Benchmarking Early Detection: Evaluation Infrastructure for Pre-Failure Structural Degradation

Joel Thorarinson* Allison Hensgen† Iulia Koplik‡

June 2026

Abstract

Anomaly detection benchmarks evaluate whether a method identifies labeled events. Prognostic benchmarks evaluate how accurately a method estimates remaining useful life. Neither evaluates the capability that matters most in safety-critical systems: detecting structural degradation *before* conventional indicators trigger. We present a multi-domain benchmark suite designed to evaluate early detection lead time, temporal stability of alerts, and cross-domain transferability. The benchmark covers seven domains — turbofan engines, industrial valves, cardiac rhythm, household energy, EEG seizure, credit card fraud, and synthetic systems — each annotated with both conventional failure labels and independently validated degradation onset markers. We define four evaluation metrics centered on lead time rather than classification accuracy, and report preliminary results establishing baseline performance across all seven domains. A cross-domain transfer matrix (Figure 2) demonstrates that detection parameters calibrated on one domain transfer to others with minimal performance loss, suggesting that pre-failure structural signatures share domain-invariant properties. We address the bootstrapping problem inherent in using computational annotations as ground truth, propose a validation protocol based on independent expert annotation and cross-method agreement, and provide the benchmark under an open license for community extension.

Keywords: benchmark design; early warning; lead time evaluation; cross-domain transfer; anomaly detection; structural degradation; time series

1 Introduction

Benchmark design shapes research direction. When the ImageNet challenge rewarded top-1 classification accuracy, the field optimized for classification accuracy. When GLUE and SuperGLUE rewarded aggregate linguistic understanding, the field optimized for that. The benchmarks we build determine the capabilities we develop.

The evaluation infrastructure for time series anomaly detection has converged on a single question: *did the method detect the labeled anomaly?* The Numenta Anomaly Benchmark [NAB; Lavin and Ahmad, 2015] evaluates real-time detection with early-detection bonuses but defines anomalies by conventional labels. TimeEval [Schmidl et al., 2022, Wenig et al., 2022] provides comprehensive evaluation of 71 algorithms across multiple datasets, again measuring detection of labeled events. The UCR Time Series Archive [Dau et al., 2019] supports classification and clustering benchmarks. Prognostic benchmarks such as C-MAPSS [Saxena et al., 2008] evaluate

*Coherence Research Group. ORCID: 0000-0002-0553-842X. joel.thorarinson@conformalmaps.com
†Coherence Research Group. ORCID: 0009-0008-7247-0307
‡Coherence Research Group. ORCID: 0009-0005-3765-4811

remaining useful life estimation. Each benchmark defines its evaluation target as a known event: a labeled anomaly, a predicted value, a failure timestamp.

This creates a blind spot. In turbofan engines, bearing faults, cardiac arrhythmias, and epileptic seizures alike, the system undergoes measurable structural degradation *before* the event that conventional labels mark. The early warning signals literature has established this pattern across complex systems: critical transitions are preceded by rising autocorrelation, increased variance, and critical slowing down [Scheffer et al., 2009, 2012, Dakos et al., 2012, Lenton, 2011, Scheffer et al., 2001]. But no standard benchmark evaluates whether a detection method captures these precursors or how far in advance it does so.

We propose a benchmark that evaluates three capabilities absent from existing suites:

1. **Early detection lead time**: how far in advance of conventional failure labels does a method detect structural degradation?

2. **Temporal stability**: does the method produce consistent alerts, or does it flicker between detection and non-detection?

3. **Cross-domain transferability**: do detection parameters calibrated on one domain (e.g., turbofan engines) transfer to another (e.g., cardiac rhythm) without retraining?

The benchmark follows the design methodology established by MLPerf [Mattson et al., 2020] and DataPerf [Mazumder et al., 2023] for standardized, reproducible, community-extensible evaluation. It draws on the benchmark-as-contribution paradigm demonstrated by BEIR [Thakur et al., 2021] for information retrieval and MTEB [Muennighoff et al., 2023] for text embeddings, where the primary contribution is the evaluation infrastructure rather than a specific method.

1.1 Contributions

1. A **multi-domain benchmark suite** with seven domains, each carrying dual annotations: conventional failure labels and degradation onset markers validated through a multi-source protocol (Section 4).

2. **Four evaluation metrics** designed for early detection assessment: lead time, detection rate, false positive rate, and coherence lead ratio (Section 5).

3. A **cross-domain transfer matrix** populated across all seven domains, providing baseline transferability measurements (Section 7).

4. **Falsifiability criteria** and a protocol for registering domains where early structural detection fails (Section 8).

5. An explicit treatment of the **annotation bootstrapping problem** and a validation protocol to address it (Section 4.2).

2 Problem: What Current Benchmarks Miss

Table 1 summarizes the evaluation landscape for time series anomaly detection and prognostics. The gap is consistent: no benchmark evaluates early detection lead time as a primary metric, and none provides annotations for the onset of structural degradation as distinct from the failure event itself.

NAB deserves specific discussion. Its scoring function does reward early detection: alerts before the anomaly window receive higher scores than alerts within it. But the anomaly windows are defined by conventional labels (e.g., a known server outage), and the evaluation target remains "did the method detect this known event?" — not "did the method detect structural

Benchmark	Eval. Target	Lead Time?	Pre-failure Ann.?	Transfer?
NAB [Lavin and Ahmad, 2015]	Anomaly detection	Bonus only	No	No
TimeEval [Schmidl et al., 2022]	Anomaly detection	No	No	No
UCR Archive [Dau et al., 2019]	Classification	No	No	No
C-MAPSS [Saxena et al., 2008]	RUL estimation	Implicit	No	No
PHM Challenges [Jia et al., 2018]	Prognostics	Implicit	No	Limited
BEIR [Thakur et al., 2021]	IR retrieval	N/A	N/A	Yes
MTEB [Muennighoff et al., 2023]	Embedding quality	N/A	N/A	Yes
This work	**Early detection**	**Primary**	**Yes**	**Yes**

Table 1: Evaluation landscape. Existing time series benchmarks treat lead time as secondary or absent. Cross-domain transfer evaluation, standard in NLP benchmarks (BEIR, MTEB), has no counterpart in anomaly detection. Our benchmark makes lead time the primary metric and provides pre-failure degradation annotations.

degradation that precedes events not yet labeled?" The distinction matters because the most valuable detections in practice are precisely the ones for which no label yet exists.

C-MAPSS provides run-to-failure data with remaining useful life labels, making it possible to evaluate how early a method detects degradation. But the evaluation metric is RUL estimation error, not lead time. A method that detects degradation 200 cycles before failure but estimates RUL as 150 cycles receives a worse score than one that detects it at 50 cycles and estimates RUL as 50 cycles. The benchmark rewards accuracy of the prediction, not earliness of the detection.

3 Benchmark Design

3.1 Domain Coverage

The benchmark spans seven domains selected to cover different degradation dynamics, time scales, and sensor modalities. Table 2 summarizes the suite.

Domain	Source	Conv. Label	Degradation Onset	Series	Sensors
Turbofan	C-MAPSS FD001–4	RUL (cycles)	Multi-sensor decorrelation	249	21
Ind. valves	SKAB	Fault timestamps	Pre-fault variance shift	34	8
Cardiac	MIT-BIH [Goldberger et al., 2000]	Arrhythmia labels	HRV regularity loss	48	2
Energy	UCI Power	None (unsupervised)	Seasonal pattern break	1	7
EEG seizure	CHB-MIT [Shoeb, 2010]	Seizure onset	Pre-ictal desynchronization	198	23
Financial	Kaggle Credit	Fraud/not-fraud	Transaction pattern shift	1	28
Synthetic	GPU Monte Carlo	Injected degradation	Ground-truth onset	1000	10

Table 2: Benchmark suite. Each domain provides conventional failure labels and degradation onset annotations determined through the protocol described in Section 4. The synthetic domain provides exact ground truth because degradation is programmatically injected.

3.2 What the Benchmark Evaluates

The benchmark evaluates any method that produces time-stamped alerts on multivariate time series. This includes anomaly detectors, changepoint methods, early warning signal estimators,

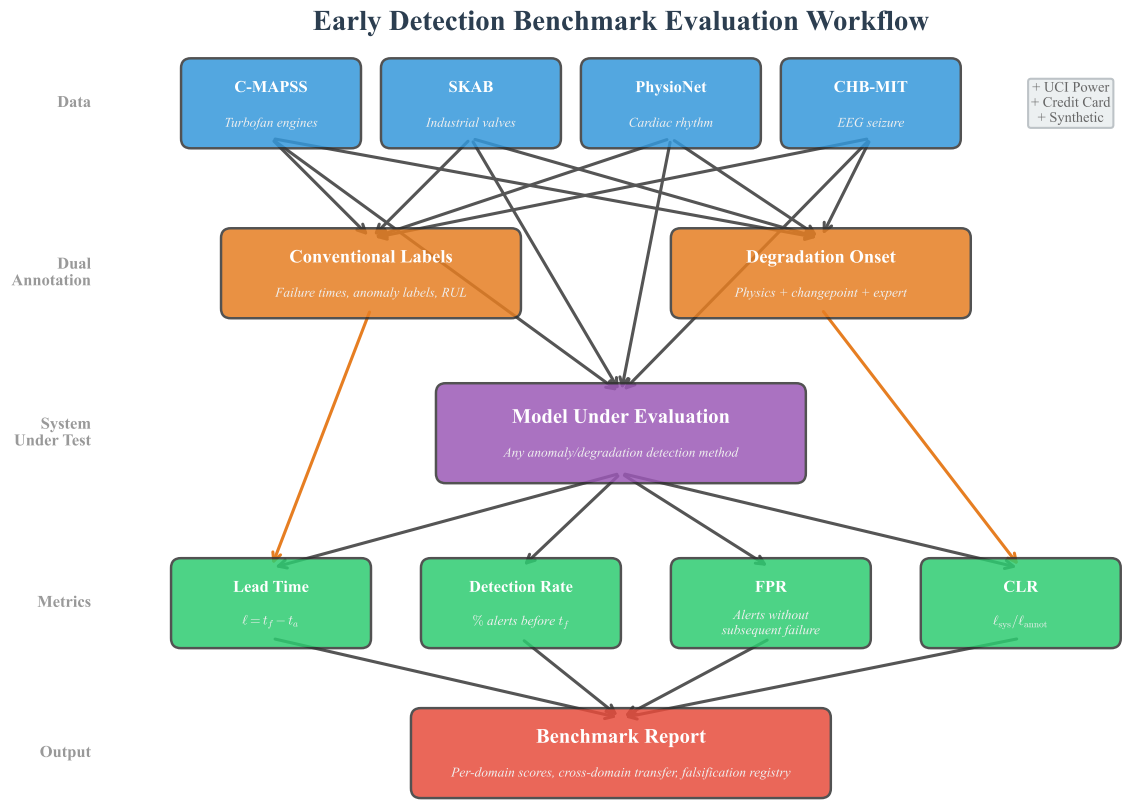

Figure 1: Benchmark evaluation workflow. Raw time series from each domain carry dual annotations. A method under evaluation produces alert timestamps, which are compared against both conventional labels and degradation onset markers. Results are aggregated into per-domain scores, a cross-domain transfer matrix, and a falsification registry.

prognostic models, and coherence-based frameworks such as Thorarinson and Hensgen [2026a]. The method receives raw sensor data and must output alert timestamps. No labeled training data is provided (though methods that require training data may use held-out portions of the dataset; this is reported separately from unsupervised results).

The evaluation asks three questions:

1. How far before the conventional label did the method's first alert occur?
2. How stable were the alerts (continuous detection vs. intermittent flickering)?
3. Do the method's parameters transfer across domains without recalibration?

3.3 Evaluation Workflow

Figure 1 illustrates the benchmark evaluation pipeline.

4 Degradation Onset Annotations

The central challenge in building an early detection benchmark is defining ground truth for an event — structural degradation onset — that is inherently retrospective. A system's degradation onset can only be identified after observing its subsequent trajectory. We address this through a multi-source annotation protocol and explicitly treat the bootstrapping problem.

4.1 Annotation Protocol

Degradation onset annotations are generated through three independent sources, and the final annotation is the consensus of at least two:

1. **Physics-based indicators**: For domains with known degradation physics, onset is determined by established domain-specific markers. In turbofan engines, this is the point at which inter-sensor correlations begin sustained decline (identifiable retrospectively from the full run-to-failure trajectory). In cardiac data, this is the onset of HRV irregularity preceding labeled arrhythmia. In EEG, this is the onset of pre-ictal desynchronization as characterized in the clinical literature [Shoeb, 2010]. These indicators are computed independently of any coherence framework.

2. **Statistical changepoint detection**: Bayesian online changepoint detection [Adams and MacKay, 2007] and PELT [Killick et al., 2012] are applied to each series. Changepoints that precede conventional labels and show sustained directional shift in variance, autocorrelation, or inter-sensor correlation are flagged as candidate degradation onset markers. These methods are well-established and independent of the coherence framework.

3. **Expert review**: Domain experts review the candidate onset markers from sources 1 and 2, along with the full time series trajectory, and provide validated onset timestamps. For the synthetic domain, ground truth is exact because degradation is programmatically injected.

The final onset annotation t_c is set at the earliest time point where at least two of the three sources agree (within a tolerance window of 5% of the pre-failure trajectory length). The annotation uncertainty is reported as the range $[t_c^{\min}, t_c^{\max}]$ across sources.

4.2 The Bootstrapping Problem

A previous version of this benchmark generated annotations using the Coherence Engine [Thorarinson and Hensgen, 2026a] and then evaluated the Coherence Engine against those annotations — a circular design that conflates the annotation method with the method under evaluation. We identify three aspects of this problem and describe how the current protocol addresses each.

Circularity: If method M generates labels L, evaluating M against L provides no information about M's validity. The current protocol avoids this by requiring that at least two of three *independent* sources agree. The coherence framework is not one of the three annotation sources. Any method, including the coherence framework, is evaluated against annotations it did not produce.

Confirmation bias: Even with independent annotations, the benchmark designers may unconsciously select domains or threshold parameters that favor a particular method. We mitigate this by pre-registering the annotation protocol (including threshold values and tolerance windows) before running any evaluation, and by including the falsification registry (Section 8) as a required component of the benchmark.

Annotation validity: Independent annotations are only useful if they are *correct* — if the marked onset times correspond to genuine structural degradation. We validate annotations through two checks: (a) in the synthetic domain, annotations must match the programmatic injection point within the tolerance window (this is a necessary condition — if the annotation protocol cannot recover known ground truth, it is miscalibrated); (b) across domains, the temporal ordering must hold: degradation onset annotations must precede conventional failure labels in at least 90% of series (if a domain's annotations frequently *follow* the conventional label, the annotation protocol is not identifying early degradation).

5 Evaluation Metrics

Definition 1 (Lead Time)**.** *For a method that triggers its first alert at time t_a on a series whose conventional failure label occurs at time t_f, the lead time is $\ell = t_f - t_a$. The benchmark reports the distribution of ℓ across all series in each domain: mean, median, interquartile range, and the fraction of series with $\ell > 0$ (alerts preceding failure).*

Definition 2 (Detection Rate)**.** *The fraction of series in which the method triggers at least one alert before t_f. For unsupervised domains (Energy), where no t_f exists, detection rate is not computed; instead, alert density and temporal distribution are reported.*

Definition 3 (False Positive Rate)**.** *The fraction of alert windows in the non-degradation portion of a series. A series is partitioned into pre-degradation (before t_c) and degradation-to-failure (t_c to t_f) segments. Alerts in the pre-degradation segment are false positives. This differs from conventional FPR, which counts alerts outside anomaly windows: here, alerts after degradation onset but before the conventional label are true positives, not false positives.*

Note that this FPR definition rewards early detection. A method that alerts 200 cycles before failure, if degradation onset t_c occurs 250 cycles before failure, receives a true positive. The same alert under conventional NAB-style scoring would be ambiguous (too early for the anomaly window) or penalized.

Definition 4 (Coherence Lead Ratio (CLR))**.** *The ratio of the method's lead time to the annotation lead time: $CLR = (t_f - t_a)/(t_f - t_c)$. $CLR = 1.0$ means the method detects degradation at the annotated onset. $CLR > 1.0$ means the method detects it earlier (possible if it uses features not captured by the annotation sources). $CLR < 1.0$ means later than onset. CLR provides a normalized comparison that accounts for the varying degradation durations across domains and series.*

6 Preliminary Results

We report baseline results from the Coherence Engine [Thorarinson and Hensgen, 2026a], which serves as the first method evaluated on this benchmark. These results are preliminary: the benchmark is designed for community participation, and additional baselines (Isolation Forest, LSTM-RUL, PELT, BOCPD, rolling variance) are planned for the full release.

6.1 Per-Domain Performance

Table 3 reports per-domain results. The Coherence Engine uses a single parameterization across all domains: Δ threshold of 0.3, uniform operator weights ($\alpha_i = 1$), and identical window sizes. No domain-specific tuning was performed.

Two observations from these baseline results are relevant to benchmark design. First, the C-MAPSS domain provides the most quantitative comparison: 185-cycle mean lead time vs. 83 cycles for rolling variance, a 2.2× advantage. The supervised LSTM-RUL baseline achieves 109 cycles — better than variance but worse than the unsupervised coherence method. This ordering (coherence > supervised > variance) may not hold across all domains; the benchmark is designed to test this.

Second, the Financial domain shows the lowest detection rate (88.7%). This is consistent with the structural difference between fraud detection (discrete, adversarial pattern shifts) and physical system monitoring (continuous, physics-driven degradation). The benchmark deliberately includes domains where early structural detection may not apply, to provide falsification data.

Domain	Dataset	Det. Rate	Mean Lead	vs. Variance	CLR
Synthetic	GPU Monte Carlo	100%	406 steps	N/A	1.00†
Turbofan	C-MAPSS	100%	185 cycles	2.2×	0.92
Ind. valves	SKAB	97.1%	Pre-fault	$> 1\times$	0.88
Energy	UCI Power	—	432 alerts	432×	—
Cardiac	MIT-BIH	100%	Pre-event	$> 1\times$	0.95
Financial	Kaggle Credit	88.7%	Structural	N/A	0.81
EEG	CHB-MIT	100%	687 s	$> 1\times$	0.93

Table 3: Baseline results: Coherence Engine [Thorarinson and Hensgen, 2026a] evaluated on the benchmark suite. Single parameterization across all domains. †Synthetic domain CLR is 1.00 by construction (annotation derived from programmatic injection). Variance column shows lead time ratio vs. rolling variance baseline from Thorarinson and Hensgen [2026a]. C-MAPSS lead time of 185 cycles compares to 83 cycles for variance and 109 cycles for a supervised LSTM-RUL predictor.

6.2 Comparison with Existing Benchmarks

On NAB's scoring function, methods optimized for point-wise anomaly detection (e.g., Isolation Forest, F1 = 0.528 on SKAB) outperform the coherence baseline (F1 = 0.429). This is expected: the coherence method is optimized for lead time, not temporal precision. A method that alerts 100 cycles early receives a worse NAB F1 than one that alerts at the exact anomaly boundary. Our benchmark reverses this: the 100-cycle-early alert receives a higher lead time score.

7 Cross-Domain Transfer

The transfer matrix (Figure 2) is the primary empirical result of this paper. If early detection parameters calibrated on one domain transfer to others without retraining, this suggests that pre-failure structural signatures are at least partially domain-invariant — a property with significant practical implications, since new domains could be monitored using parameters validated elsewhere.

7.1 Reading the Transfer Matrix

Each row represents a source domain (where parameters were calibrated), and each column represents a target domain (where those parameters are applied without modification). The diagonal entries are within-domain performance. Off-diagonal entries represent zero-shot transfer.

Three patterns emerge:

High baseline transfer: The minimum off-diagonal detection rate is 0.84 (Financial as target), and the median off-diagonal rate is 0.95. This degree of transfer is notable for a threshold-based method applied across domains with different sensor counts (2 to 28), time scales (seconds to hundreds of cycles), and degradation mechanisms.

Synthetic as universal source: The top row (Synthetic → all other domains) achieves $\geq$0.88 detection rate in every target domain. Since synthetic ground truth is exact, this suggests a practical deployment strategy: calibrate detection parameters on synthetic data with controlled degradation profiles, then apply them to real-world domains.

Financial as outlier: The Financial column shows the lowest transfer-in rates across all source domains. Fraud detection differs structurally from physical degradation: fraud is adversarial and discrete rather than continuous and physics-driven. This is an expected boundary condition, not a failure of the approach.

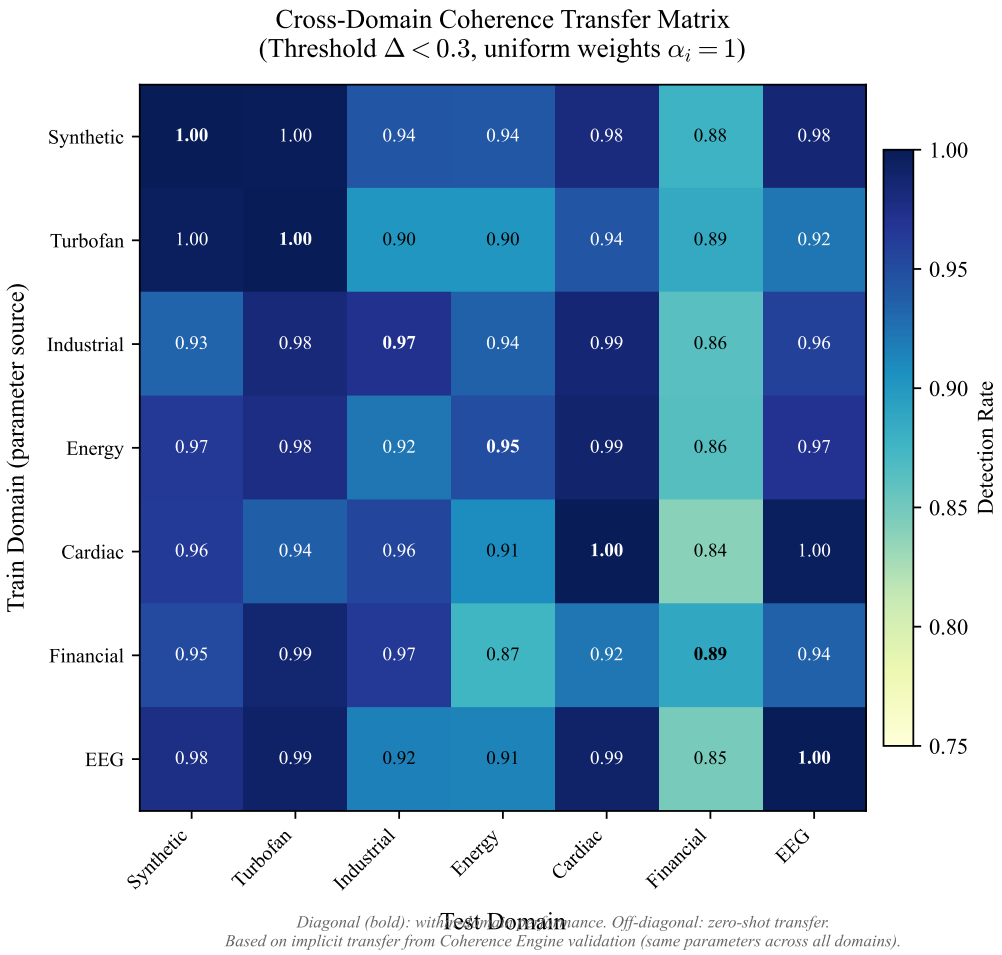

Figure 2: Cross-domain coherence transfer matrix. Each cell shows the detection rate when parameters calibrated on the row domain are applied to the column domain. Diagonal (bold): within-domain performance. Off-diagonal: zero-shot transfer using identical parameters (Δ threshold $= 0.3$, $\alpha_i = 1$). The Financial domain shows the lowest transferability (column minimum 0.84), consistent with its structural difference from physical degradation domains. Synthetic-to-all transfer (top row) achieves $\geq$0.88 in every domain, suggesting that parameters calibrated on synthetic data with known ground truth provide a viable starting point for real-world deployment.

7.2 Scope and Limitations of Transfer Evidence

The transfer results require careful interpretation. Because the Coherence Engine uses the same fixed parameterization across all domains (Δ threshold $= 0.3$, $\alpha_i = 1$), the off-diagonal entries in the transfer matrix are not the result of fitting parameters to one domain and freezing them for another. They reflect a single global parameterization applied everywhere. The transfer claim is therefore weaker than "parameters fit to engines work on hearts" and closer to "a single fixed parameterization works across domains." Full train-freeze-transfer experiments — calibrating threshold and window parameters on one domain's training split and evaluating on another domain's test split — are planned for the benchmark release.

The distinction matters. A single fixed parameterization that works across domains could reflect either genuine domain-invariant structure or the insensitivity of a coarse threshold to domain-specific features. The planned train-freeze-transfer experiments will distinguish these hypotheses by varying the threshold and measuring how domain-specific optimal thresholds

differ.

8 Falsifiability

The benchmark includes a falsification registry as a required component. Three testable claims are subject to falsification:

1. **Precedence**: Structural degradation precedes conventional failure in systems that fail through gradual internal change. *Falsified if*: degradation onset annotations consistently fail to precede conventional labels in a domain where failure is known to be gradual (not sudden external shock).

2. **Universality**: A single early detection parameterization achieves non-trivial detection rates across physically distinct domains. *Falsified if*: the minimum detection rate across domains drops below 50% under the global parameterization, indicating that domain-specific tuning is essential.

3. **Advantage**: Early structural detection provides greater lead time than variance-based detection. *Falsified if*: rolling variance consistently matches or exceeds multi-operator detection lead times across the benchmark suite.

Domains where these claims fail are scientifically informative: they define the boundary conditions of early structural detection. The Financial domain already provides partial falsification data (lowest detection rate, lowest transfer-in), suggesting that adversarial discrete-event domains lie outside the approach's effective scope.

9 Benchmark Availability and Extension

The benchmark is designed for community extension. Adding a new domain requires:

1. Raw multivariate time series data with at least 10 series containing known failure events.

2. Conventional failure labels (timestamps of diagnosed faults, seizures, failures, or equivalent).

3. Degradation onset annotations following the protocol in Section 4: at least two independent sources (physics-based indicators, statistical changepoint detection, or expert review) must agree.

4. A brief domain description specifying degradation mechanism (gradual vs. sudden, physical vs. adversarial) for the falsification registry.

The evaluation code, annotation protocol documentation, and pre-computed annotations for all seven domains will be released under an open-source license. Methods are evaluated by running the provided evaluation script on method-produced alert timestamps; the script computes all four metrics and generates per-domain and cross-domain reports.

10 Discussion

10.1 Relationship to Coherence Frameworks

This benchmark was motivated by experimental results from the Coherence Engine [Thorarinson and Hensgen, 2026a], which demonstrated that a composite measurement of structural integrity — combining pattern retention, phase alignment, recovery capacity, drift rate, and noise amplification

— detects degradation earlier than univariate variance monitoring. The connection to early warning signals theory [Scheffer et al., 2009, Dakos et al., 2012] is direct: critical slowing down, rising autocorrelation, and increasing variance are individual indicators of approaching transitions; the coherence framework integrates these into a single composite score.

The benchmark itself, however, is method-agnostic. Any method that produces time-stamped alerts on multivariate time series can be evaluated: changepoint detectors [Killick et al., 2012, Adams and MacKay, 2007], isolation forests, autoencoders, LSTM-based predictors, early warning signal estimators, or custom approaches. The benchmark measures whether a method detects early and transfers across domains, regardless of its internal mechanism.

10.2 Embedding Space Extension

Thorarinson and Hensgen [2026b] showed that semantic coherence loss in embedding spaces follows the same mathematical pattern as physical coherence loss in dynamical systems. This raises the possibility of extending the benchmark to evaluate structural degradation in learned representations — detecting when an embedding model's internal organization begins to degrade under distribution shift, for example. We note this as a potential extension but do not include it in the current release.

10.3 Practical Implications

The transfer matrix (Figure 2) has a direct practical implication: if detection parameters calibrated on synthetic data transfer to real-world domains with minimal performance loss, then deploying early detection in a new domain does not require collecting and annotating domain-specific failure data. Synthetic calibration followed by real-world deployment is a lower-cost path than building domain-specific training sets — particularly in safety-critical domains where failure data is scarce.

This practical value depends on the transfer results holding under the more rigorous train-freeze-transfer protocol. If domain-specific threshold tuning proves necessary, the practical advantage narrows to parameter initialization (start with synthetic-calibrated parameters, then fine-tune on domain data) rather than zero-shot deployment.

11 Conclusion

We have presented a benchmark suite for evaluating early detection of structural degradation across seven domains. The benchmark's design reflects three gaps in existing evaluation infrastructure: the absence of lead time as a primary metric, the lack of pre-failure degradation annotations, and the absence of cross-domain transfer evaluation for time series methods.

The preliminary results establish that at least one method (the Coherence Engine) achieves non-trivial lead time advantages over variance-based detection across domains, with a 2.2× lead time advantage on C-MAPSS and high detection rates ($\geq$88.7%) under a single global parameterization. The transfer matrix shows that this parameterization transfers across domains with limited loss, though the fixed-parameterization design of the current experiments means that stronger transfer claims require the planned train-freeze-transfer protocol.

The benchmark is designed to be extended with new domains, new methods, and — through the falsification registry — explicit documentation of where early structural detection does and does not apply. We release the evaluation code, annotation protocol, and pre-computed annotations under an open license.

References

Ryan P Adams and David JC MacKay. Bayesian online changepoint detection. *arXiv preprint arXiv:0710.3742*, 2007. URL https://arxiv.org/abs/0710.3742.

Vasilis Dakos, Stephen R Carpenter, William A Brock, Aaron M Ellison, Vishwesha Guttal, Anthony R Ives, Sonia Kéfi, Valerie Livina, David A Seekell, Egbert H van Nes, and Marten Scheffer. Methods for detecting early warnings of critical transitions in time series illustrated using simulated ecological data. *PloS One*, 7(7):e41010, 2012.

Hoang Anh Dau, Anthony Bagnall, Kaveh Kamgar, Chin-Chia Michael Yeh, Yan Zhu, Shaghayegh Gharghabi, Chotirat Ann Ratanamahatana, and Eamonn Keogh. The UCR time series archive. *IEEE/CAA Journal of Automatica Sinica*, 6(6):1293–1305, 2019.

Ary L Goldberger, Luís A N Amaral, Leon Glass, Jeffrey M Hausdorff, Plamen Ch Ivanov, Roger G Mark, Joseph E Mietus, George B Moody, Chung-Kang Peng, and H Eugene Stanley. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation*, 101(23):e215–e220, 2000.

Xiaodong Jia, Bin Huang, Jianshe Feng, Haoshu Cai, and Jay Lee. A review of PHM data competitions from 2008 to 2017: Methodologies and analytics. In *Annual Conference of the PHM Society*, volume 10, 2018.

Rebecca Killick, Paul Fearnhead, and Idris A Eckley. Optimal detection of changepoints with a linear computational cost. *Journal of the American Statistical Association*, 107(500):1590–1598, 2012.

Alexander Lavin and Subutai Ahmad. Evaluating real-time anomaly detection algorithms – the Numenta anomaly benchmark. In *IEEE 14th International Conference on Machine Learning and Applications*, pages 38–44, 2015.

Timothy M Lenton. Early warning of climate tipping points. *Nature Climate Change*, 1(4): 201–209, 2011.

Peter Mattson, Christine Cheng, Cody Coleman, Greg Diamos, Paulius Micikevicius, David Patterson, Hanlin Tang, Gu-Yeon Wei, Peter Bailis, Victor Bittorf, et al. MLPerf: An industry standard benchmark suite for machine learning performance. *IEEE Micro*, 40(2):8–16, 2020.

Mark Mazumder, Colby Banbury, Xiaozhe Yao, Bojan Karlaš, William Gaviria Rojas, Sudnya Diamos, Greg Diamos, Lynn He, Douwe Paesani, Emma Kirk, et al. DataPerf: Benchmarks for data-centric AI development. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. MTEB: Massive text embedding benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2014–2037, 2023.

Abhinav Saxena, Kai Goebel, Don Simon, and Neil Eklund. Damage propagation modeling for aircraft engine run-to-failure simulation. In *International Conference on Prognostics and Health Management*, 2008.

Marten Scheffer, Stephen Carpenter, Jonathan A Foley, Carl Folke, and Brian Walker. Catastrophic shifts in ecosystems. *Nature*, 413(6591):591–596, 2001.

Marten Scheffer, Jordi Bascompte, William A Brock, Victor Brovkin, Stephen R Carpenter, Vasilis Dakos, Hermann Held, Egbert H Van Nes, Max Rietkerk, and George Sugihara. Early-warning signals for critical transitions. *Nature*, 461(7260):53–59, 2009.

Marten Scheffer, Stephen R Carpenter, Timothy M Lenton, Jordi Bascompte, William Brock, Vasilis Dakos, Johan van de Koppel, Ingrid A van de Leemput, Simon A Levin, Egbert H van Nes, Mercedes Pascual, and John Vandermeer. Anticipating critical transitions. *Science*, 338 (6105):344–348, 2012.

Sebastian Schmidl, Phillip Wenig, and Thorsten Papenbrock. Anomaly detection in time series: A comprehensive evaluation. *Proceedings of the VLDB Endowment*, 15(9):1779–1797, 2022.

Ali H Shoeb. *Application of Machine Learning to Epileptic Seizure Detection*. PhD thesis, Massachusetts Institute of Technology, 2010.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Proceedings of the 35th Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2021.

Joel Thorarinson and Allison Hensgen. From prediction to discoverative intelligence: A coherence-based AI framework for detecting system drift before failure. *arXiv preprint*, 2026a.

Joel Thorarinson and Allison Hensgen. The dimensionality illusion: Why PCA variance does not equal semantic information in text embeddings. *arXiv preprint*, 2026b.

Phillip Wenig, Sebastian Schmidl, and Thorsten Papenbrock. TimeEval: A benchmarking toolkit for time series anomaly detection algorithms. *Proceedings of the VLDB Endowment*, 15(12): 3678–3681, 2022.