL FDD and 0.47 on N-CMAPSS. The last is high because that benchmark has five components and its root cause set often names several of them. The deranged column of Table 15 thus sits at or below chance on all three. The graded sweep of Appendix G.8 measures the range between a correct association and that floor.

G.6 An Optional Ranking-to-Set Cut

One threshold on the combined score Ψ of Equation 10 cuts the ranking into a set,

$$\widehat{S} = \{\, c \in \mathcal{C} : \Psi(c) > \kappa \,\}, \qquad \kappa = 5, \tag{14}$$

with the top-ranked component alone when none clears it. Both scores in Ψ are divided by their median over the components, hence $\kappa = 5$ keeps a component that scores five times the median component on either score. The same value serves all datasets. Table 17 reports the cut for CULPRIT

Table 14: Each baseline given Stage 1, Top-1 per dataset as raw / tolerance ball / regime. On Bank and Market the ball matches almost no case and repeats the raw value. DoWhy-GCM is partial on DDAHU under both rules. The last row is CULPRIT.

Method	N-CMAPSS	SDAHU	DDAHU	Bank	Telecom	Market
BARO	0.50 / 0.58 / 0.47	0.29 / 0.54 / 0.17	0.00 / 0.11 / 0.08	0.31 / 0.31 / 0.38	0.04 / 0.06 / 0.15	0.08 / 0.08 / 0.25
Cholesky	0.47 / 0.44 / 0.46	0.17 / 0.04 / 0.00	0.64 / 0.27 / 0.34	0.27 / 0.26 / 0.29	0.09 / 0.07 / 0.08	0.12 / 0.12 / 0.13
CIRCA	0.58 / 0.59 / 0.62	0.33 / 0.33 / 0.33	0.20 / 0.16 / 0.27	0.34 / 0.34 / 0.28	0.06 / 0.03 / 0.03	0.16 / 0.16 / 0.14
RCD	0.46 / 0.46 / 0.46	0.10 / 0.10 / 0.10	0.04 / 0.04 / 0.04	0.09 / 0.14 / 0.16	0.04 / 0.00 / 0.02	0.12 / 0.14 / 0.14
DoWhy-GCM	0.38 / 0.63 / 0.40	0.17 / 0.67 / 0.33	0.01 / 0.16 / 0.17	0.25 / 0.25 / 0.40	0.04 / 0.12 / 0.12	0.22 / 0.22 / 0.31
Traversal	0.36 / 0.46 / 0.45	0.09 / 0.10 / 0.26	0.09 / 0.26 / 0.29	0.27 / 0.27 / 0.18	0.14 / 0.08 / 0.10	0.04 / 0.04 / 0.04
Smooth Traversal	0.35 / 0.40 / 0.39	0.17 / 0.50 / 0.25	0.16 / 0.27 / 0.41	0.16 / 0.16 / 0.09	0.20 / 0.15 / 0.14	0.06 / 0.06 / 0.06
Score Ordering	0.54 / 0.53 / 0.52	0.25 / 0.25 / 0.11	0.53 / 0.25 / 0.34	0.07 / 0.07 / 0.08	0.02 / 0.02 / 0.01	0.02 / 0.02 / 0.02
LatentScope	0.44 / 0.36 / 0.46	0.58 / 0.75 / 0.25	0.24 / 0.00 / 0.05	0.32 / 0.32 / 0.24	0.21 / 0.23 / 0.10	0.19 / 0.18 / 0.20
CULPRIT	0.68	0.92	0.72	0.41	0.27	0.34

Table 15: **Deranging the sensor-to-component association collapses Top-1 on all three benchmarks.** Deranged is a fixed-point-free permutation of the association, averaged over three draws on Causal Chamber and N-CMAPSS and ten on the air handlers. p is a paired Wilcoxon of each case's correct and draw-averaged Top-1.

		Top-1			Avg@5			
Dataset	Cases	Correct	Deranged	Drop	Correct	Deranged	Drop	p
Causal Chamber	28	0.39	0.04	0.35↓	0.68	0.14	0.54↓	3.8×10^{-5}
LBNL FDD	49	0.77	0.03	0.74↓	0.89	0.12	0.77↓	8.3×10^{-9}
N-CMAPSS	99	0.68	0.38	0.30↓	0.84	0.70	0.14↓	2.2×10^{-5}

per dataset. The set is tight where the ranking is strong. On SDAHU it has 2.2 components and reaches F1 0.70. On the wide systems it has 3.6 to 5.8 components, with F1 between 0.20 and 0.42. With one true root cause, a set of five caps F1 at 0.33 even when it contains the answer. The F1 column therefore mostly reports the size.

A cut with a guarantee exists in the conformal literature, where outlier p-values passed through Benjamini–Hochberg control the false discovery rate of the selected set (Bates et al., 2023). Hore & Ramdas (2026) also build finite-sample confidence sets for the stream whose distribution changes first. Neither transfers to our benchmark. On seven of the datasets in Table 17, the healthy prefix of a case yields four to six disjoint windows of the fault window's length. The smallest attainable conformal p-value is then 0.14 to 0.20, which leaves nothing to select at a significance level of 0.1. The temporal definition also needs staggered onsets across streams, whereas our cases share one. The route to a guarantee is a calibration bank of healthy windows gathered from the cases of one system, which trades the per-case guarantee for a per-dataset one.

G.7 Why the Baselines Fail

We find three weaknesses in the baseline field.

Table 16: Deranging the association for the baselines, Top-1 correct → deranged, averaged over the draws of Table 15. The last row is CULPRIT.

Method	Causal Chamber	SDAHU	DDAHU	N-CMAPSS
BARO	0.15 → 0.05	0.29 → 0.07	0.00 → 0.09	0.50 → 0.44
Cholesky	0.13 → 0.04	0.17 → 0.12	0.64 → 0.01	0.47 → 0.44
CIRCA	0.15 → 0.04	0.33 → 0.17	0.20 → 0.04	0.58 → 0.52
LatentScope	0.15 → 0.05	0.58 → 0.32	0.24 → 0.04	0.44 → 0.40
Score Ordering	0.09 → 0.04	0.25 → 0.06	0.53 → 0.03	0.54 → 0.45
CULPRIT	0.39 → 0.04	0.92 → 0.07	0.72 → 0.02	0.68 → 0.38

Table 17: The set cut of Equation 14 for CULPRIT. $|\mathcal{C}|$ counts the components, and $|S^\star|$ and $|\widehat{S}|$ are the mean sizes of the true and returned sets. F1 is the mean per-case F1 of $\widehat{S}$, and F1 all is the F1 of naming all components.

Dataset	Cases	$\lvert\mathcal{C}\rvert$	$\lvert S^\star\rvert$	$\lvert\widehat{S}\rvert$	F1	F1 all	Top-1
RCAEval	735	31	1.0	5.2	0.40	0.10	0.71
Causal Chamber	28	21	1.0	4.0	0.42	0.09	0.39
SDAHU	12	10	1.0	2.2	0.70	0.18	0.92
DDAHU	37	26	1.0	3.6	0.41	0.07	0.72
N-CMAPSS	99	5	2.3	1.0	0.44	0.58	0.68
SWaT	20	51	1.1	4.8	0.25	0.04	0.25
Bank	134	17	1.0	2.2	0.34	0.11	0.41
Market	148	70	1.0	5.8	0.25	0.03	0.34
Telecom	48	55	1.0	5.5	0.20	0.04	0.27

A maximum over sensors inherits the instrumentation. LatentScope scores each component as a latent variable and still reaches 0.00 Top-1 on Telecom pods and 0.11 on Market services. It keeps the maximum over a component's sensors and only reweights it by a pairwise regression between neighbours. To confirm that the bias lives in the sensor-pod-service hierarchy and not in the fault, we map each Market sensor to its own component. LatentScope's node-level Top-1 moves from 0.30 to 0.53 and its service-level Top-1 from 0.11 to 0.00. Two baselines cannot rank on this data. RCD assigns the same score to all components in all cases on N-CMAPSS, the Causal Chamber and both LBNL FDD air handlers. Traversal ties 9.25 of SDAHU's 10 components on average. The expected metric of Section 4.1 makes these ties visible where a first-listed tie-break would hide them.

A near-zero per-sensor spread inflates the steadiest sensors. AERCA reaches 0.10 to 0.15 Top-1 across seeds on SWaT, below all the statistical baselines. Its per-sensor exogenous z-score divides by a near-zero spread on sensors that barely move. It therefore ranks the lowest-variance healthy-prefix sensors at the top, with a median prefix standard deviation of 0.055 against 0.168 across all sensors. The true attack sensors, in contrast, sit at the high-variance end.

Onset faults and slow faults reward different windows. Cholesky reaches 0.64 Top-1 on DDAHU by diagnosing the first post-onset sample alone. A one-line predictor that reads only that sample reproduces Cholesky's top-ranked component in 37 of 37 cases. The same design collapses to 0.17 on SDAHU, where actuators sweep their full annual range. CULPRIT's velocity score reads the same dynamics signature within the window and scores 1.00 on all 15 DDAHU damper-stuck cases.

G.8 Cost, Robustness, and Sensitivity

This appendix collects the tables behind Section 4.4, i.e., runtime, the three perturbation sweeps, and the tolerance sweep. Each sweep varies one setting and fixes the rest at their default.

Efficiency. Table 18 reports wall-clock seconds per case for each method on each dataset under the compute allocation of Appendix E. The CULPRIT row times the configuration with the latent velocity form alone. The reported union adds one pass of first differences over the raw windows.

Robustness. CULPRIT runs in its reported configuration. Each column uses one perturbation draw with a fixed seed. RCD and CIRCA are absent by cost, at roughly 1000 s per RCAEval case (Section 4.4). Two variants of CULPRIT are flat under noise. The latent-velocity-only configuration reaches 0.71 to 0.73 Top-1 at all noise strengths, and the no-velocity configuration 0.73 to 0.74. This places the noise sensitivity in the per-sensor velocity form. Table 19 reports Gaussian noise, and Table 20 reports sensor dropout. The dropout draws are nested, hence a sensor removed at one drop rate stays removed at each higher rate.

BARO's RCAEval Top-1 rises from 0.23 unperturbed to 0.61–0.62 under noise (Table 19), and its zero-perturbation column reproduces the unperturbed run bitwise. Several RCAEval sensors are quantized and flat in the healthy window. Such a sensor likely gives BARO's IQR-based scoring a zero interquartile range and tied scores, which injected noise breaks.

Table 18: Runtime per failure case in seconds, fastest in bold, second fastest underlined. The deep learning baselines train at diagnosis time and sit outside the decoration.

Method	RCAEval		Causal Chamber		N-CMAPSS	
	Mean	Median	Mean	Median	Mean	Median
BARO	**0.013**	**0.007**	**0.00095**	**0.00085**	**0.014**	**0.012**
Cholesky	1.84	0.246	0.375	0.062	3.17	1.06
CIRCA	998	13.6	0.193	0.024	<u>0.114</u>	<u>0.107</u>
LatentScope	0.405	0.230	<u>0.036</u>	0.032	1.11	1.05
RCD	103	18.1	0.739	0.052	352	272
AERCA	31.4	15.4	9.06	3.71	22.6	9.45
StaR	7261	6115	175	137	1980	1845
RUN	8300	8527	311	86.6	1886	1949
CULPRIT	<u>0.018</u>	<u>0.011</u>	0.085	<u>0.003</u>	1.59	1.60

Table 19: Top-1 on RCAEval by Gaussian noise strength in units of each sensor's healthy spread.

Method	RCAEval				
	0	0.25	0.5	1	2
BARO	0.23	0.61	0.62	0.62	0.61
Cholesky	0.06	0.06	0.06	0.07	0.07
CULPRIT	0.71	0.64	0.62	0.61	0.59

The saturation sweep bends the sensor map on purpose. Each reading $x_{t,s}$ becomes $\bar{\mu}_s + (\bar{\sigma}_s/\gamma)\tanh\big(\gamma(x_{t,s} - \bar{\mu}_s)/\bar{\sigma}_s\big)$, with $\bar{\mu}_s$ and $\bar{\sigma}_s$ the sensor's healthy-window mean and standard deviation. At $\gamma = 0$ the map is the identity, its limit as $\gamma \to 0$. It compresses each deviation beyond about $\bar{\sigma}_s/\gamma$. The sweep therefore stresses the linearization around the healthy state. Conditions are untouched, and CULPRIT runs its reported configuration, since these datasets carry condition channels. Table 21 reports Top-1 by γ on the synthetic chain of Appendix D and the four physical datasets, and Section 4.4 discusses the curves. Cholesky draws a random permutation per case from a run-order-dependent stream, hence its $\gamma = 0$ column differs from the main tables by up to 0.02.

Parameter sensitivity. Table 22 sweeps the condition-match tolerance under pointwise matching and under case-level matching, which uses one reference per case near the fault window's condition support. Case-level matching is flat in the tolerance on all three datasets. The ridge of the structural fit is swept in Section 4.4. The separation-off cell at tolerance 0.10 is the last row of Table 11.

A partly wrong association. The derangement of Table 15 misassigns every sensor at once. A real document is more often right in most rows and wrong in a few. We therefore reassign a fraction of the sensors to a uniformly drawn wrong component and leave the rest correct. The draw is nested, i.e., a sensor wrong at one level stays wrong the same way at the levels above it, and each level averages ten draws. Table 23 sweeps one percent at a time to twenty percent of the system. Three of the four systems carry 9 to 23 mapped sensors, where one percent reaches no sensor at all. The table therefore prints the number of wrong sensors at each percentage, and a repeated count is a repeated

Table 20: Top-1 by sensor-drop rate (fraction of sensors removed, nested across rates), RCAEval.

Method	RCAEval			
	0	0.1	0.25	0.5
BARO	0.23	0.24	0.27	0.30
Cholesky	0.06	0.05	0.05	0.04
CULPRIT	0.71	0.67	0.63	0.52

Table 21: Top-1 by sensor saturation strength γ on the synthetic chain and the four physical datasets. The saturation applies tanh in units of each sensor's healthy spread, and $\gamma = 0$ is the identity.

Method	Synthetic					Causal Chamber					N-CMAPSS					SDAHU					DDAHU				
	0	0.25	0.5	1	2	0	0.25	0.5	1	2	0	0.25	0.5	1	2	0	0.25	0.5	1	2	0	0.25	0.5	1	2
BARO	0.44	0.46	0.32	0.17	0.15	0.15	0.13	0.08	0.08	0.09	0.50	0.53	0.53	0.51	0.48	0.29	0.54	0.58	0.58	0.50	0.00	0.00	0.00	0.00	0.00
Cholesky	0.23	0.25	0.25	0.26	0.24	0.13	0.12	0.15	0.10	0.06	0.49	0.46	0.54	0.54	0.48	0.17	0.17	0.17	0.25	0.46	0.65	0.62	0.64	0.59	0.45
CULPRIT	0.98	0.95	0.88	0.75	0.61	0.39	0.39	0.39	0.39	0.39	0.68	0.68	0.68	0.69	0.70	0.92	1.00	1.00	0.67	0.67	0.72	0.72	0.72	0.72	0.77

Table 22: Top-1 / Avg@5 as a function of condition-match tolerance for CULPRIT under pointwise and under case-level matching on the three condition-labelled datasets, separation held at its default. Tolerance 0.10 is the reported configuration.

Matching	Dataset	0.10 (default)	0.25	0.50	1.00
Pointwise	N-CMAPSS	0.68 / 0.84	0.68 / 0.80	0.65 / 0.77	0.52 / 0.73
Pointwise	SDAHU	0.92 / 0.98	0.92 / 0.98	0.92 / 0.98	0.67 / 0.67
Pointwise	DDAHU	0.72 / 0.85	0.73 / 0.85	0.74 / 0.85	0.73 / 0.85
Case-level	N-CMAPSS	0.36 / 0.67	0.36 / 0.66	0.36 / 0.67	0.35 / 0.66
Case-level	SDAHU	0.67 / 0.67	0.67 / 0.67	0.67 / 0.67	0.67 / 0.67
Case-level	DDAHU	0.80 / 0.89	0.77 / 0.85	0.77 / 0.85	0.77 / 0.85

state. Top-1 stays far above the deranged floor at all levels of the sweep. DDAHU is the one system with enough sensors for a curve at this resolution, and its Top-1 is 0.69 at one wrong sensor of 97 and 0.63 at five, against 0.72 correct and 0.02 deranged. On N-CMAPSS one or two wrong sensors of nine change Top-1 by at most 0.03, and the change is not significant. Each corrupted case records whether the draw misdocumented a sensor of its own root cause, and the effect follows that flag. On N-CMAPSS the cases whose root cause is untouched keep their accuracy in full. The projection predicts this, since a component is read through the sensors the document assigns to it. It also explains why the systems differ. For instance, a Causal Chamber root cause is assigned 6 of the 11 sensors, hence a wrong sensor there often touches the answer.

G.9 Fault Displacement off the Projection Direction

The projection keeps the part of a component's displacement that lies along w_c and sets aside the part orthogonal to it. This subsection states when a fault lies along w_c, measures the captured share of real faults, and evaluates a score for the orthogonal part.

Lemma G.1 (Displacement in the loading span)**.** *Let Assumption C.1 hold in both environments with the same W and residual noise of mean zero, and let $\mu_{c'} = \mathbb{E}_1[z_t^{(c')}] - \mathbb{E}_0[z_t^{(c')}]$ for each component c'. The mean displacement of the sensors of c is*

$$\mathbb{E}_1\big[r_t[\mathcal{S}_c]\big] - \mathbb{E}_0\big[r_t[\mathcal{S}_c]\big] = a_c\,\mu_c + \sum_{c' \neq c} W[\mathcal{S}_c, c']\,\mu_{c'},$$

which lies in the span of a_c and the loadings of the components sharing its sensors, and equals $a_c\mu_c$ when all sensors of c are local.

Proof. Take expectations of $r_t = W z_t + \nu_t$ on $\mathcal{S}_c$, where row s of W vanishes outside A_s. □

A fault on c, whether it changes $f^{(c)}$ or the law of $\zeta^{(c)}$, therefore displaces the mean of $r_t[\mathcal{S}_c]$ only within the span of a_c when all sensors of c are local. A downstream effect moves these sensors the same way, since it reaches them through $z^{(c)}$. Stage 1 removes the variance the operating condition drives before Σ_c is estimated, hence w_c follows the component's own fluctuation at a fixed operating point. The population direction equals the direction of a_c when the noise covariance on $\mathcal{S}_c$ is a multiple of the identity. Otherwise it lies within the angle of Proposition C.10. When all sensors of c are local, a displacement orthogonal to a_c comes from a biased sensor, a change of W between the windows, a component with more than one state variable, or the curvature of g_s. The efficiency and the flow capacity of a turbine module (Li, 2002) are two such state variables. The curvature

Table 23: **Top-1 under a partly wrong association, one percent at a time.** Each value averages ten nested draws that assign the row's percentage of sensors to a wrong component. The subscript counts the wrong sensors on that system. Deranged relabels the whole association (Table 15). † marks a paired Wilcoxon difference from the correct association at $p < 0.05$.

Misdocumented	Causal Chamber (11 sensors)	SDAHU (23 sensors)	DDAHU (97 sensors)	N-CMAPSS (9 sensors)
0%	$0.39_{(0)}$	$0.92_{(0)}$	$0.72_{(0)}$	$0.68_{(0)}$
1%	$0.39_{(0)}$	$0.92_{(0)}$	$0.69_{(1)}$	$0.68_{(0)}$
2%	$0.39_{(0)}$	$0.92_{(0)}$	$0.68_{(2)}$	$0.68_{(0)}$
3%	$0.39_{(0)}$	$0.96_{(1)}$	$0.68_{(3)}$	$0.68_{(0)}$
4%	$0.39_{(0)}$	$0.96_{(1)}$	$0.66_{(4)}$	$0.68_{(0)}$
5%	$0.34^{\dagger}_{(1)}$	$0.96_{(1)}$	$0.63^{\dagger}_{(5)}$	$0.68_{(0)}$
6%	$0.34^{\dagger}_{(1)}$	$0.96_{(1)}$	$0.60^{\dagger}_{(6)}$	$0.65_{(1)}$
7%	$0.34^{\dagger}_{(1)}$	$0.93_{(2)}$	$0.58^{\dagger}_{(7)}$	$0.65_{(1)}$
8%	$0.34^{\dagger}_{(1)}$	$0.93_{(2)}$	$0.56^{\dagger}_{(8)}$	$0.65_{(1)}$
9%	$0.34^{\dagger}_{(1)}$	$0.93_{(2)}$	$0.58^{\dagger}_{(9)}$	$0.65_{(1)}$
10%	$0.34^{\dagger}_{(1)}$	$0.93_{(2)}$	$0.55^{\dagger}_{(10)}$	$0.65_{(1)}$
11%	$0.34^{\dagger}_{(1)}$	$0.88^{\dagger}_{(3)}$	$0.57^{\dagger}_{(11)}$	$0.65_{(1)}$
12%	$0.34^{\dagger}_{(1)}$	$0.88^{\dagger}_{(3)}$	$0.56^{\dagger}_{(12)}$	$0.65_{(1)}$
13%	$0.34^{\dagger}_{(1)}$	$0.88^{\dagger}_{(3)}$	$0.56^{\dagger}_{(13)}$	$0.65_{(1)}$
14%	$0.30^{\dagger}_{(2)}$	$0.88^{\dagger}_{(3)}$	$0.54^{\dagger}_{(14)}$	$0.65_{(1)}$
15%	$0.30^{\dagger}_{(2)}$	$0.88^{\dagger}_{(3)}$	$0.54^{\dagger}_{(15)}$	$0.65_{(1)}$
16%	$0.30^{\dagger}_{(2)}$	$0.69^{\dagger}_{(4)}$	$0.53^{\dagger}_{(16)}$	$0.65_{(1)}$
17%	$0.30^{\dagger}_{(2)}$	$0.69^{\dagger}_{(4)}$	$0.53^{\dagger}_{(16)}$	$0.65_{(2)}$
18%	$0.30^{\dagger}_{(2)}$	$0.69^{\dagger}_{(4)}$	$0.53^{\dagger}_{(17)}$	$0.65_{(2)}$
19%	$0.30^{\dagger}_{(2)}$	$0.69^{\dagger}_{(4)}$	$0.51^{\dagger}_{(18)}$	$0.65_{(2)}$
20%	$0.30^{\dagger}_{(2)}$	$0.58^{\dagger}_{(5)}$	$0.51^{\dagger}_{(19)}$	$0.65_{(2)}$
Deranged	0.04	0.07	0.02	0.38

bends a large shift out of the tangent a_c by a term of second order in the shift. The captured share below is measured against w_c, hence a projection direction away from a_c lowers it as well, inside Assumption C.1.

Captured share of real faults. Table 24 reports, for the true root cause, the median share $(w_c^{\top}\bar{r}_c)^2/\|\bar{r}_c\|^2$ of its mean fault-window displacement $\bar{r}_c$ that w_c captures. It lists the same share for downstream components next to it, together with the share $1/|\mathcal{S}_c|$ that a uniformly random direction captures in expectation. The share is 0.71 on N-CMAPSS and 0.52 on SDAHU, more than twice the random level. Where the share is lower, the ranking holds because root shifts are large. The captured part alone leaves the root latent a median of 2.7 to 111 healthy standard deviations from its centre. A fault-informed direction helps only when an oracle names the root cause (Table 25). Projecting the sensors of each component on their own fault-window mean direction gives 0.51 Top-1, against 0.52 for w_c, since downstream components gain as much as the root.

A score for the residual subspace. The part w_c sets aside is the squared prediction error of process monitoring, $Q_c(t) = \|(I - w_c w_c^{\top})\, r_t[\mathcal{S}_c]\|^2$, which monitoring reads next to the principal score (MacGregor & Kourti, 1995; Joe Qin, 2003). We score it as a displacement, the fault-window mean of $|\sqrt{Q_c(t)} - m_c|/v_c$ with m_c and v_c the healthy mean and standard deviation of $\sqrt{Q_c(t)}$. We join it to Equation 10 by the same median-normalised maximum. A single biased sensor raises Q_c on its own. The screened score therefore takes the minimum of that displacement over the sensors j of the component after removing the reconstruction along $(I - w_c w_c^{\top})e_j$ (Dunia & Joe Qin, 1998). This screening returns a lone biased sensor to its healthy level and needs at least three sensors per component. A shared sensor carries the latents of the other components it is assigned to into the same subspace, hence the deflated score also projects out their projection directions.

Synthetic faults at a controlled angle. We extend the synthetic chain of Appendix D to three or six local sensors per component and give the root's fault a fixed norm at an angle θ to a_c. The part along a_c is a mechanism shift, and the orthogonal part displaces the root's sensors alone. A second family adds a bias of two or four healthy standard deviations to one sensor of another component. Table 26 shows CULPRIT at 0.99 Top-1 at $\theta = 0$ and at 0.97 or above at $\theta = 45°$, with three and with six sensors. The plain residual-subspace score recovers all orthogonal faults but ranks the biased sensor's component first in 0.97 and 0.78 of the cases at two standard deviations. The screened score matches CULPRIT at $\theta = 0$ and on the sensor faults at both bias sizes, and recovers 1.00 at $\theta = 90°$ with six sensors and 0.89 with three.

The benchmarks. Table 25 joins each score to the reported configuration. The deflated score raises RCAEval from 0.71 to 0.79. On LBNL FDD it can rank a downstream component first, e.g., the mixing box below a stuck cooling-box damper, as the fault drives that component into states outside the healthy window. The reported configuration therefore keeps the projection alone. The screened score suits systems such as RCAEval, whose faults stay within downstream components' healthy support.

Table 24: Median share of the mean fault-window displacement of a component's sensors that w_c captures, over components with at least two sensors. Downstream are the components the declared edges reach from the root cause, and Causal Chamber and RCAEval declare none. Random direction is the expected share $1/|\mathcal{S}_c|$. SWaT has one sensor per component.

Benchmark	Roots	Root	Downstream	Random direction
Chamber	11	0.12	–	0.12
SDAHU	12	0.52	0.45	0.20
DDAHU	37	0.20	0.36	0.14
N-CMAPSS	171	0.71	0.73	0.33
RCAEval	735	0.32	–	0.17
SWaT	–	–	–	–
OpenRCA	327	0.01	0.15	0.02

Table 25: Top-1 with the residual-subspace scores of Appendix G.9 joined to CULPRIT. The oracle root projects the root cause's sensors on their own fault-window mean direction. An arrow marks a paired Wilcoxon difference from CULPRIT at $p < 0.05$.

	Causal	LBNL FDD		N-CMAPSS	RCAEval	SWaT	OpenRCA			
Configuration	Chamber	SDAHU	DDAHU				Bank	Telecom	Market	Avg
CULPRIT	0.39	0.92	0.72	0.68	0.71	0.25	0.41	0.27	0.34	0.52
+ Q	0.39	0.58	$0.36^{\downarrow}$	0.69	0.69	0.25	$0.51^{\uparrow}$	0.27	$0.22^{\downarrow}$	0.44
+ screened	0.39	0.83	$0.45^{\downarrow}$	0.66	$0.79^{\uparrow}$	0.25	$0.60^{\uparrow}$	0.29	0.35	0.51
+ deflated	0.39	0.92	$0.53^{\downarrow}$	0.68	$0.79^{\uparrow}$	0.25	$0.60^{\uparrow}$	0.29	0.37	0.54
oracle root	$0.64^{\uparrow}$	0.92	0.82	$0.89^{\uparrow}$	$0.85^{\uparrow}$	0.25	$0.87^{\uparrow}$	$0.48^{\uparrow}$	$0.78^{\uparrow}$	0.72

Table 26: Synthetic chain with the scores of Appendix G.9, 20 seeds over 5 root causes per cell. The right block gives the share of cases that rank first the component whose sensor carries a bias of two or four healthy standard deviations.

	Top-1, fault at angle θ to a_c						Sensor fault ranked first			
	3 sensors			6 sensors			3 sensors		6 sensors	
Configuration	0°	45°	90°	0°	45°	90°	2σ	4σ	2σ	4σ
CULPRIT	0.99	0.97	0.49	0.99	0.98	0.34	0.01	0.82	0.00	0.04
+ Q	0.99	1.00	1.00	0.99	1.00	1.00	0.97	1.00	0.78	1.00
+ screened	0.99	0.98	0.89	0.99	0.99	1.00	0.01	0.82	0.00	0.04

H Diagnosis Artifacts

This appendix follows one diagnosis end to end through the artifacts an engineer checks. The case is N-CMAPSS engine DS02-006 unit 5, whose true root cause is the high-pressure turbine. The appendix then gives the parent-share formula and the call rule behind the evidence line of Stage 3. Figures 3, 4 and 5 draw the artifacts of the three stages on this engine.

Stage 1 artifact. Stage 1 emits a coverage report stating which fault samples have a matched reference and which sit at operating points the healthy window never visited. A score with no matched reference is therefore flagged instead of hidden.

Stage 2 artifact. Stage 2 emits the projection matrix $\hat{W}$ and each component's standardization constant s_c, which an engineer checks against the instrumentation drawings as one reads a loading matrix. Figure 4 draws it on the engine of this appendix.

Stage 3 artifacts. Stage 3 emits an indicator table, a separation subgraph and one evidence line per component. The table (Table 27) lists $\Delta(c)$, $\tilde{\Delta}(c)$, $V(c)$ and $\Psi(c)$ with the rank, the parents with their coefficients, and the *parent share* defined below. The subgraph (Figure 5) shades each component by its parent share. A downstream effect appears as a dark component downstream of a light one. The evidence line lists the component's sensors with their mean fault-window residual, flags those loading on another component, and closes with the call defined below.

Coverage report. The report summarizes the matched fault samples by the *condition coverage* and the *point coverage*,

$$\text{coverage} = \frac{|\{t \in \mathcal{T}_1 : \mathcal{M}_\tau(t) \neq \emptyset\}|}{|\mathcal{T}_1|}, \qquad \text{coverage}_{n_{\min}} = \frac{|\{t \in \mathcal{T}_1 : |\mathcal{M}_\tau(t)| \geq n_{\min}\}|}{|\mathcal{T}_1|}. \quad (15)$$

At $\text{coverage} = 1$ the case never leaves the region the healthy window covers, and $\text{coverage}_{n_{\min}}$ counts the fault samples with enough matched samples to standardize against. Figure 3 reads the report on this engine, where 141 of the 150 unmatched fault samples still have healthy neighbours in altitude and Mach alone. Coverage is therefore lost jointly across all four condition channels.

The trail. The coverage report opens the trail with point coverage 0.63 and a median of 36 matched healthy samples per fault sample. T48, the turbine-exit gas temperature and the textbook HPT health sensor, dominates the sensor evidence at $d_{\text{T48}} = 147.3$. Because T48 sits at the station between HPT and LPT, the aggregation renders the ambiguity as an exact tie {HPT, LPT} at 147.3 instead of resolving it silently. The projection reads T48 as HPT's whole direction and as a loading of 0.61 inside LPT's direction next to T50 and P50. The structural coefficient on the edge HPT→LPT is 0.96. Separation lifts HPT from a raw near tie at 2.80 against 2.59 to 5.25 against 2.30. Stage 3 closes the trail with the indicator table and its generated evidence lines (Table 27, drawn on the gas-path graph in Figure 5), and each phrase of those lines is filled from a cell of that table.

Parent share and the call rule. The parent share of a component in the indicator table is the fraction of its fault-window displacement that its parents' latents predict through Equation 7,

$$\min\Big(1,\ \sum_{t\in\mathcal{T}_1} \Big|\textstyle\sum_{c'\in \text{Pa}(c)} \beta_{cc'}\hat{z}_t^{(c')}\Big| \Big/ \sum_{t\in\mathcal{T}_1} \big|\hat{z}_t^{(c)}\big|\Big).$$

The generated evidence line closes with a call read off that number. The call is root cause for the top-ranked component. It is downstream effect of the parent when the parent share is at least one half, and own deviation with its rank otherwise.

I Scope and Applicability

CULPRIT reads a system model, i.e., a component list, a sensor-to-component association and optional edges (Section 2). This appendix names real systems whose association is fixed by design, the routes that obtain it elsewhere, and how a practitioner verifies it.

Systems with a fixed association. In many physical systems the association is written when the system is built and revised only at a retrofit, and four of our six benchmarks are of that kind.

Gas turbines. A turbofan's sensors sit at numbered gas-path stations, e.g., T24 at the low-pressure compressor outlet and T48 at the high-pressure turbine outlet. The numbering is published with the

Table 27: **Stage 3's indicator table, in rank order.** N-CMAPSS engine DS02-006 unit 5. $\Delta(c)$ and $\tilde{\Delta}(c)$ are in healthy-sigma units, and the parent share is the fraction of the fault-window displacement the parents predict. Below the table are the generated evidence lines for the root cause and its downstream neighbour.

Component	$\Delta(c)$	$\tilde{\Delta}(c)$	$V(c)$	$\Psi(c)$	Rank	Parent (β)	Parent share
HPT	2.80	5.25	1.24	2.66	1	HPC (0.86)	0.35
FAN	1.34	1.34	1.85	1.40	2	–	0.00
LPT	2.59	2.30	1.30	1.17	3	HPT (0.96)	1.00
LPC	1.27	1.97	1.51	1.14	4	FAN (0.97)	1.00
HPC	1.13	1.72	1.32	1.00	5	LPC (0.95)	1.00

```
HPT: separated displacement 5.25 sigma, velocity 1.24, score 2.66,
rank 1.  Deviating sensors (mean residual, sigma):  T48 +1.5.  T48
also loads on LPT. Parent HPC (beta 0.86) predicts 35% of its
fault-window displacement:  root cause.
LPT: separated displacement 2.30 sigma, velocity 1.30, score 1.17,
rank 3.  Deviating sensors (mean residual, sigma):  T50 +1.9, T48
+1.5, P50 +0.1.  T48 also loads on HPT. Parent HPT (beta 0.96)
predicts 100% of its fault-window displacement:  downstream effect of
HPT.
```

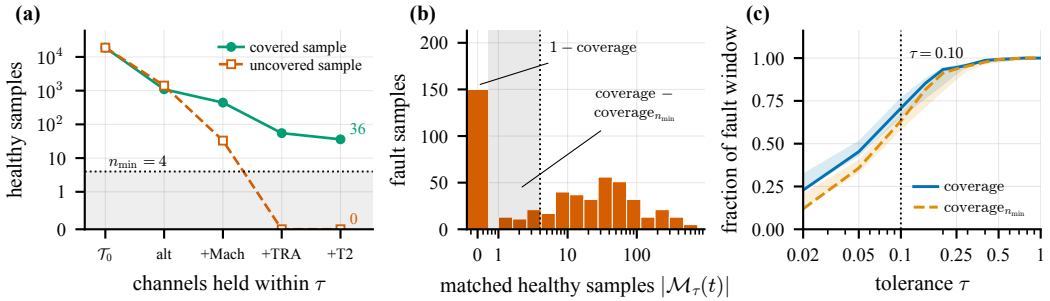


Figure 3: **Coverage is lost jointly, not per channel.** N-CMAPSS engine DS02-006 unit 5 at $\tau = 0.10$ and $n_{\min} = 4$. (a) Healthy samples left as each condition channel is held within τ in turn. (b) Matched-set sizes over the fault window. (c) Both coverages against τ, with min–max bands over the six development engines.

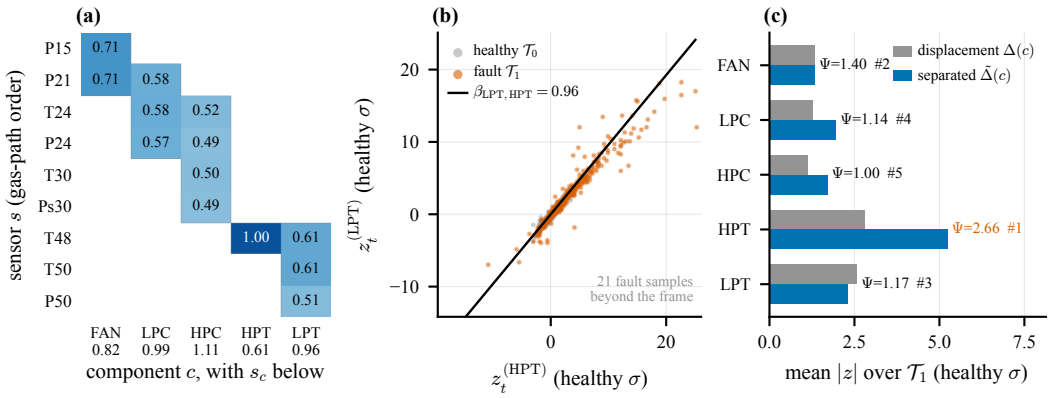


Figure 4: **Structural separation breaks the HPT-LPT tie.** N-CMAPSS engine DS02-006 unit 5. (a) The projection matrix $\hat{W}$ of Equation 6, where T48, HPT's only sensor, is shared with LPT. (b) The fault window follows the healthy fit of the structural coefficient of Equation 7 on the edge HPT to LPT. (c) Displacement, separated displacement and combined score $\Psi(c)$ with rank. HPT's healthy regression predicts LPT's displacement, hence the separated displacement ranks HPT first.

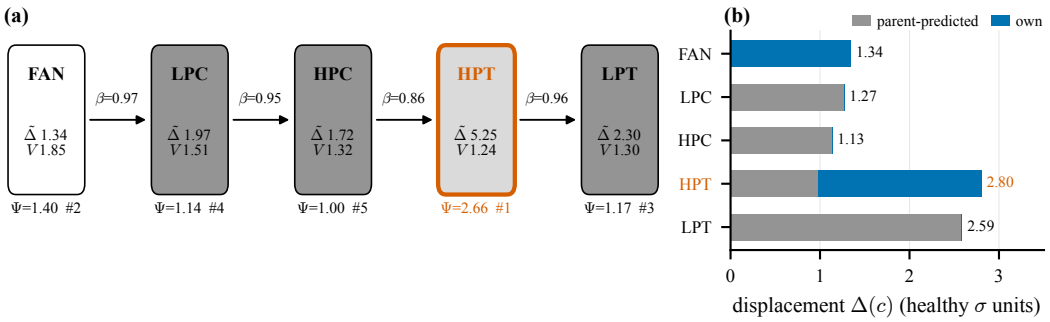


Figure 5: **Parent shares on one engine.** N-CMAPSS engine DS02-006 unit 5. (a) The gas-path edges with their coefficients β, the top-ranked component outlined in orange and the others shaded by their parent share. (b) Each component's displacement split into the part its parents predict and its own remainder. HPC predicts 35% of HPT's displacement, while HPT predicts all of LPT's.

engine model (Arias Chao et al., 2021). The station a sensor occupies is set by the thermodynamic layout of the engine, hence the association is fixed for the life of the engine type.

Buildings. An air handler ships a schema file, built from the mechanical drawing, that names which points belong to which equipment. The LBNL FDD datasets carry such a file for each system (Granderson et al., 2023). The file lists each component, its points, and the order in which air flows through them. These give the association and edges CULPRIT reads.

Water and process plants. A process plant documents its instruments on piping and instrumentation diagrams, and the tag of an instrument names the process stage it serves. In the SWaT testbed the first digit of each channel name is its stage (Goh et al., 2016), and Steude et al. (2026) derive the stage membership automatically from that convention.

Laboratory apparatus. The Causal Chamber wind tunnel is a fixed apparatus with published blueprints, and its graph is validated by randomized experiments (Gamella et al., 2025b). Deranging that association (Table 15) is a controlled counterfactual against a ground-truth document.

Independent work on diagnosing cyber-physical systems takes the same two inputs, a subsystem-signals map and a subsystem graph, and obtains them from documentation (Steude et al., 2026).

Obtaining the association from a simulator. Where the document is missing or incomplete, a simulator of the system yields the association by intervention. A thermodynamic engine model carries state variables per component, e.g., an efficiency and a flow modifier for each of the fan, the compressors and the turbines in C-MAPSS. The simulator produces degraded behaviour by moving them (Arias Chao et al., 2021). Moving one component's modifier and recording which sensors respond gives that component's column of A. Open gas-turbine models such as T-MATS and pyCycle carry a flow scalar and an efficiency scalar for each compressor and turbine (Chapman et al., 2014; Hendricks & Gray, 2019). A software system has the same route through fault injection in a preproduction environment (Basiri et al., 2016). RCAEval collected its cases by injecting a fault into one running service and recording the telemetry (Pham et al., 2025). The sensors that respond to a fault injected at a component are that component's sensors. In both cases the association is read from interventions on a copy of the system, hence it can be refreshed whenever the system changes.

Software systems. Two of our benchmarks are microservice deployments, where services are created and retired within days. On RCAEval each metric name carries the service it belongs to. The association is therefore one sensor to one service and needs no document. A log template or trace span of the RE2 and RE3 datasets (Pham et al., 2025) enters Stage 2 the same way, as a sensor of its service. On OpenRCA the association between hosts, pods and services comes from the deployment inventory shipped with the data, and the edges come from that inventory and the service call traces.

Verifying the association. Before a diagnosis, the graphical conditions of Appendix C determine which components the association resolves. After a diagnosis, the evidence artifact of Appendix H shows which sensors carried each component's displacement, exposing any sensor that loads on a component it is not assigned to.