PARTITION RECOVERY FOR STRUCTURED REPRESENTATION LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Representation learning is typically evaluated through predictive accuracy, reconstruction quality, or information-theoretic objectives. However, in many scientific and decision-making domains, the true objective is to preserve latent distinguishability structure rather than labels alone. A representation can therefore be predictively sufficient while collapsing meaningful latent modes. We formalize this problem as *partition recovery*: recovering the latent equivalence relation induced by task-relevant distinguishability.

We develop a statistical theory of partition recovery, establishing consistency guarantees, finite-sample recovery bounds governed by representation complexity and partition margin, and a characterization showing that recovery under imbalance is controlled by the least-sampled latent mode. These results motivate a stronger notion of representation quality, *task-structure sufficiency*, which requires preservation of the underlying partition rather than predictive performance alone.

To optimize this objective, we propose a differentiable representation-learning framework (DPT) that directly targets partition preservation. DPT can be trained using either domain-derived partitions or partitions statistically bootstrapped from data. Across synthetic and real-world benchmarks, we show that conventional supervised objectives can preserve coarse predictive performance while destroying genuine latent distinctions, whereas DPT retains the underlying structure and reveals representational failures that remain invisible to standard accuracy metrics. Together, these results suggest that representation learning should be evaluated not only by what predictions are preserved, but also by what latent structure survives compression.

1 INTRODUCTION

Consider a standard hierarchical classification task, e.g., CIFAR-100. A model trained on coarse "superclass" labels can achieve 100% accuracy while completely destroying the fine-grained subclass structure. The representation is predictively sufficient for the coarse task but structurally insufficient for the fine one. This phenomenon is not a quirk of image datasets: in scientific and physical domains—fault diagnosis, materials science, cell typing—the "labels" are often arbitrary granularities of an underlying latent partition, and preserving the full distinguishability structure is the actual objective.

Modern self-supervised and contrastive methods (e.g., SimCLR, SwAV, TS2Vec (Chen et al., 2020; Da Costa et al., 2022; Yue et al., 2022)) learn representations without labels, yet they optimize for continuous embedding geometry rather than discrete partition structure. Information bottleneck methods preserve predictive information $I(Z;Y)$, but when Y is a coarse label, $I(Z;Y)$ can be maximized while the underlying partition is destroyed. What is missing is a criterion for task-structure sufficiency: when does a representation preserve all task-relevant distinguishability, independent of a particular label granularity?

To enable partition-learning, we must address three questions. *First*, what object is being learned? We argue that the primary object is the confusable-fibre partition induced by the realized representation. *Second*, when does a learned partition converge to the population partition? This motivates a theory of partition consistency. *Third*, how much data is required to recover a task-specific sufficient partition? This motivates a sample-complexity analysis based on partition margins.

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.

We develop a framework connecting partition-based inference with statistical learning theory, sufficient representations, information bottlenecks, and representation learning, and empirically validate this using realization-theoretic diagnosis examples.

We propose a unifying framework in which any system, whether a first-principles physical model or a neural network, induces a partition on the latent class space through the information resources it employs (sensors, domain knowledge, data, hypothesis class). Two systems are equivalent not when they share architecture or loss function, but when they recover the same partition. This allows direct comparison of symbolic and learned methods through the single lens of partition recovery.

The principal contributions are:

1. *Partition recovery theory*: a theory based on consistency and sample-complexity for recovering latent equivalence relations (not labels). We provide a partition consistency theorem and finite-sample recovery guarantees, including a corollary showing recovery is governed by the least-sampled latent class, $\min_f n_f$, rather than aggregate sample size;
2. *Task-structure sufficiency*: a sufficiency criterion for representations that is strictly stronger than predictive sufficiency.
3. *DPT*: a differentiable partition-learning objective that trains representations to preserve structure without requiring fine-grained labels.
4. *Empirical validation*: cross-benchmark evidence that partition-based metrics reveal structure invisible to methods based on traditional classification accuracy.

Sections 3–7 translate these contributions into six testable hypotheses, H1–H6 (Section 8.2), spanning partition equivalence across realization mechanisms, resource monotonicity, a sufficiency threshold in latent dimension, and robustness to imbalance, noise, and missing labels. Section 8 validates our hypotheses through well-structured experiments.

2 Related Work

We survey the literature across three primary axes, with a comprehensive discussion deferred to Appendix J.

Statistical Learning of Combinatorial Objects. Classical uniform convergence and sample-complexity theory (Vapnik, 1998; Shalev-Shwartz & Ben-David, 2014) focus on the convergence of prediction risk. In contrast, we study the statistical convergence of a combinatorial object—the latent partition—where the partition margin plays a role analogous to the classification margin.

Representation Learning and Information Bottlenecks. Task-structure sufficiency is distinct from classical statistical sufficiency, as it targets partition preservation rather than parameter information. It is strictly stronger than predictive sufficiency: standard supervised and self-supervised objectives (Chen et al., 2020; Da Costa et al., 2022; Yue et al., 2022) optimize for continuous embedding geometry or coarse label prediction $I(Z;Y)$, which can inadvertently collapse confusable-fibre structures (Theorem 2). While the Information Bottleneck (Tishby et al., 1999; 2000) compresses inputs while retaining label information, our Partition Information Bottleneck (Appendix E) optimizes directly for partition preservation. Similarly, while deep clustering (Caron et al., 2020) groups data points heuristically, DPT provides a principled statistical framework grounded in partition consistency and margin bounds.

State Abstraction and Bisimulation. Our framework shares conceptual roots with minimal sufficient state abstractions in Markov Decision Processes (Li et al., 2006) and behavioral bisimulation metrics (Ferns et al., 2004). However, rather than relying on system dynamics or transition kernels, we specialize these concepts to a purely distribution- and signature-based notion of equivalence.

3 Population Partitions and Partition Consistency

We view learning algorithms as inducing partitions on latent modes through the information resources available to them (data, labels, inductive biases, and hypothesis class). Different construc-

tion mechanisms are therefore compared through the partitions they recover rather than through architectural details.

Let Z be a latent space; for each label $f \in F$, assume observations are drawn from an unknown P_f on O. A representation $\phi : O \to Z$ induces a pushforward $\Phi(f) = \phi_{\#}P_f = \mathrm{Law}(\phi(O) \mid f)$, the *latent signature map* $\Phi : F \to \mathcal{P}(Z)$.

Exact equality of the $\Phi(f)$ is unverifiable from finite data, so we define a partition structure via a metric on distributions.

Definition 1 (Signature Metric, ε-Confusability, Population Partition)**.** *A signature metric $d : \mathcal{P}(Z) \times \mathcal{P}(Z) \to \mathbb{R}_{\geq 0}$ satisfies the metric axioms (e.g. Wasserstein, MMD, Jensen-Shannon distance). For $\varepsilon > 0$, faults f_i, f_j are ε-confusable if $d(\Phi(f_i), \Phi(f_j)) \leq \varepsilon$; writing $f_i \sim_\varepsilon f_j$ for this relation, the quotient $\Pi^\star_\varepsilon = F/\sim_\varepsilon$ is the* population partition *– the primary object of task-structure learning, which a representation is useful only to the extent it recovers.*

Definition 2 (Partition Margin)**.** *If $\Pi^\star_\varepsilon$ has more than one block, the partition margin is $\gamma = \min_{[f_i]\neq[f_j]} d(\Phi(f_i), \Phi(f_j))$.*

If $\gamma > \varepsilon$, every pair in distinct population blocks remains separated under $\sim_\varepsilon$ (immediate from the definition of γ).

$\Pi^\star_\varepsilon$ depends on the generally-unknown Φ; this section defines empirical estimators from finite samples $\mathcal{D}_f = \{o_1, \ldots, o_{n_f}\}$ per fault, with latent samples $z_i = \phi(o_i)$.

Definition 3 (Empirical Signature and Partition)**.** *The empirical signature is $\widehat{\Phi}_n(f) = \frac{1}{n_f}\sum_i \delta_{z_i}$ (the empirical distribution of ϕ under fault f). The empirical partition $\widehat{\Pi}_n$ is induced by an empirical confusability relation $\widehat{\sim}$ (below).*

Applying the population threshold ε directly to *empirical* signatures is not sound: an empirical distance can lie strictly between ε and γ from estimation error alone, so no rule tested against ε itself can be guaranteed correct no matter how small that error becomes. The decision rule instead uses a threshold inside the gap (ε, γ):

Definition 4 (Decision Threshold)**.** $\tau = (\varepsilon + \gamma)/2$*. Faults f_i, f_j are* empirically confusable *if $d(\widehat{\Phi}_n(f_i), \widehat{\Phi}_n(f_j)) \leq \tau$.*

Partition Consistency. To establish partition consistency, we first specify our assumptions, which denote the conditions under which $\widehat{\Pi}_n \to \Pi^\star_\varepsilon$: (A1) F finite; (A2) positive margin $\gamma = \min_{[f_i]\neq[f_j]} d(\Phi(f_i), \Phi(f_j)) > \varepsilon$ (excludes partition blocks arbitrarily close); (A3) uniform consistency, $\sup_{f\in F} d(\widehat{\Phi}_n(f), \Phi(f)) \xrightarrow{a.s.} 0$ (satisfied by common distribution estimators under standard regularity). We can now state our primary statistical theorem.

Theorem 1 (Partition Consistency)**.** [1] *Under (A1)–(A3), $\Pr(\widehat{\Pi}_n = \Pi^\star_\varepsilon) \to 1$ as $n \to \infty$; equivalently $R_n \to 0$.*

4 Representation Gap Theorem

The results of Section 3 are statistical: they characterize *when* an empirical partition converges to the population partition, with margin playing the role of a classification margin in classical uniform-convergence theory. This section is different in kind. It establishes *that* the gap between predictive and task-structure sufficiency exists at all (Theorem 2), and *what* the object a task-sufficient representation must recover structurally is (Theorem 3) – neither result involves concentration, sample size, or a margin. Together they motivate the statistical machinery that follows: recovering $\Pi^\star$ is worth a consistency theory only because representations that are otherwise perfectly good (predictively sufficient, Bayes-optimal) can fail to recover it.

Assumption 1 (Label Determinism)**.** *The coarse label Y is a deterministic function of the observation: there exists a measurable $\ell : X \to Y$ with $Y = \ell(X)$ almost surely.*

This is realistic rather than restrictive: without Assumption 1, $H(Y \mid X)$ need not be zero and the equality driving the proof below does not hold in general.

[1] We present all proofs in Appendix D.

Theorem 2 (Representation Gap). *Let $F = \{f_1, \ldots, f_m\}$ be a latent mode set and let $Y = g(F)$ be a non-injective coarse labeling, i.e., $\exists f_i \neq f_j$ such that $g(f_i) = g(f_j)$. Suppose observations X are generated from latent modes F, satisfy Assumption 1, and consider representations $\phi : X \to Z$ used only for prediction of Y. Then there exists a representation $\phi^\star$ such that $I(\phi^\star(X); Y) = I(X; Y)$, so that $\phi^\star$ is predictively sufficient for Y, yet $\Pi_{\phi^\star} \neq \Pi^\star$, where $\Pi^\star$ is the latent task partition that distinguishes f_i and f_j.*

Moreover, the induced partition collapses every pair of latent modes sharing the same coarse label: $g(f_i) = g(f_j) \implies f_i \sim_{\phi^\star} f_j$. Consequently, predictive sufficiency does not imply task-structure sufficiency.

This is an *existence* statement – it exhibits one predictively-sufficient representation that fails to be task-sufficient, not a claim about what representations gradient-based training typically finds. Section 8's coarse-label experiments (the synthetic `f_pump`/`f_valve` pair and the real `IR007`/`IR028` pair, both under CIFAR-100-style location-only coarse labels) show the same collapse occurring in practice under ordinary supervised training, not only for the constructed $\phi^\star$ of the proof above.

Corollary 1 (Accuracy Does Not Certify Structure). *Let $\Pi^\star$ be a task partition on latent modes F and let $Y = g(F)$ be a non-injective coarse labeling satisfying Assumption 1. Then there exists a representation ϕ such that $R_{\mathrm{cls}}(\phi) = R^\star_{\mathrm{Bayes}}$, where $R^\star_{\mathrm{Bayes}}$ is the Bayes-optimal classification risk for predicting Y, yet $VI(\Pi_\phi, \Pi^\star) > 0$. Consequently, achieving Bayes-optimal classification performance does not certify recovery of the latent task structure.*

4.1 Relationship to Latent-State Abstractions

Theorem 3 (Partition Recovery as Abstraction Recovery). *Let F be a latent state space and let $\Pi^\star = \{C_1, \ldots, C_k\}$ be the task partition induced by confusability. Define the abstraction map $\alpha : F \to \{1, \ldots, k\}$ by assigning each latent state to its partition block.*

Then:

1. *α is sufficient for task structure;*
2. *every task-sufficient representation factors through α, and is related to it by an injective relabeling of blocks;*
3. *consequently, α is the unique task-sufficient abstraction up to relabeling: no strictly coarser map is task-sufficient, and no other task-sufficient map is strictly finer than a relabeling of α.*

Hence recovering $\Pi^\star$ is equivalent to recovering the (essentially unique) task-sufficient latent abstraction. This is the partition-recovery analogue of minimal sufficient state abstraction in MDPs (Li et al., 2006) and of exact bisimulation (Ferns et al., 2004), specialized to a distribution-based rather than a dynamics-based notion of equivalence.

5 Approximate Partition Recovery

Exact recovery (Section 3) assumes the empirical partition eventually equals the population partition, which may be unrealistic under measurement noise, model uncertainty, or intrinsically similar faults. This section formulates the approximate partition recovery task.

Assumption 2 (Uniform Concentration). *There exists $\psi(n, \delta)$, depending on the distribution estimator, representation class, and metric d, such that $\Pr(\sup_{f \in F} d(\widehat{\Phi}_n(f), \Phi(f)) \leq \psi(n, \delta)) \geq 1 - \delta$.*

Definition 5 (Variation of Information). *For partitions Π_1, Π_2 of F, $VI(\Pi_1, \Pi_2) = H(\Pi_1) + H(\Pi_2) - 2I(\Pi_1; \Pi_2)$, where H is partition entropy and I is mutual information. VI is a metric on finite partitions: $VI(\Pi_1, \Pi_2) = 0$ iff $\Pi_1 = \Pi_2$.*

Definition 6 (Partition Loss and Approximate Recovery). *The partition loss is $\Delta_\Pi(n) = VI(\widehat{\Pi}_n, \Pi^\star_\varepsilon)$. A sequence of learned partitions satisfies approximate partition recovery if $\Delta_\Pi(n) \to 0$ in probability as $n \to \infty$.*

Assumption 3. *We assume: (A1) F is finite; (A2) d is continuous; (A3)* $\sup_{f \in F} d(\widehat{\Phi}_n(f), \Phi(f)) \xrightarrow{P} 0$ *– neither exact separation nor a positive margin is required; (A4)* genericity*: no two distinct faults have population signature distance exactly ε, i.e. $d(\Phi(f_i), \Phi(f_j)) \neq \varepsilon$ for all $f_i \neq f_j$.*

(A4) is needed because confusability is a *thresholded* indicator, $\mathbf{1}[d(\cdot, \cdot) \leq \varepsilon]$, which is discontinuous exactly at ε: convergence in probability of a pairwise distance to its population value only forces the corresponding indicator to converge if that population value is not itself sitting on the threshold. (A4) rules this out; it is a mild genericity condition, violated only for a knife-edge choice of ε against a fixed set of signature distributions, and does not require a uniform margin across pairs the way (A2) of Section 3 does. Note that continuous distribution metrics like MMD or Wasserstein rarely land on a knife-edge boundary unless engineered to do so.

The population and empirical confusability matrices are $c^\star_{ij} = \mathbf{1}[d(\Phi(f_i), \Phi(f_j)) \leq \varepsilon]$ and $\hat{c}_{ij} = \mathbf{1}[d(\widehat{\Phi}_n(f_i), \widehat{\Phi}_n(f_j)) \leq \tau]$ (decision threshold τ from Section 3); together they determine the corresponding partitions.

Theorem 4 (Approximate Partition Recovery)**.** *Under (A1)–(A4), $VI(\widehat{\Pi}_n, \Pi^\star_\varepsilon) \xrightarrow{P} 0$, equivalently $\Delta_\Pi(n) \xrightarrow{P} 0$.*

6 Sample Complexity of Partition Recovery

This section characterizes the number of samples required for exact partition recovery, $R_n = \Pr(\widehat{\Pi}_n \neq \Pi^\star_\varepsilon) \to 0$. As $\gamma \to \varepsilon$, the effective margin $\gamma - \varepsilon$ – the width of the gap into which the decision threshold $\tau = (\varepsilon + \gamma)/2$ (Section 3) must be placed – shrinks, and recovery becomes harder.

Theorem 5 (Finite-Sample Partition Recovery)**.** [2] *If $\psi(n, \delta) < (\gamma - \varepsilon)/4$ and $\widehat{\Pi}_n$ uses threshold $\tau = (\varepsilon + \gamma)/2$, then $\Pr(\widehat{\Pi}_n = \Pi^\star_\varepsilon) \geq 1 - \delta$.*

Let $C(H)$ be a capacity measure of hypothesis class H (VC dimension, Rademacher complexity, covering numbers, ...). Assuming $\psi(n, \delta) = O(\sqrt{(C(H) + \log(1/\delta))/n})$:

Theorem 6 (Sample Complexity Bound)**.** *Exact partition recovery occurs whenever $n = \Omega\Big(\frac{C(H) + \log(1/\delta)}{(\gamma - \varepsilon)^2}\Big)$.*

This states complexity in terms of total n, but the empirical signature for fault f is estimated from only the n_f observations for that fault, and real fault occurrence is rarely balanced.

Corollary 2 (Imbalance-Limited Recovery)**.** *Suppose the per-fault concentration rate is $\psi(n_f, \delta/m) = O(\sqrt{(C(H) + \log(m/\delta))/n_f})$. Then, under $\gamma > \varepsilon$, exact recovery with probability $\geq 1 - \delta$ occurs whenever $\min_{f \in F} n_f = \Omega\Big(\frac{C(H) + \log(m/\delta)}{(\gamma - \varepsilon)^2}\Big)$. Theorem 6 is the special case $n_f \equiv n/m$, and understates the true requirement under imbalance: it can be satisfied by large total n while $\min_f n_f$ remains far below threshold.*

A dataset can be arbitrarily large in total while a single rare class remains outside the regime that guarantees recovery, because its own n_f is small – the statistical-learning counterpart of incipient or rarely observed faults being the hard case in fleet data because they are least sampled, not because they are intrinsically less separable. Section 8 tests this corollary directly. Sample complexity is governed jointly by representation complexity $C(H)$, the partition margin γ, and confidence δ: learning a task sufficient representation is equivalent to recovering the population partition, and is fundamentally a partition-recovery problem rather than a classification one.

7 Partition Learning

This section presents DPT, a weakly-supervised representation-learning method: DPT requires only a pairwise equivalence matrix $P^\star$ (which can be bootstrapped from data or derived from domain

[2] Full proof in Appendix G.

knowledge) and no fine-grained class labels. It is related to contrastive learning, but instead of pulling 'positive' pairs and pushing 'negatives' in embedding space, it directly minimizes a soft variation-of-information to a target partition. This makes it suitable for any domain where the supervision is a partition rather than a label vector.

Task-specific partitions are discrete, so direct optimization is non-differentiable; this section develops a differentiable relaxation for end-to-end representation learning.

Definition 7 (Soft Partition Matrix)**.** *For empirical latent signatures $\widehat{\Phi}_i = \widehat{\Phi}(f_i)$ under $\phi_\theta : O \to Z$, let $P_{ij} = \exp(-d(\widehat{\Phi}_i, \widehat{\Phi}_j)^2/\tau) \in [0, 1]$, temperature $\tau > 0$: P_{ij} is the soft probability that f_i, f_j share a fibre.*

Proposition 1 (Hard Partition Limit)**.** $\lim_{\tau \to 0} P_{ij} = \mathbf{1}[d(\widehat{\Phi}_i, \widehat{\Phi}_j) = 0]$*: soft partitions converge to classical partitions as $\tau \to 0$.*

7.1 Reference Partition Matrix

This section shows how a reference partition can be bootstrapped statistically, with no symbolic model at all. The raw-channel pipeline of Sections 3–6 already constructs an empirical partition $\widehat{\Pi}_n$ from finite raw samples alone; given a large pilot n_{pilot}, the Partition Consistency Theorem guarantees $\widehat{\Pi}_{n_{\text{pilot}}}$ converges to the true partition, and using it in place of $\Pi^\star$ trains ϕ_θ against a purely data-derived target. Section 8 reports this end to end: a pilot of $n_{\text{pilot}} = 500$ recovers the exact population partition with no symbolic model, and DPT trained against it is evaluated at sample sizes far below the pilot.

Under this reading, DPT's value is not recovering task structure a raw-channel estimator could not – by a data-processing argument, $\phi_\theta(O)$ cannot exceed O's partition information, so it cannot outperform raw-channel recovery asymptotically. Its value is finite-sample and compression-driven: estimating a signature directly in a high-dimensional, nuisance-heavy space is statistically harder than in a compressed embedding that has learned which directions carry partition information – the Partition Compression Principle of Appendix C, made load-bearing here.

A symbolic $\Pi^\star$ remains available where one exists (a strong-fault Reiter, diagnoser, observer, or geometric realization, or the companion DSS framework's minimal sufficient realization). Where DSS has already established this for a benchmark and sensor configuration, it supplies a provably-exact $\Pi^\star$ for *evaluation* (Section 8) independently of whichever source trains DPT – DSS characterizes $\Pi^\star$ exactly in the symbolic-model limit; the bootstrap estimates the same object statistically. Define $P^\star_{ij} = \mathbf{1}[f_i \sim f_j]$.

7.2 Partition-Based Bottleneck Loss

The reconstruction loss $L_{\text{part}} = \frac{1}{m^2}\|P_\theta - P^\star\|_F^2$ (or a BCE variant) matches the soft partition to $P^\star$. A differentiable variation-of-information surrogate $\widetilde{VI} = \widetilde{H}(\Pi_O) + \widetilde{H}(\Pi_Z) - 2\widetilde{I}(\Pi_O; \Pi_Z)$ is built from soft cluster memberships $q_{ik} \geq 0$, $\sum_k q_{ik} = 1$ in the same way as classical entropy/mutual information. The partition bottleneck loss is $L_{\text{diag}} = \widetilde{VI} + \beta\, C(Z)$, with $C(Z)$ a complexity penalty (latent dimension, covariance trace, or $I(X; Z)$) trading off partition preservation against compression.

Definition 8 (Task-structure Sufficiency)**.** *$\phi : O \to Z$ is* task sufficient *if $\Pi_Z = \Pi_O$. (Appendix C develops equivalent characterizations and consequences for detectability/isolability.)*

Theorem 7 (Global Optimum and task-structure sufficiency)**.** *If H contains a task sufficient representation, every global minimizer of $\widetilde{VI}$ satisfies $\Pi_Z = \Pi_O$, hence is task sufficient.*

Remark 1 (The Theorem Is About the Optimum, Not the Optimizer)**.** *The theorem characterizes the* global *minimizer of a non-convex objective over neural network parameters θ; it says nothing about whether gradient-based training of ϕ_θ (Algorithm 1, Section H) reaches it, only that reaching it would suffice. This is the same distinction already made explicit for the unlabeled setting's Proposition 5 (k-means finding a global, not merely local, optimum), and it is not merely hypothetical here either: Section 8 reports DPT's own overall recovery as imperfect ($\text{ARI} \approx 0.80$) at a narrow embedding width, attributed there to the optimization budget of a small network on a short training schedule rather than to this theorem being wrong. A convergence-rate or optimization-landscape*

analysis of L_{diag} is not attempted here and is left to future work; what the empirical results support is narrower and still useful – L_{diag} is a training signal whose global optimum is provably correct, and whose practical optimization gap is capacity- and budget-dependent rather than fundamental (Section 8's finding that the coarse-label counterexample itself disappears at a wider embedding is the same phenomenon from the other direction).

7.3 Training (DPT)

ϕ_θ is trained by: fixing a signature metric (MMD with an RBF kernel throughout, Wasserstein/JS as a robustness ablation, Section I.9); at each step, sampling a per-fault mini-batch, forming P_θ via a differentiable MMD estimator, and taking a gradient step on L_{diag} against $P^\star$ (symbolic or bootstrapped) while annealing $\tau \to 0$; stopping on held-out $\widetilde{VI}$. No fault-specific labels are used at any step, only the pairwise $P^\star$. Full algorithm in Appendix H.

8 Experimental Methodology

This section evaluates partition recovery, task-structure sufficiency, and the six hypotheses below, rather than classification accuracy alone.

8.1 Benchmark Systems and Experimental Design

We study two benchmark systems: an enriched two-tank process (8 fault modes: pump, valve, leak, a topological guard fault, two sensor biases, nominal, and a multi-fault case, read through several sensor/field configurations that induce different confusable-fibre structures) and the IEEE 7-bus radial feeder under LinDistFlow (12 fault modes: line, injection, and sensor faults, read under four sensor placements M_1–M_4). Both benchmarks are deliberately chosen so that the exact confusable-fibre partition $\Pi^\star_\varepsilon$ is computable in closed form (Section 7.1), giving a provable ground truth to score against rather than an assumed one. Appendix H gives full benchmark and factor specifications.

We compare five realization classes: strong- and weak-fault (Feldman et al., 2010) consistency-based (symbolic) realizations, a learned supervised realization (fine-grained fault labels), a learned anomaly-detection realization (nominal-vs-abnormal), and the DPT realization of Section 7 (no fault-specific labels, trained via the bottleneck loss against a reference partition obtained either symbolically or via the statistical bootstrap of Section 7.1).

Given these benchmarks, our experimental design varies 7 factors, and applies metrics over pairwise agreement (PA), adjusted Rand index (ARI), variation of information (VI), and *fibre purity* and *completeness* – the two conditional-entropy components whose sum is VI, reported separately because they distinguish two failure modes a single VI number conflates: purity falls when an estimator wrongly *merges* distinguishable faults, completeness falls when it wrongly *fragments* a truly confusable fibre. We describe full factor levels, design, and metrics in Appendix H.5.

8.2 Hypotheses

This section summarizes our experimental hypotheses. Due to space limitations, we present full formal descriptions in Appendix H.10; Appendix H.1 maps each hypothesis to the specific figures, tables, and statistical tests that address it.

H1 (Partition Equivalence): task-equivalent realizations satisfy $\mathrm{ARI} \approx 1$.

H2 (Partition Preservation): partition equivalence implies equal detectability and isolability.

H3 (Knowledge-Induced Refinement): stronger fault knowledge yields a strictly finer partition.

H4 (Resource Monotonicity): removing sensors, fault knowledge, training data, or latent dimension coarsens the recovered partition.

H5 (Sufficiency Threshold): there exists a latent dimension $d^\star$ above which $\Pi_Z = \Pi^\star$.

H6 (Robustness): imbalance and noise degrade purity more than completeness (false merges), while reduced labels degrade completeness more than purity (fragmentation), with imbalance-driven

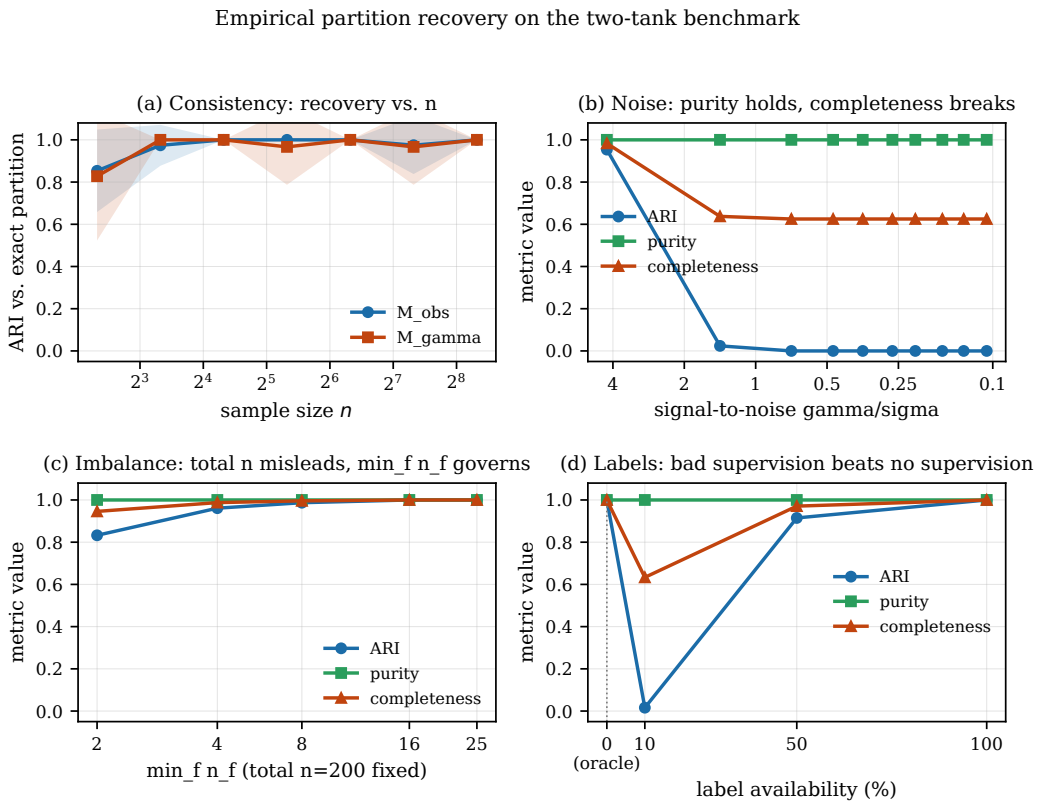

Figure 1: Empirical partition recovery on the two-tank benchmark against the exact ground-truth partition, prior to introducing a learned representation. (a) Recovery converges to exact as n grows, for two diagnosable-inequivalent fields (H4). (b)–(d) Recovery under noise, imbalance, and reduced labels (H6). Shaded regions and error bars, where visible, are ± 1 standard deviation over 30 random seeds per condition.

degradation following the Imbalance-Limited Recovery corollary's rate in $\min_f n_f$ rather than in total n.

8.3 Results

H4 and H6, raw channels. Figure 1 evaluates the statistical machinery of Sections 3–6 directly on raw noisy observations, prior to introducing a learned ϕ_θ – the necessary precondition for attributing any subsequent learned-representation result to ϕ_θ itself. Panel (a) shows two sensor/field configurations converging to two *different* exact partitions as n grows (**H4**: sensor information changes partition granularity), exactly as the Partition Consistency Theorem predicts. Panels (b)–(d) test **H6**: recovery degrades once noise exceeds the closed-form margin γ (b); at a *fixed total* sample budget, increasing imbalance degrades recovery via $\min_f n_f$ collapsing, not via any change in aggregate n (c), exactly the Imbalance-Limited Recovery corollary's rate; and reduced label availability degrades recovery sharply at $\ell = 10\%$, with an oracle unsupervised baseline at $\ell = 0\%$ *outperforming* the $\ell = 10\%$ semi-supervised estimator (d) – unreliable supervision can corrupt a signature more than having none at all. A single mechanism underlies (b)–(d), confirming **H6** precisely: purity stays at or near 1.0 throughout, while completeness is what degrades – none of the three stressors causes wrongful *merging*, all three cause *fragmentation* of a confusable fibre.

The same three sweeps on the structurally unrelated IEEE 7-bus feeder reproduce the identical purity-holds/ completeness-breaks signature and the oracle-beats-bad- supervision reversal (**H6** again, on a completely different fault taxonomy), evidence that the mechanism belongs to the estimator, not to the tank benchmark's particular structure (Appendix H.13).

H1, H3, H4: realization-class comparison. We evaluate the four realization classes directly with a small trained ϕ_θ, held out on a test split never seen during training. *DPT and fine-grained supervision recover partitions equivalent to raw channels* (confirms **H1**), while coarse-label supervision does not: two genuinely distinguishable synthetic faults sharing a coarse label collapse toward the true population margin γ under coarse supervised training, but stay separated under DPT, which – using strictly less fault-specific information than the coarse-label classifier – nonetheless recovers a strictly finer partition (**H3/H4**: less task-relevant knowledge yields a coarser partition). This effect is capacity-conditional (it requires a narrow embedding; a wider one has enough spare capacity to preserve the distinction unasked). The same mechanism replicates on real bearing-vibration data (CWRU, 16 experimentally-verified fault conditions), where it is sharper and there is no synthetic ground truth to lean on: grouping two inner-race faults of different severity under one location-only coarse label collapses their pairwise MMD from 0.69 (raw features) to 0.14 under coarse supervision, while DPT retains 0.49 – and overall $\mathrm{ARI} = 1.0$ throughout for all four realizations, so the failure is entirely invisible in that aggregate metric, an argument for reporting pairwise distances and purity/completeness rather than ARI alone. Full figures (synthetic and real-data) and exact numbers are in Appendix H.13.

H5: DPT and sample efficiency. Training DPT against a purely data-bootstrapped reference partition (no symbolic model, Section 7.1) in a 24-dimensional ambient space (4 true signal channels, 20 nuisance channels) recovers the true partition at $\mathrm{ARI} \approx 0.87$–$0.93$ for $n \in \{5, 10, 20\}$, where raw-channel recovery in the same space fails outright ($\mathrm{ARI} \approx 0$–0.06), with both converging by $n = 40$ – consistent with **H5**: a compressed, task-sufficient embedding needs far less data than the full ambient space. A wrong-quotient ablation and two further design-choice ablations (full results in Appendix H.12.1) validate the training procedure itself: an incorrect reference partition produces an irreducible bias invariant to n rather than shrinking estimation error, and the DPT loss's separation term and reference-partition construction are both necessary for the recovery reported above.

9 Conclusion

This paper argues that representation learning should be evaluated not only by predictive performance, but also by preservation of latent distinguishability structure. A representation can achieve high accuracy while collapsing meaningful latent modes, making predictive sufficiency an incomplete notion of representation quality. We formalized this problem through *partition recovery* and introduced *task-structure sufficiency*, a criterion requiring preservation of the underlying partition induced by the data.

We developed a statistical theory of partition recovery, including consistency guarantees, finite-sample recovery bounds governed by representation complexity and partition margin, and an imbalance-sensitive characterization showing that recovery is controlled by the least-sampled latent mode. These results provide a learning-theoretic foundation for evaluating representations through the structure they preserve rather than solely through downstream prediction.

To operationalize this objective, we proposed Diagnostic Partition Training (DPT), a differentiable representation-learning framework that optimizes partition preservation directly. Empirical results demonstrate that conventional supervised objectives can preserve coarse predictive performance while destroying genuine latent distinctions, whereas partition-oriented training retains the underlying structure and reveals failures invisible to standard accuracy metrics.

More broadly, the partition-recovery perspective covers scientific machine learning, self-supervised learning, latent-state discovery, and other settings where the objective is recovery of meaningful latent structure rather than prediction alone. We hope this work motivates further study of representation learning through the lens of structure preservation and partition recovery.

Appendix I presents a full discussion of limitations and threats to validity. These issues include learnability of approximate partitions, sensitivity of the calibrated decision threshold to pilot sample size, the impact of unlabeled and weakly supervised data, etc.

REFERENCES

Sanjeev Arora, Rong Ge, Behnam Neyshabur, and Yi Zhang. Stronger generalization bounds for deep nets via a compression approach. In *International conference on machine learning*, pp. 254–263. PMLR, 2018.

Peter Bartlett, Dylan J Foster, and Matus J Telgarsky. Spectrally-normalized margin bounds for neural networks. *Advances in neural information processing systems*, 30, 2017.

Michael M. Bronstein, Joan Bruna, Yann LeCun, Arthur Szlam, and Pierre Vandergheynst. Geometric deep learning: Going beyond euclidean data. *IEEE Signal Processing Magazine*, 34(4): 18–42, 2017.

Michael M. Bronstein, Joan Bruna, Taco Cohen, and Petar Velickovic. Geometric deep learning: Grids, groups, graphs, geodesics, and gauges. *arXiv:2104.13478*, 2021.

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 9912–9924, 2020.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. A simple framework for contrastive learning of visual representations. *CoRR*, abs/2002.05709, 2020. URL `https://arxiv.org/abs/2002.05709`.

Victor Guilherme Turrisi Da Costa, Enrico Fini, Moin Nabi, Nicu Sebe, and Elisa Ricci. solo-learn: A library of self-supervised methods for visual representation learning. *Journal of Machine Learning Research*, 23(56):1–6, 2022.

Johan De Kleer, Alan K Mackworth, and Raymond Reiter. Characterizing diagnoses and systems. *Artificial intelligence*, 56(2-3):197–222, 1992.

Steven X. Ding. *Data-Driven Design of Fault Diagnosis and Fault-Tolerant Control Systems*. Springer, 2014.

Gintare Karolina Dziugaite and Daniel M Roy. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. *arXiv preprint arXiv:1703.11008*, 2017.

Alexander Feldman, Gregory Provan, and Arjan Van Gemund. Approximate model-based diagnosis using greedy stochastic search. *Journal of Artificial Intelligence Research*, 38:371–413, 2010.

Norm Ferns, Prakash Panangaden, and Doina Precup. Metrics for finite Markov decision processes. In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 162–169, 2004.

Jean-Yves Franceschi, Aymeric Dieuleveut, and Martin Jaggi. Unsupervised scalable representation learning for multivariate time series. *Advances in Neural Information Processing Systems (NeurIPS)*, 32, 2019.

Janos Gertler. *Fault Detection and Diagnosis in Engineering Systems*. Marcel Dekker, 1998.

Lihong Li, Thomas J Walsh, and Michael L Littman. Towards a unified theory of state abstraction for MDPs. In *Proceedings of the Ninth International Symposium on Artificial Intelligence and Mathematics (AI&M)*, pp. 531–539, 2006.

Bryan Lim and Stefan Zohren. Time-series forecasting with deep learning: a survey. *Philosophical Transactions of the Royal Society A*, 379(2194):20200209, 2021.

Yuxuan Nie, Nam H Nguyen, Pichayoot Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. In *International Conference on Learning Representations (ICLR)*, 2023.

David Pollard. Strong consistency of k-means clustering. *The annals of statistics*, pp. 135–140, 1981.

Gregory Provan. A realization-theoretic foundation for fault diagnosis, with diagnosability equivalence predictions. *Proc. Conf. on Diagnosis and Resilient Systems*, 2026.

Maziar Raissi. Deep hidden physics models: Deep learning of nonlinear partial differential equations. *Journal of Machine Learning Research*, 19(25):1–24, 2018.

Raymond Reiter. A theory of diagnosis from first principles. *Artificial Intelligence*, 32(1):57–95, 1987.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.

Ben Taskar, Vassil Chatalbashev, Daphne Koller, and Carlos Guestrin. Learning structured prediction models: A large margin approach. In *Proceedings of the 22nd International Conference on Machine Learning (ICML)*, pp. 896–903, 2005.

Naftali Tishby, Fernando Pereira, and William Bialek. The information bottleneck method. In *Proceedings of the 37th Annual Allerton Conference on Communication, Control and Computing*, pp. 368–377, 1999.

Naftali Tishby, Fernando Pereira, and William Bialek. The information bottleneck method. *arXiv:physics/0004057*, 2000.

Vladimir Vapnik. *Statistical Learning Theory*. Wiley, 1998.

Long Wen, Xiang Li, Liang Gao, and Yaran Zhang. Transformers in fault diagnosis: A review. *IEEE Transactions on Industrial Informatics*, 2023.

Jan C. Willems. The behavioral approach to open and interconnected systems. *IEEE Control Systems Magazine*, 27(6):46–99, 2007.

Zhihan Yue, Yujing Wang, Jianghao Duan, Tianyi Yang, Conghui Huang, Yunqi Cao, Yang Wang, Bo Chen, and Jianxin Zhang. TS2Vec: Towards universal representation of time series. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 8985–8993, 2022.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*, 2016.

A Required Statements

A.1 AI Use Statement

In this article, we used AI to provide critiques of the paper and identify technical holes and literature gaps. All theory and experimental design is the work of the authors. Partial AI assistance was used for implementation and debugging.

A.2 Ethics Statement

We certify that we fully adhere to the ICLR Code of Ethics.

A.3 Reproducibility Statement

All efforts have been made to ensure reproducibility. Proofs for all theorems are presented in the Appendices. We provide complete experimental designs in the Appendix, together with all necessary parameters and hyperparameters required. Each hypothesis validated is stated clearly, and where appropriate we employ suitable statistical tests.

For novel models or algorithms, we attach anonymous downloadable source code as supplementary materials; for theoretical results, clear explanations of any assumptions and a complete proof of the claims has been included in the appendix; for any datasets used in the experiments, a complete

description of the data processing steps is provided in the supplementary materials (as needed). For public data we provide the link/reference used to download the data.

Guide to Appendices

The three subsections below develop task-structure sufficiency, a partition information theory, and a partition information bottleneck built on top of the partition-recovery framework of the main text. They are not required for the main text's central claims, experimental results, or the DPT training procedure of Section 7 (which uses only the one-line task-structure sufficiency definition restated there); they are included here for completeness and because Appendix E and Appendix F are referenced once each, in Section 8 and Section 7 respectively, as the source of the task-structure counterexample argument and the information-theoretic metrics reported there.

B Realization Construction and Partitions

Let $F = \{f_1, \ldots, f_m\}$ be a finite label set and O an observation space. We construct a realization $R : (B, \Lambda) \to S$ from a behavioural specification B and inductive bias specification (auxiliary information) Λ.

Definition 9 (Inductive Bias Structure)**.** *An inductive bias structure is $\Lambda = (S, K, D, H)$: S is input data, K label knowledge, D observational data, and H a hypothesis class. We express symbolic vs. learning-based definitions by controlling which Λ generates the realization.*

A standard representation-learning realization uses only $\Lambda = (S, D, H)$: it has no explicit generative model or domain-knowledge inductive bias over the latent class structure, and must discover all distinctions from data via the hypothesis class H. By contrast, a realization with full domain knowledge (strong structural prior) employs $K = \{M_N, M_{f_1}, \ldots, M_{f_m}\}$, where both the nominal data manifold and each latent-class-conditional distribution are explicitly modeled; a weak structural prior uses only $K = \{M_N\}$, modeling latent classes as perturbations of nominal behavior—equivalent to learning with anomaly-detection or one-class constraints.

Remark 2 (Calibration)**.** *γ is unknown in advance; τ is calibrated from a pilot sample independent of the evaluation sample (reusing the evaluation sample to estimate γ biases the reported recovery rate optimistically). Appendix H specifies this calibration step.*

The learning problem is recovery of $\Pi^\star_\varepsilon$ in variation-of-information loss, $\Delta_\Pi(n) := VI(\widehat{\Pi}_n, \Pi^\star_\varepsilon) \to 0$ (defined formally in Section 5); partition consistency, sample complexity, task-structure sufficiency, and the information bottleneck of Appendix F are all statements about this same quantity Δ_Π. The partition recovery error is $R_n = \Pr(\widehat{\Pi}_n \neq \Pi^\star_\varepsilon)$; its convergence to zero is the central statistical problem of the remainder of the paper.

C Task-Structure Sufficiency

The consistency results of the previous sections establish conditions under which an empirical partition converges to a population partition.

A deeper question remains. Given a representation $\phi : O \to Z$, when does the representation preserve all partition information present in the original observations?

This section introduces the notion of task-structure sufficiency. The resulting concept plays a role analogous to statistical sufficiency in classical estimation theory.

C.1 Population Partitions

Let Π_O denote the population partition induced by the original observation space O. Let Π_Z denote the population partition induced by the latent representation space Z. The representation ϕ is task-structure useful only insofar as it preserves the distinguishability structure encoded by Π_O.

C.2 Task-Structure Sufficiency

Definition 10 (Task-Structure Sufficiency)**.** *A representation $\phi : O \to Z$ is task sufficient with respect to fault set F if $\Pi_Z = \Pi_O$.*

Thus a task sufficient representation preserves the entire confusable-fibre structure of the original observations.

Unlike reconstruction objectives, task-structure sufficiency does not require preserving all information contained in O. Only information relevant to fault distinguishability must be retained.

C.3 Equivalent Characterizations

Let $\Phi_O : F \to \mathcal{P}(O)$ denote the population signature map in observation space.

Likewise let $\Phi_Z : F \to \mathcal{P}(Z)$ denote the latent signature map.

Theorem 8 (Equivalent Characterizations of Task-Structure Sufficiency)**.** *The following statements are equivalent.*

1. *ϕ is task sufficient.*
2. $\Pi_Z = \Pi_O$.
3. $\mathrm{Ker}(\Phi_Z) = \mathrm{Ker}(\Phi_O)$.
4. *For every pair of faults $f_i, f_j \in F$, $f_i \sim_O f_j \iff f_i \sim_Z f_j$.*

Proof. (1)⇒(2):

This is the definition of task-structure sufficiency.

(2)⇒(3):

Partitions are generated by the fibres of their corresponding signature maps. If the partitions are equal, then the corresponding equivalence classes are identical. Hence the kernels coincide.

(3)⇒(4):

The kernel of a signature map determines which fault pairs belong to the same fibre. Equality of kernels therefore implies identical pairwise confusability relations.

(4)⇒(2):

If every pairwise confusability relation is preserved, the induced equivalence classes are preserved. Hence $\Pi_Z = \Pi_O$. Therefore all four statements are equivalent.

□

C.4 Latent Mode Preservation

The principal importance of task-structure sufficiency is that it guarantees preservation of task-structure capability.

Theorem 9 (Latent Mode Preservation)**.** *Suppose ϕ is task sufficient. Then $D_Z = D_O$ and $I_Z = I_O$, where D denotes detectability and I denotes a partition is possible.*

Proof. Because $\Pi_Z = \Pi_O$, the fault pairs occupying common fibres are identical in both spaces. Detectability depends upon separation from the nominal fibre. Isolability depends upon pairwise fibre separation. Since fibre memberships are unchanged, detectability and ability to induce partitions are unchanged. Therefore $D_Z = D_O$, and $I_Z = I_O$. □

C.5 Minimally Task-Sufficient Representations

Statistical sufficiency motivates the search for the smallest representation preserving task structure.

Definition 11 (Minimally Task-Sufficient Representation)**.** *A task sufficient representation $\phi^\star$ is minimal if every task sufficient representation ψ satisfies* $\dim(\phi^\star(O)) \leq \dim(\psi(O))$.

The existence and uniqueness of such representations is a separate question. The realization-theoretic framework suggests that they should correspond to a quotient induced by the partition.

C.6 Partition Compression

Task-structure sufficiency permits aggressive compression of observations provided partition structure is preserved.

Proposition 2 (Partition Compression Principle)**.** *Let $\phi : O \to Z$ be task sufficient. Then any information removed by ϕ is irrelevant to fault distinguishability.*

Proof. Task-structure sufficiency implies $\Pi_Z = \Pi_O$. Hence every pair of faults distinguishable in O remains distinguishable in Z, and every confusable pair remains confusable.

Therefore all information discarded by ϕ lies outside the partition structure.

□

C.7 Relationship to Statistical Sufficiency

Task-structure sufficiency is analogous to classical statistical sufficiency but differs in its objective. A statistically sufficient statistic preserves all information about a parameter. A task sufficient representation preserves all information about fault distinguishability. Consequently the target object is not a parameter estimator but a partition of fault space.

The representation-learning problem therefore becomes $\Pi_O \longrightarrow \Pi_Z$ rather than $X \longrightarrow Y$.

C.8 Task-Structure Sufficiency Risk

The consistency results of Sections 4–6 motivate a natural notion of representation-learning risk.

Definition 12 (Task-Structure Sufficiency Risk)**.** *Let $\hat{\Pi}_n$ denote the empirical partition.*

Define

$$R_D(n) = \Pr\left(\hat{\Pi}_n \neq \Pi_O\right).$$

A representation-learning procedure is partition-consistent if $R_D(n) \to 0$ as $n \to \infty$. By the Partition Consistency Theorem (Section 3), this property holds whenever the assumptions of partition consistency are satisfied.

C.9 Interpretation

The previous sections demonstrate that a learned representation should not be evaluated solely through reconstruction error, classification accuracy, mutual information, or latent dimensionality.

The primary criterion is preservation of partition structure.

Representation learning succeeds precisely when the learned partition converges to the task sufficient partition induced by the original observation space.

D Representation Gap Theorem

Theorem 10 (Representation Gap)**.** *Let $F = \{f_1, \ldots, f_m\}$ be a latent mode set and let $Y = g(F)$ be a non-injective coarse labeling, i.e., $\exists f_i \neq f_j$ such that $g(f_i) = g(f_j)$. Suppose observations X*

are generated from latent modes F, satisfy Assumption 1, and consider representations $\phi : X \to Z$ used only for prediction of Y. Then there exists a representation $\phi^\star$ such that $I(\phi^\star(X); Y) = I(X; Y)$, so that $\phi^\star$ is predictively sufficient for Y, yet $\Pi_{\phi^\star} \neq \Pi^\star$, where $\Pi^\star$ is the latent task partition that distinguishes f_i and f_j.

Moreover, the induced partition collapses every pair of latent modes sharing the same coarse label: $g(f_i) = g(f_j) \implies f_i \sim_{\phi^\star} f_j$. Consequently, predictive sufficiency does not imply task-structure sufficiency.

Proof. Since g is non-injective, there exists at least one pair $f_i \neq f_j$ such that $g(f_i) = g(f_j) = y$.

Define the representation $\phi^\star(X) = \ell(X)$ under Assumption 1 (equivalently $\phi^\star(X) = g(F)$, since $Y = g(F) = \ell(X)$ a.s.), i.e., the representation retains only the coarse label and discards all remaining information about the latent mode. Because $Y = \ell(X) = \phi^\star(X)$ a.s. by Assumption 1, $H(Y \mid \phi^\star(X)) = 0$, so $I(\phi^\star(X); Y) = H(Y) - H(Y \mid \phi^\star(X)) = H(Y)$. Since $Y = \ell(X)$ is itself a deterministic function of X, the same identity applied to X gives $H(Y \mid X) = 0$ and hence $I(X; Y) = H(Y)$ as well. Therefore $I(\phi^\star(X); Y) = H(Y) = I(X; Y)$, which shows that $\phi^\star$ preserves all information relevant to predicting Y and is therefore predictively sufficient.

However, whenever $g(f_i) = g(f_j)$, the representation maps both latent modes to the same value. Therefore their latent signatures coincide: $\Phi_{\phi^\star}(f_i) = \Phi_{\phi^\star}(f_j)$. By Definition 1, they belong to the same partition block. Consequently $f_i \sim_{\phi^\star} f_j$. Since the task partition $\Pi^\star$ distinguishes f_i and f_j by assumption, it follows that $\Pi_{\phi^\star} \neq \Pi^\star$. Therefore a representation may be perfectly predictive while failing to preserve the latent task structure. □

Corollary 3 (Accuracy Does Not Certify Structure)**.** *Let $\Pi^\star$ be a task partition on latent modes F and let $Y = g(F)$ be a non-injective coarse labeling satisfying Assumption 1. Then there exists a representation ϕ such that $R_{\mathrm{cls}}(\phi) = R^\star_{\mathrm{Bayes}}$, where $R^\star_{\mathrm{Bayes}}$ is the Bayes-optimal classification risk for predicting Y, yet $VI(\Pi_\phi, \Pi^\star) > 0$. Consequently, achieving Bayes-optimal classification performance does not certify recovery of the latent task structure.*

Proof. Since g is non-injective, there exist latent modes $f_i \neq f_j$ satisfying $g(f_i) = g(f_j) = y$.

Consider the representation $\phi(X) = \ell(X) = g(F)$ (Assumption 1). By construction, Y is a deterministic function of $\phi(X)$, namely $Y = \phi(X)$. Therefore the Bayes classifier operating on $\phi(X)$ predicts Y perfectly and attains Bayes-optimal risk,

$$R_{\mathrm{cls}}(\phi) = R^\star_{\mathrm{Bayes}}.$$

However, because ϕ retains only the coarse label, all latent modes sharing the same label collapse into the same representation.

In particular, $\phi(f_i) = \phi(f_j)$. Hence their latent signatures coincide, $\Phi_\phi(f_i) = \Phi_\phi(f_j)$, and therefore $f_i \sim_\phi f_j$.

The induced partition Π_ϕ thus merges at least one pair of latent modes that are distinct in the task partition $\Pi^\star$. Therefore $\Pi_\phi \neq \Pi^\star$. Since variation of information is a metric on the finite partition space, $VI(\Pi_\phi, \Pi^\star) = 0 \iff \Pi_\phi = \Pi^\star$. Because the partitions differ, $VI(\Pi_\phi, \Pi^\star) > 0$. Thus a representation may achieve Bayes-optimal classification performance while failing to recover the underlying task structure. □

D.1 Relationship to Latent-State Abstractions

Theorem 11 (Partition Recovery as Abstraction Recovery)**.** *Let F be a latent state space and let $\Pi^\star = \{C_1, \ldots, C_k\}$ be the task partition induced by confusability. Define the abstraction map $\alpha : F \to \{1, \ldots, k\}$ by assigning each latent state to its partition block.*

Then:

1. *α is sufficient for task structure;*

2. *every task-sufficient representation factors through α, and is related to it by an injective relabeling of blocks;*

3. *consequently, α is the unique task-sufficient abstraction up to relabeling: no strictly coarser map is task-sufficient, and no other task-sufficient map is strictly finer than a relabeling of α.*

Hence recovering $\Pi^\star$ is equivalent to recovering the (essentially unique) task-sufficient latent abstraction. This is the partition-recovery analogue of minimal sufficient state abstraction in MDPs (Li et al., 2006) and of exact bisimulation (Ferns et al., 2004), specialized to a distribution-based rather than a dynamics-based notion of equivalence.

Proof. Let α assign each latent mode to its partition block. By construction, two latent modes lie in the same block iff they are task-confusable. Therefore all task-relevant distinctions are preserved by α, proving sufficiency.

Now let $\psi : F \to Z$ be any task-sufficient representation. Task sufficiency implies $\Pi_\psi = \Pi^\star$. Hence whenever $\alpha(f_i) = \alpha(f_j)$, we also have $\psi(f_i) = \psi(f_j)$ up to an injective relabeling of blocks.

Therefore $\exists$ a mapping h such that $\psi = h \circ \alpha$. Thus every task-sufficient representation factors through the partition abstraction, proving (2). For (3): since task-structure sufficiency is defined by exact partition equality ($\Pi_\psi = \Pi^\star$), any task-sufficient ψ already induces exactly $\Pi^\star$ – there is no family of task-sufficient abstractions at varying granularity to range over, only representations of $\Pi^\star$ that differ by how blocks are labeled. Any strictly coarser map merges at least two blocks of $\Pi^\star$ and so destroys a distinction $\Pi^\star$ requires, hence is not task-sufficient; and by (2) no task-sufficient map is strictly finer than α composed with a relabeling. This is what "coarsest" and "minimal" amount to here: uniqueness up to relabeling, not optimization over a genuine coarse/fine spectrum of sufficient abstractions (the latter is available in the state-abstraction literature (Li et al., 2006) under weaker, approximate notions of sufficiency than the exact equality used throughout this paper).

□

E Partition Information Theory

This section develops an information-theoretic interpretation of task-structure capability. The central observation is that latent mode characterization depends not on the specific representation itself but on the latent partition induced by that representation. Partition information therefore becomes a property of partitions rather than random variables.

E.1 Partition Complexity

Let $\Pi = \{C_1, \ldots, C_k\}$ denote a partition of fault set F. Each block C_i contains fault modes that are indistinguishable.

The number of partition blocks therefore measures the number of distinguishable fault classes.

Definition 13 (Partition Information)**.** *Let $|\Pi|$ denote the number of partition blocks.*

The partition information of a partition is $I_D(\Pi) = \log |\Pi|$.

The logarithm is chosen so that information is additive under independent partition refinements.

E.2 Interpretation

The quantity $I_D(\Pi)$ measures the amount of task-structure distinction represented by a realization.

Several special cases are instructive.

1. If every fault belongs to a single fibre, $|\Pi| = 1$, then $I_D(\Pi) = 0$.
   No task-structure distinctions are possible.
2. If every fault occupies its own fibre, $|\Pi| = |F|$, then $I_D(\Pi) = \log |F|$.
   All faults are distinguishable.

3. Intermediate values correspond to partial diagnosability.

Thus partition information quantifies task-structure resolution.

E.3 Refinement Order

Definition 14 (Partition Refinement)**.** *Let* Π_1 *and* Π_2 *be partitions of the same fault set.*

We write $\Pi_1 \preceq \Pi_2$ *if every block of* Π_1 *is contained in a block of* Π_2.

In this case Π_1 *is finer than* Π_2.

Partition refinement induces a partial order on task-structure realizations.

E.4 Monotonicity of Partition Information

Partition information behaves monotonically with respect to refinement.

Theorem 12 (Partition Information Monotonicity)**.** *If* $\Pi_1 \preceq \Pi_2$, *then* $I_D(\Pi_1) \geq I_D(\Pi_2)$.

Proof. Because Π_1 is finer than Π_2, every block of Π_2 contains one or more blocks of Π_1. Therefore $|\Pi_1| \geq |\Pi_2|$.

Since the logarithm is monotone increasing, $\log |\Pi_1| \geq \log |\Pi_2|$. Hence $I_D(\Pi_1) \geq I_D(\Pi_2)$. □

Thus partition refinement always increases partition information.

E.5 Partition Information Loss

The effect of realization simplification can now be quantified.

Definition 15 (Partition Information Loss)**.** *Let* $\Pi_1 \preceq \Pi_2$. *The information loss from* Π_1 *to* Π_2 *is*

$$L_D(\Pi_1, \Pi_2) = I_D(\Pi_1) - I_D(\Pi_2).$$

The quantity L_D measures the reduction in task-structure capability induced by sensor removal, weaker fault models, reduced training data, or reduced representation capacity.

E.6 Information Loss Under Resource Reduction

Let $\Lambda = (S, K, D, H)$ denote a realization-construction structure.

Theorem 13 (Resource Monotonicity)**.** *Suppose* $\Lambda_2 \subseteq \Lambda_1$. *Then* $I_D(\Lambda_2) \leq I_D(\Lambda_1)$.

Proof. By the fibre-merger theorem, reducing realization-construction information produces a coarsening of the induced partition. Therefore $\Pi(\Lambda_1) \preceq \Pi(\Lambda_2)$.

Applying the Partition Information Monotonicity theorem above yields $I_D(\Lambda_1) \geq I_D(\Lambda_2)$. The claim follows.

□

Thus loss of realization-construction resources cannot increase partition information.

E.7 Partition Entropy

The previous definitions assume all faults are equally likely. To incorporate fault probabilities, let $p_i = \Pr(C_i)$ denote the probability mass associated with partition block C_i.

Definition 16 (Partition Entropy)**.** *The partition entropy of a partition is*

$$H_D(\Pi) = -\sum_{i=1}^{|\Pi|} p_i \log p_i.$$

Partition entropy measures uncertainty over distinguishable task-structure outcomes.

Proposition 3. *For fixed fault probabilities, $H_D(\Pi)$ is maximized when all partition blocks are equally probable.*

Proof. This is the classical maximum-entropy property of Shannon entropy.

□

E.8 Partition Mutual Information

Partition structure also permits a mutual-information interpretation. Let Π_O and Π_Z denote the observation-space and latent-space partitions.

Definition 17 (Partition Mutual Information)**.** *Define*

$$I_D(\Pi_O; \Pi_Z) = H_D(\Pi_O) - H_D(\Pi_O|\Pi_Z).$$

This quantity measures how much task-structure distinction present in the observation space survives in the latent representation.

E.9 Task-Structure Sufficiency Revisited

Task-structure sufficiency admits an information-theoretic characterization.

Theorem 14 (Information-Theoretic Sufficiency)**.** *A representation ϕ is task sufficient if and only if*

$$I_D(\Pi_O; \Pi_Z) = H_D(\Pi_O).$$

Proof. Task-structure sufficiency implies $\Pi_O = \Pi_Z$. Hence $H_D(\Pi_O|\Pi_Z) = 0$. Substituting into the definition of partition mutual information gives $I_D(\Pi_O; \Pi_Z) = H_D(\Pi_O)$.

Conversely, if $I_D(\Pi_O; \Pi_Z) = H_D(\Pi_O)$, then $H_D(\Pi_O|\Pi_Z) = 0$. Therefore the latent partition determines the observation partition exactly. Hence $\Pi_O = \Pi_Z$, which is task-structure sufficiency. □

E.10 Connection to Information Bottlenecks

Classical information bottleneck methods seek representations maximizing $I(Z;Y)$ while minimizing $I(Z;X)$. The realization-theoretic framework suggests an alternative objective.

Rather than preserving labels, the objective is to preserve partitions.

This leads naturally to the constrained optimization

$$\max\ I_D(\Pi_O; \Pi_Z) \text{ subject to } \dim(Z) \leq d.$$

In contrast to conventional bottleneck methods, the objective explicitly preserves task-structure distinguishability rather than generic predictive information.

E.11 Interpretation

The results of this section identify partition information as a partition-level quantity. Sensors, fault models, training data, latent dimension, and inductive bias influence diagnosability only through the partitions they induce. Consequently, realization-construction resources can be compared through I_D, rather than through architecture-specific measures. This observation provides the information-theoretic foundation for the experimental studies introduced in the following section.

F Partition Information Bottleneck

The previous section established that task-structure capability is determined by the partition induced on the fault space. This observation suggests a realization-theoretic reinterpretation of information bottleneck methods. Classical information bottleneck theory seeks a representation $Z = \phi(X)$ that preserves information about a target variable Y while discarding irrelevant information about the input X.

Here, the primary object of interest is not a label variable but the partition. Consequently, the relevant information-preservation objective concerns fault distinguishability rather than prediction accuracy.

F.1 Classical Information Bottleneck

Given observations X, representation Z, and target variable Y, the classical information bottleneck solves

$$\max_{\phi} \Big(I(Z;Y) - \beta I(Z;X) \Big).$$

The resulting representation seeks to preserve predictive information while minimizing representation complexity. From a partition perspective this objective possesses a fundamental limitation. Preservation of predictive information does not necessarily imply preservation of task structure.

F.2 Task-Structure Counterexample

Suppose faults f_1 and f_2 possess identical labels but correspond to distinct physical failure modes. A representation may preserve perfect prediction performance $I(Z;Y) = I(X;Y)$ while merging f_1 and f_2 into a common confusable fibre. In such a situation predictive sufficiency is preserved but task-structure sufficiency fails. This demonstrates that task structure is generally finer than standard predictive information.

F.3 Partition Information Bottleneck

We therefore introduce a bottleneck objective defined directly on fault partitions.

Definition 18 (Partition Bottleneck Objective)**.** *Let Π_O denote the population partition induced by the original observation space and let Π_Z denote the partition induced by the learned representation.*

The partition information bottleneck problem is

$$\max_{\phi} \Big(I_D(\Pi_O; \Pi_Z) - \beta C(\phi) \Big),$$

where $C(\phi)$ is a measure of representation complexity.

The objective differs fundamentally from the classical information bottleneck. Rather than preserving prediction information, it seeks to preserve task-structure distinguishability.

F.4 Task-Structure Sufficiency as Bottleneck Optimality

Theorem 15 (Sufficient Bottleneck Optimum)**.** *Suppose $\phi^\star$ is task sufficient. Then $\phi^\star$ maximizes $I_D(\Pi_O; \Pi_Z)$.*

Proof. By task-structure sufficiency, $\Pi_Z = \Pi_O$. Applying the Information-Theoretic Sufficiency theorem of Appendix E gives $I_D(\Pi_O; \Pi_Z) = H_D(\Pi_O)$.

The mutual information between two partitions cannot exceed the entropy of either partition. Hence $H_D(\Pi_O)$ is the maximum attainable value. Therefore $\phi^\star$ is an optimizer. □

F.5 Partition Compression

A desirable representation should preserve partition information while minimizing dimensionality. This leads to the constrained optimization problem

$$\min_{\phi} \dim(Z)$$

subject to

$$I_D(\Pi_O; \Pi_Z) = H_D(\Pi_O).$$

This formulation characterizes minimal task sufficient representations.

F.6 Partition Distortion

Partition information loss can be interpreted as partition distortion.

Definition 19 (Partition Distortion)**.** *The distortion induced by a representation is*

$$\Delta_D(\phi) = H_D(\Pi_O) - I_D(\Pi_O; \Pi_Z).$$

A task sufficient representation satisfies $\Delta_D(\phi) = 0$. Positive values indicate fibre mergers and therefore loss of task-structure resolution.

F.7 Interpretation

The partition bottleneck perspective shifts the objective of representation learning. Conventional objectives seek low reconstruction error, classification accuracy, or predictive information. The present framework instead seeks preservation of partitions. Thus the target object becomes Π_O rather than a label variable.

G Proofs

Full proofs of the theorems and corollaries stated in Sections 3–6, given in the order they appear there.

G.1 Proof of the Partition Consistency Theorem

Proof. Let

$$\eta_n = \sup_{f \in F} d(\widehat{\Phi}_n(f), \Phi(f)).$$

By assumption (A3), $\eta_n \xrightarrow{a.s.} 0$. Let $\tau = (\varepsilon + \gamma)/2$ denote the decision threshold introduced above (Definition of the Decision Threshold), and fix

$$\delta = \frac{\gamma - \varepsilon}{4} > 0,$$

which is well defined since (A2) now reads $\gamma > \varepsilon$. Since $\eta_n \to 0$, there exists N such that $\eta_n < \delta$ for all $n \geq N$.

Consider two cases.

Case 1: f_i and f_j belong to the same population partition block. Then $d(\Phi(f_i), \Phi(f_j)) \leq \varepsilon$. Applying the triangle inequality,

$$d(\widehat{\Phi}_n(f_i), \widehat{\Phi}_n(f_j)) \leq \varepsilon + 2\eta_n < \varepsilon + 2\delta = \varepsilon + \frac{\gamma - \varepsilon}{2} = \tau.$$

Since $d(\widehat{\Phi}_n(f_i), \widehat{\Phi}_n(f_j)) < \tau$, the pair is empirically confusable and remains in the same empirical block.

Case 2: f_i and f_j belong to different population partition blocks. Then $d(\Phi(f_i), \Phi(f_j)) \geq \gamma$.

Using the triangle inequality,

$$d(\widehat{\Phi}_n(f_i), \widehat{\Phi}_n(f_j)) \geq \gamma - 2\eta_n > \gamma - 2\delta = \gamma - \frac{\gamma - \varepsilon}{2} = \tau.$$

Since $d(\widehat{\Phi}_n(f_i), \widehat{\Phi}_n(f_j)) > \tau$, the pair is not empirically confusable and thus remains separated.

Since F is finite, all pairwise relationships are eventually preserved. Therefore $\widehat{\Pi}_n = \Pi_\varepsilon^\star$ for sufficiently large n. Consequently $\Pr(\widehat{\Pi}_n = \Pi_\varepsilon^\star) \to 1$.

□

G.2 Proof of the Approximate Partition Recovery Theorem

Proof. Since the fault set is finite, there are only

$$M = \frac{m(m-1)}{2}$$

distinct fault pairs.

Consider any pair (f_i, f_j). By assumption (A3),

$$d\Big(\widehat{\Phi}_n(f_i), \Phi(f_i)\Big) \xrightarrow{P} 0$$

and

$$d\Big(\widehat{\Phi}_n(f_j), \Phi(f_j)\Big) \xrightarrow{P} 0.$$

Using continuity of the metric together with the triangle inequality,

$$\Big|d(\widehat{\Phi}_n(f_i), \widehat{\Phi}_n(f_j)) - d(\Phi(f_i), \Phi(f_j))\Big| \xrightarrow{P} 0.$$

Hence every empirical pairwise distance converges in probability to its population counterpart.

Fix a pair (f_i, f_j) with $i \neq j$. By (A4), $d(\Phi(f_i), \Phi(f_j)) \neq \varepsilon$; let

$$\delta_{ij} = |d(\Phi(f_i), \Phi(f_j)) - \varepsilon| > 0.$$

Since $d(\widehat{\Phi}_n(f_i), \widehat{\Phi}_n(f_j)) \xrightarrow{P} d(\Phi(f_i), \Phi(f_j))$,

$$\Pr\Big(\big|d(\widehat{\Phi}_n(f_i), \widehat{\Phi}_n(f_j)) - d(\Phi(f_i), \Phi(f_j))\big| < \delta_{ij}\Big) \to 1.$$

On this event, $d(\widehat{\Phi}_n(f_i), \widehat{\Phi}_n(f_j))$ lies strictly on the same side of ε as $d(\Phi(f_i), \Phi(f_j))$ does, since the two are within δ_{ij} of each other while ε is at distance exactly δ_{ij} from the latter. Hence, on this event, $\hat{c}_{ij} = c_{ij}^\star$, so

$$\Pr(\hat{c}_{ij} = c_{ij}^\star) \to 1.$$

This is the step where continuity of $d(\cdot, \cdot)$ and convergence of the raw distance are not, by themselves, sufficient: the confusability indicator is discontinuous exactly at ε, and (A4) is what keeps $d(\Phi(f_i), \Phi(f_j))$ off that discontinuity so the argument above goes through.

Because there are finitely many fault pairs, a union bound over the pairwise events $\{\hat{c}_{ij} = c^\star_{ij}\}$ implies

$$\Pr\left(\widehat{C}_n = C^\star\right) \to 1.$$

Consequently, the pairwise confusability relations induced by $\widehat{\Phi}_n$ and Φ agree with probability approaching one.

Since partitions are uniquely determined by their confusability matrices,

$$\Pr\left(\widehat{\Pi}_n = \Pi^\star_\varepsilon\right) \to 1.$$

Finally, variation of information is a metric on the space of finite partitions and satisfies

$$VI(\Pi, \Pi') = 0 \quad \Longleftrightarrow \quad \Pi = \Pi'.$$

Therefore

$$VI\left(\widehat{\Pi}_n, \Pi^\star_\varepsilon\right) \xrightarrow{P} 0.$$

Thus

$$\Delta_\Pi(n) \xrightarrow{P} 0.$$

□

G.3 Proof of the Finite-Sample Partition Recovery Theorem

Proof. The concentration assumption implies that, with probability at least $1 - \delta$, every empirical signature lies within $\psi(n, \delta)$ of its population signature.

Condition on this event. For a same-block pair, the triangle inequality gives $d(\widehat{\Phi}_n(f_i), \widehat{\Phi}_n(f_j)) \leq \varepsilon + 2\psi(n, \delta) < \varepsilon + (\gamma - \varepsilon)/2 = \tau$, so the pair is empirically confusable at threshold τ. For a different-block pair, $d(\widehat{\Phi}_n(f_i), \widehat{\Phi}_n(f_j)) \geq \gamma - 2\psi(n, \delta) > \gamma - (\gamma - \varepsilon)/2 = \tau$, so the pair is not. This is exactly the argument of the Partition Consistency Theorem with $\psi(n, \delta)$ in place of η_n.

Therefore the empirical partition equals the population partition whenever the concentration event holds. The concentration event occurs with probability at least $1 - \delta$. □

G.4 Proof of the Sample Complexity Bound Theorem

Proof. Partition recovery requires

$$\psi(n, \delta) < \frac{\gamma - \varepsilon}{4}.$$

Substituting the assumed concentration rate yields

$$\sqrt{\frac{C(H) + \log(1/\delta)}{n}} < c(\gamma - \varepsilon)$$

for some constant $c > 0$.

Squaring both sides and rearranging gives

$$n = \Omega\left(\frac{C(H) + \log(1/\delta)}{(\gamma - \varepsilon)^2}\right).$$

□

G.5 Proof of the Imbalance-Limited Recovery Corollary

Proof. Apply a union bound over the m faults to the per-fault concentration events, each at confidence level δ/m, giving $\Pr(\sup_f d(\widehat{\Phi}_{n_f}(f), \Phi(f)) \leq \max_f \psi(n_f, \delta/m)) \geq 1 - \delta$. Since $\psi(n_f, \delta/m)$ is decreasing in n_f, the maximum over f is attained at $n_f = \min_{f'} n_{f'}$. Substituting this into the Finite-Sample Partition Recovery Theorem's requirement $\psi < (\gamma - \varepsilon)/4$ and solving for $\min_f n_f$, exactly as n was solved for in the proof of the Sample Complexity Bound theorem above, gives the stated rate. □

H Full Empirical Analysis

All experiments we conducted on an Intel Core i7-14700K paired with an NVIDIA GeForce RTX 4070 Super, 32GB of DDR5 RAM, and a 2TB NVMe SSD.

Full specification of the benchmark systems, realization classes, experimental factors and levels, experimental design, partition-estimation calibration protocol, evaluation metrics, formal hypothesis statements, statistical analysis, and complete results (figures, tables, and ablations) summarized in Section 8. Appendix H.1 maps each hypothesis to the specific evidence for it below.

H.1 Hypothesis-to-Evidence Map

Table 1: Where each hypothesis (Section 8.2) is tested. Formal statements are in Appendix H.10.

H	Evidence	Location
H1	DPT and fine-grained supervision recover partitions equivalent to raw channels; coarse supervision does not (synthetic and real bearing-vibration data)	Fig. 5, 6; Sec. 8
H2	Follows definitionally once H1 holds: detectability and isolability are defined directly in terms of the recovered partition (Appendix E), not measured by a separate statistic	Appendix E
H3	Coarse-label supervision (less fault-specific knowledge) yields a strictly coarser partition than fine-grained supervision or DPT, on both benchmarks; the same relation holds between strong- and weak-fault symbolic realizations in the Design 1 factorial	Fig. 5; Sec. 8, H.11
H4	Sensor field/placement changes recovered partition granularity (Fig. 1a); the same coarse-vs-fine effect as H3	Fig. 1, 5
H5	A compressed, DPT-trained embedding (dimension 3) recovers the true partition at far smaller n than the full ambient space it was trained from	Fig. 3; Sec. 7.1
H6	Purity-preserved, completeness-degrading pattern under noise, imbalance ($\min_f n_f$ rate), and reduced labels; oracle-beats-bad-supervision reversal; replicated on a second, structurally unrelated benchmark	Fig. 1, 4

H.2 Benchmark Systems

Two benchmark systems are employed.

H.2.1 MULTI-TANK SYSTEM

The multi-tank process provides

- continuous dynamics,
- observer-based realizations,
- consistency-based realizations,
- geometric realizations,
- learned latent realizations.

Fault classes include:

- leakage faults,
- blockage faults,
- pump faults,
- sensor faults.

The primary use of this benchmark is visualization of partition geometry and latent manifolds.

H.2.2 IEEE 7-BUS FEEDER

The feeder benchmark provides a system with known confusable classes under multiple sensor placements.

Fault classes include:

- line faults,
- injection faults,
- sensor faults.

The benchmark is particularly suitable for testing fibre mergers predicted by the realization-construction theory.

H.3 REALIZATION CLASSES

Five realization classes are considered.

1. Strong-fault consistency-based realization.
2. Weak-fault consistency-based realization.
3. Learned supervised realization.
4. Learned anomaly-detection realization.
5. Diagnostic-partition-trained (DPT) realization.

The first four are baselines against which the paper's own proposal is measured. The fifth is trained directly against the partition bottleneck loss L_{diag} of Section 7 and is the only class that optimizes for partition preservation rather than for a proxy objective (label prediction or anomaly score); it is the class that empirically tests the Global Optimum and Task-Structure Sufficiency theorem of that section.

H.3.1 STRONG-FAULT REALIZATIONS

Strong realizations (Feldman et al., 2010) employ $K = \{M_N, M_{f_1}, \ldots, M_{f_m}\}$. Individual fault-mode behaviour is explicitly modeled.

H.3.2 Weak-Fault Realizations

Weak realizations (De Kleer et al., 1992) employ $K = \{M_N\}$ only. Specific fault behaviour is not provided.

H.3.3 Supervised Learned Realizations

The supervised learning setting employs fault-specific labels $y \in \{f_1, \ldots, f_m\}$.

H.3.4 Anomaly-Detection Realizations

The anomaly setting employs labels $y \in \{\text{normal}, \text{abnormal}\}$. This setting provides the learning-theoretic analogue of weak-fault diagnosis (Reiter, 1987).

H.3.5 Diagnostic-Partition-Trained (DPT) Realizations

The DPT setting trains ϕ_θ using the partition bottleneck loss $L_{\text{diag}} = \widetilde{VI} + \beta\, C(Z)$ of Section 7, against a reference partition $P^\star$ obtained either from the exact diagnosability realization of the companion Diagnostic System Specification (DSS) framework, where available for the benchmark and sensor configuration in question, or from a strong-fault consistency-based realization otherwise (Algorithm 1). No fault-specific class labels are used; only the pairwise reference partition matrix $P^\star$ is required. This setting directly targets the object the theory identifies as the correct one, and is the class expected to close the task-structure counterexample of Appendix F: a supervised realization can preserve label accuracy while merging two same-label, physically distinct fault fibres, whereas a DPT realization trained against a reference partition that separates them is not subject to this failure mode by construction.

H.4 Experimental Factors

The study examines seven realization-construction and data-condition resources. Factors A and B (sensors, fault knowledge) apply to all five realization classes, since sensor placement and fault-model strength are meaningful for both symbolic and learned realizations. Factors C and D (training-data fraction, latent dimension) and the new Factors E–G below apply only to the three learned realization classes (supervised, anomaly-detection, DPT): the strong- and weak-fault consistency-based realizations have no training data, latent representation, class imbalance, missing labels, or observation noise model of their own within this framework, and are held fixed as reference points rather than swept.

H.4.1 Factor A: Sensor Information

Sensor availability is varied independently of all other factors.

For the feeder benchmark:

$$M_1 = \{0, 1, 2, 3, 4, 5, 6\}$$

(full placement),

$$M_2 = \{3, 5, 6\}$$

(diagnosability-equivalent sparse placement): $M_3 = \{3, 6\}$, and $M_4 = \{1, 2, 4\}$.

For the tank benchmark:

- full sensing,
- reduced sensing,
- minimum sensing.

H.4.2 Factor B: Fault Knowledge

Two levels are considered: K_{strong} and K_{weak}. This factor tests the Knowledge-Induced Refinement Theorem.

H.4.3 Factor C: Training Data

Training fractions:

$$100\%, 75\%, 50\%, 25\%, 10\%.$$

This factor evaluates the sample-complexity predictions of Section 6.

H.4.4 Factor D: Latent Dimension

Representation dimensions

$$d_z = 2, 4, 8, 16, 32.$$

This factor evaluates the existence of a task sufficient latent dimension.

H.4.5 Factor E: Class Imbalance

Per-fault sample counts n_f are drawn according to an imbalance ratio $r \in \{1, 10, 100\}$, where r is the ratio between the most- and least-sampled fault modes under a fixed total budget $n = \sum_f n_f$ (a log-uniform allocation across faults is used to realize each ratio). This factor evaluates the Imbalance-Limited Recovery corollary of Section 6.

H.4.6 Factor F: Label Availability

Fault-mode labels are retained for a fraction $\ell \in \{100\%, 50\%, 10\%, 0\%\}$ of training observations; at $\ell = 0\%$ the supervised and DPT realizations receive only unlabeled latent samples and must be paired with an unsupervised clustering step prior to matching clusters to reference fault identities (a remark in Appendix I sketches this two-stage estimator; it is not treated with the same theoretical completeness as the labeled setting and is included here as an empirical probe of that gap).

H.4.7 Factor G: Observation Noise

Additive sensor noise is scaled by $\sigma \in \{0.01, 0.1, 0.5, 1.0\}$ (in units of the nominal signal range for each benchmark). Per the distinction drawn in Section I.3, this factor is designed to probe margin-shrinking noise (its effect on the population quantity γ, estimated via the pilot calibration of Section H.6) separately from the estimation-noise effect already probed by Factor C.

H.5 Experimental Design

A single factorial design crossing all seven factors with all five realization classes is not well defined, since Factors C–G do not apply to the two symbolic realization classes (Section H.4). The study is instead organized as three targeted designs, each aimed at a distinct subset of the hypotheses of Section 8.2.

- **Design 1 (partition equivalence).** Sensors $\times$ Fault Knowledge, crossed with all five realization classes and both benchmark systems. Benchmark and Realization Class are included as explicit fixed factors rather than reported as separate side-by-side tables, since cross-benchmark generalization is part of what H1 claims. This design tests H1–H4.
- **Design 2 (sufficiency threshold and sample complexity).** Training Fraction $\times$ Latent Dimension, run within the three learned realization classes only (supervised, anomaly-detection, DPT), with Sensors and Fault Knowledge held at their strongest level (M_1, K_{strong}). This design tests H5 and the sample-complexity prediction of the Sample Complexity Bound theorem.

- **Design 3 (robustness to realistic data conditions).** Imbalance Ratio $\times$ Label Availability $\times$ Noise Level (Factors E–G), run within the three learned realization classes only, at the single best Sensors/Fault Knowledge condition identified in Design 1, with fibre purity and completeness (Section H.7) as the primary metrics and ARI as a secondary summary. This design tests H6 (Section 8.2).

Each machine-learning configuration in every design is repeated using $N = 30$ independent random initializations.

Reported quantities include

- mean,
- standard deviation,
- 95% confidence intervals.

H.6 Partition Estimation

For each realization class:

1. Construct latent signatures $\widehat{\Phi}_n(f)$.
2. Estimate pairwise signature distances
$$d(\widehat{\Phi}_n(f_i), \widehat{\Phi}_n(f_j)).$$
3. **Calibrate the decision threshold.** Estimate $\widehat{\gamma}$ from a large pilot sample, generated independently of the sample used for the reported recovery statistics, and set $\tau = (\varepsilon + \widehat{\gamma})/2$ as in Section 3. This step is performed once per experimental condition (benchmark, sensor placement, fault-knowledge level, and realization class), not once per evaluation run, and the same τ is then applied to all $N = 30$ repetitions of that condition.
4. Construct empirical confusability relations at threshold τ.
5. Construct empirical partitions $\widehat{\Pi}_n$.

The resulting partition forms the primary object of evaluation.

H.7 Primary Evaluation Metrics

H.7.1 Partition Agreement

For partitions Π_1, Π_2, pairwise agreement is

$$PA = \frac{\#\text{ matching fault pairs}}{\#\text{ all fault pairs}}.$$

H.7.2 Adjusted Rand Index

Fibres are treated as clusters.

Partition similarity is measured using $ARI(\Pi_1, \Pi_2)$.

Perfect agreement corresponds to $ARI = 1$.

H.7.3 Variation of Information

Partition distance is measured using $VI(\Pi_1, \Pi_2)$.

Perfect agreement corresponds to $VI = 0$.

H.7.4 Fibre Purity and Completeness

PA, ARI, and VI are aggregate measures and can mask a specific failure mode of interest here: a single minority fault fibre being absorbed into a larger fibre. Treating the blocks of $\widehat{\Pi}_n$ as clusters and the blocks of $\Pi^\star_\varepsilon$ as ground-truth classes, purity of empirical block C_k is

$$\mathrm{purity}(C_k) = \frac{\max_\ell |\{f_i \in C_k : f_i \in [f_\ell]_{\Pi^\star}\}|}{|C_k|},$$

and completeness is the symmetric quantity obtained by exchanging the roles of $\widehat{\Pi}_n$ and $\Pi^\star_\varepsilon$. Equivalently, these correspond to the two conditional-entropy terms $H(\Pi^\star_\varepsilon \mid \widehat{\Pi}_n)$ and $H(\widehat{\Pi}_n \mid \Pi^\star_\varepsilon)$ whose sum constitutes VI; reporting them separately, rather than only their sum, distinguishes false merges (low purity) from spurious fragmentation (low completeness), which the imbalance, label-availability, and noise conditions of the robustness design described later in this section are expected to affect differently.

H.8 Task-Structure Metrics

Detectability:

$$D = \frac{\#\text{ detectable faults}}{\#\text{ faults}}.$$

Isolability:

$$I = \frac{\#\text{ isolable fault pairs}}{\#\text{ fault pairs}}.$$

Mean fibre size:

$$\bar{F} = \frac{1}{|F|}\sum_f |\Phi^{-1}(f)|.$$

Number of partition blocks:

$$|\Pi|.$$

H.9 Partition Information Metrics

- Partition information: $I_D(\Pi) = \log|\Pi|$.
- Partition information loss: $L_D = I_D(\Pi^\star) - I_D(\Pi)$.
- Partition entropy: $H_D(\Pi)$.
- Partition mutual information: $I_D(\Pi_O; \Pi_Z)$.

These quantities evaluate the information-theoretic predictions of Appendix E.

H.10 Hypotheses

H.10.1 H1: Partition Equivalence

Diagnosability-equivalent realizations satisfy $ARI(\Pi_{ML}, \Pi_{Symbolic}) \approx 1$.

H.10.2 H2: Diagnosability Preservation

Partition equivalence implies $D_{ML} = D_{Symbolic}$ and $I_{ML} = I_{Symbolic}$.

H.10.3 H3: Knowledge-Induced Refinement

Strong realizations satisfy $|\Pi_{strong}| > |\Pi_{weak}|$. Equivalently, $I_D(\Pi_{strong}) > I_D(\Pi_{weak})$.

H.10.4 H4: Resource Monotonicity

Reduction of sensor information, fault knowledge, training data, or latent dimension induces fibre mergers.

Therefore I_D decreases monotonically as realization-construction resources are removed.

H.10.5 H5: Task-Structure Sufficiency Threshold

There exists $d^\star$ such that $d_z \geq d^\star$ implies $\Pi_Z = \Pi^\star$.

H.10.6 H6: Robustness to Realistic Data Conditions

Increasing class imbalance and observation noise degrade fibre purity (Section H.7) more strongly than completeness, since both mechanisms preferentially merge an underrepresented or noisy fault into a neighbouring fibre rather than fragmenting a well-sampled one; conversely, reduced label availability degrades completeness more strongly than purity, since unsupervised pre-clustering tends to over-split rather than under-split in the absence of a fault-identity target. The imbalance-conditioned degradation in purity is predicted to follow the rate of the Imbalance-Limited Recovery corollary (Section 6) in $\min_f n_f$ rather than in the total sample size n.

H.11 Statistical Analysis

H.11.1 Partition Recovery

Bootstrap confidence intervals are computed for

$$ARI, \quad VI, \quad PA.$$

The primary success criterion is

$$ARI > 0.95.$$

H.11.2 Factor Effects

Because Factors C–G are undefined for the two symbolic realization classes (Section H.4), Factor Effects are assessed with two separate mixed-effects ANOVA models rather than one model spanning all factors, matching the Design 1/Design 2 split of Section H.5.

Model 1 (Design 1). Response variables

$$ARI, \quad I_D, \quad D, \quad I$$

with fixed effects Sensors, Fault Knowledge, Benchmark, and Realization Class (five levels), and random effect Random Seed. Two- and three-way interactions among the fixed effects are included; given the number of cells, interaction terms are reported with effect sizes and confidence intervals rather than significance alone, since higher-order interactions may be underpowered at $N = 30$ seeds per cell.

Model 2 (Design 2). Response variables

$$ARI, \quad I_D$$

with fixed effects Training Fraction, Latent Dimension, and Realization Class (three learned classes only), and random effect Random Seed.

H.11.3 Design 3: Robustness Analysis

For Design 3, purity and completeness (Section H.7) are the primary response variables, analyzed with a mixed-effects model with fixed effects Imbalance Ratio, Label Availability, Noise Level, and Realization Class (three learned classes only), and random effect Random Seed. The imbalance arm is additionally regressed directly against $\min_f n_f$ (rather than against the imbalance ratio r alone) to test the rate predicted by the Imbalance-Limited Recovery corollary of Section 6.

H.11.4 Partition Margin Analysis

The sample-complexity theory predicts recovery error as a function of partition margin γ.

For each experimental condition, empirical margins are estimated and regressed against R_n.

Predicted relationship:

$$R_n = O\Big((\gamma - \varepsilon)^{-2}\Big).$$

Since γ is itself estimated (see the calibration step described in the Partition Estimation subsection above), this regression is run against the pilot-estimated $\widehat{\gamma}$, not against the γ implicitly used to set τ on the same data used to compute R_n.

H.12 Empirical Validation

H.12.1 Wrong-Quotient Ablation: Full Parameters

Full specification for the wrong-quotient ablation summarized in Section 8.3 (Figure 2).

Benchmark. Enriched two-tank fault set (8 modes), field M_{full} (channels: `h1r, h2r, guard, s1, s2, coupl, inflow`), under which the true partition is 8 singleton fibres (fully diagnosable). Observation noise $\sigma = 0.15$, isotropic Gaussian, as in Section 3.

Reference partitions. $P^\star$: the true 8-singleton partition. $P^\star_{\text{wrong}}$: identical except f_{sensor1} and f_{sensor2} are merged into one block – a deliberate, chosen violation, not one induced by noise or estimation error.

DPT architecture and training. Input dimension 7 (raw M_{full} channels, no nuisance augmentation, unlike the bootstrap experiment of Section 7.1). MLP encoder: $7 \to 16$ (tanh) $\to 4$ (linear embedding). Trained via the pairwise partition-bottleneck / soft partition-matrix loss of Section 7 against $P^\star$ or $P^\star_{\text{wrong}}$ respectively, using a fresh per-fault mini-batch (size 16, resampled with replacement from a fixed 300-per-fault training pool) at each step; Adam, learning rate 0.03, 400 steps, soft-partition temperature annealed linearly $\tau : 1.0 \to 0.1$, ℓ_2 regularization 10^{-3}.

Evaluation. At each eval sample size $n \in \{5, 10, 20, 40, 80, 160, 320, 640\}$: fresh eval samples per fault, plus an independent 200-sample pilot per fault (distinct seed from the eval draw at every n) used to (i) set the RBF bandwidth via the median heuristic and (ii) calibrate the embedding-space decision threshold via the same jump-detection rule used for raw-channel calibration (Section 3), applied in embedding space since no closed-form population margin γ exists there. Recovery is scored against the true partition only, using the PA/ARI/VI/purity/ completeness definitions of Section H.7. 6 independent seeds per condition (fresh training *and* evaluation randomness per seed); means and standard deviations reported in Figure 2.

H.12.2 Design-Choice Ablations: Full Parameters

Full specification for the ablations summarized in Section 8.3 (Table 2).

Loss decomposition (λ_{sep}). The DPT loss of Section 7 sums a single BCE term over all fault pairs. We split this into L_{part} (same-fibre pairs, target $= 1$) and L_{sep} (cross-fibre pairs, target $= 0$), each normalized by its own pair count so the split is well-defined under an imbalanced reference partition, and train against $L_{\text{part}} + \lambda_{\text{sep}} L_{\text{sep}} + \ell_2$-reg. $\lambda_{\text{sep}} = 1$ is the baseline condition under this parametrization; $\lambda_{\text{sep}} = 0$ removes all gradient signal separating cross-fibre embeddings, leaving

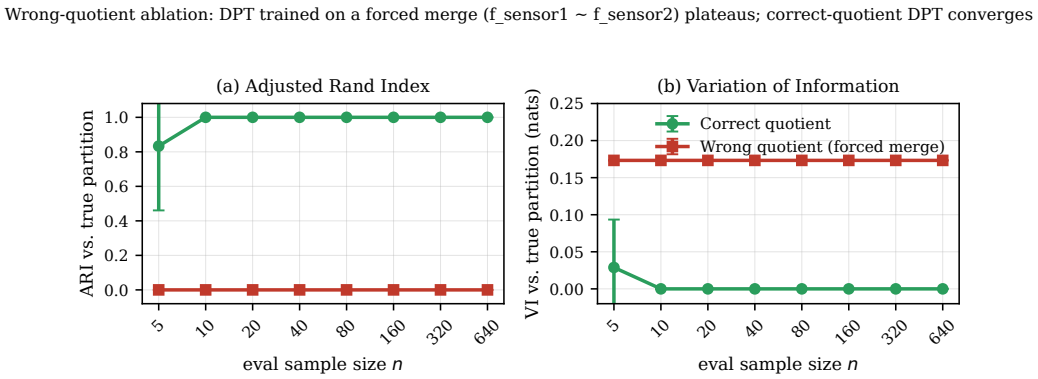

Figure 2: Wrong-quotient ablation on the enriched two-tank benchmark (M_{full} field, all 8 fault modes distinguishable). Two DPT models, identical architecture and training schedule, are trained against (i) the correct reference partition and (ii) a deliberately corrupted reference forcing $f_{\text{sensor1}} \sim f_{\text{sensor2}}$. Both are evaluated at eval sample size n against the true (uncorrupted) partition only. (a) Adjusted Rand Index and (b) Variation of Information, mean $\pm$ 1 s.d. over 6 seeds. The correct-quotient model converges to exact recovery by $n = 10$; the wrong-quotient model is flat across the full range $n \in \{5, \ldots, 640\}$, recovering exactly the corrupted partition at every sample size. ARI collapses to 0 despite only 1 of 28 fault pairs being misassigned, reflecting the known sensitivity of chance-corrected indices with mostly-singleton ground truth (confirmed by partition agreement $= 0.964$); VI is the more informative panel.

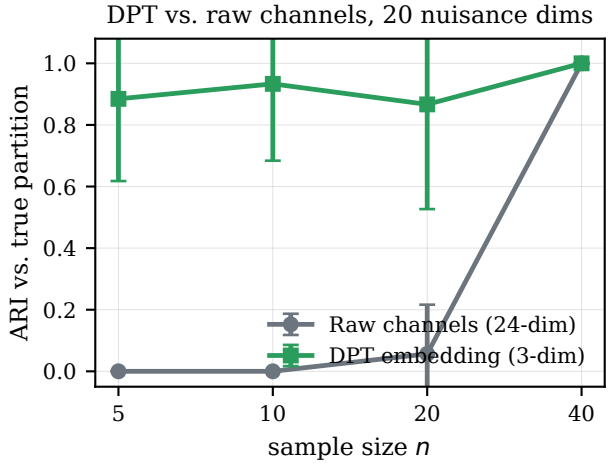

Figure 3: DPT trained against a purely data-bootstrapped reference partition (no symbolic model), versus raw-channel recovery, in a 24-dimensional ambient space (4 true signal channels, 20 nuisance channels). Both scored against the true partition. 15 random seeds per condition.

only the within-fibre attraction term. Benchmark: M_{obs} field (4 raw channels, no nuisance augmentation), $\sigma = 0.15$, symbolic exact reference partition (5 true fibres). Architecture and optimizer identical to Appendix H.12.1's baseline (4$\to$ 16 $\to$ 4 MLP – 4 input channels for M_{obs} here, versus 7 for the M_{full} field used in Appendix H.12.1's wrong-quotient ablation, as noted in Section 8.3 – Adam lr 0.03, 400 steps, τ annealed 1.0 $\to$ 0.1, $\ell_2 = 10^{-3}$, 300-sample training pool per fault). Evaluated at $n \in \{5, 10, 20, 40, 80, 160, 320, 640\}$, 6 seeds per condition; Table 2 reports the $n = 10$ point, with the effect's n-dependence discussed in the main text.

Bootstrap construction. The paper's method (`bootstrap_partition_mmd`, Section 7.1): per-fault MMD distances on a $n_{\text{pilot}} = 500$ pilot, jump- detected confusability threshold, no k required. Naive alternative (`bootstrap_partition_kmeans`): pool the same pilot samples (losing fault

Table 2: Design-choice ablations. PA†: pairwise agreement, reported as the closest available proxy to a generic "accuracy" metric (the paper does not otherwise define classification accuracy for partition recovery). $\hat{\gamma}$: pilot-estimated, embedding-space margin (jump-detected cross-fibre threshold distance; not defined for the bootstrap-quality rows, which operate on raw channels prior to any embedding). Loss-decomposition rows: $n = 10$ per fault, M_{obs} field, symbolic reference, 6 seeds. Bootstrap-construction rows: M_{obs} field with 20 nuisance channels (matching Figure 3), 12 seeds; "bootstrap quality" rows score the reference partition itself against ground truth prior to DPT training, "+DPT, $n = 10$" rows score downstream DPT recovery. See body text for the caveats these numbers carry.

Condition	PA†	ARI	d_{VI}	Purity	Completeness
Loss decomposition ($n = 10$, symbolic reference)					
Full DPT ($\lambda_{\text{sep}} = 1$, baseline)	1.000	1.000	0.000	1.000	1.000
No separation loss ($\lambda_{\text{sep}} = 0$)	0.845	0.356	0.456	0.979	0.792
Bootstrap construction (quality alone, before DPT)					
MMD-threshold (paper's method)	1.000	1.000	0.000	1.000	1.000
Naive k-means ($k = 8$)	0.878	0.526	0.359	0.958	0.844
Bootstrap construction (+DPT, $n = 10$)					
MMD-threshold $\rightarrow$ DPT	1.000	1.000	0.000	1.000	1.000
Naive k-means $\rightarrow$ DPT	0.988	0.967	0.058	0.958	1.000

identity), run vanilla Lloyd's-algorithm k-means with $k = 8$ (the number of fault classes – a naive default, not oracle- tuned or matched against the true partition), then reconstruct a fault-level partition via majority-vote of each fault's own points into the resulting clusters. Benchmark: identical to Figure 3 (M_{obs} field plus 20 nuisance channels, 24-dimensional ambient space, $\sigma = 0.15$). Bootstrap quality is scored against the true partition immediately, before any DPT training. DPT is then trained against each bootstrapped reference (embedding dimension 3, 400 steps, otherwise identical training configuration to the main pipeline) and evaluated at $n \in \{5, 10, 20, 40\}$; Table 2 reports $n = 10$. 12 seeds per condition, run as two independent 6-seed batches with disjoint seed offsets and merged, to increase statistical power on the downstream-recovery comparison specifically.

H.12.3 Temperature-Schedule Ablation: Full Parameters

Full specification and results for the temperature-schedule ablation (Table 3). `train_dpt` already exposes τ_{init}/τ_{final} as parameters, so this is a pure configuration ablation – no new training code, only different schedule endpoints passed to the existing function. Benchmark, architecture, optimizer, and seed protocol identical to the loss-decomposition ablation above (M_{obs} field, symbolic exact reference, $4 \rightarrow 16 \rightarrow 4$ MLP, Adam lr 0.03, 400 steps, $\ell_2 = 10^{-3}$, 300-sample training pool per fault, 6 seeds). Condition A ($\tau : 1.0 \rightarrow 0.1$, the paper's default) is identical in configuration to Appendix H.12.2's loss-ablation baseline and is not separately re-run; conditions B ($\tau \equiv 1.0$), C ($\tau \equiv 0.1$), and D ($\tau : 0.1 \rightarrow 1.0$) hold every other hyperparameter fixed and vary only τ_{init}/τ_{final}. Evaluated at $n \in \{5, 10, 40\}$ (a subset of the full grid, chosen to span the small-n regime where an optimization-heuristic ablation is most informative, at proportionally lower compute cost than the full eight-point sweep used elsewhere).

Responding to the question of what empirical heuristics guard L_{diag}'s non-convex optimization (Remark 1 notwithstanding) against poor local minima, we ablate the soft-partition temperature τ's annealing schedule directly – a pure configuration change, since `train_dpt` already exposes τ_{init} and τ_{final}. Four conditions, same benchmark and seed protocol as the loss-decomposition ablation above (M_{obs} field, symbolic reference, 6 seeds, $n \in \{5, 10, 40\}$): (A) the paper's default anneal $\tau : 1.0 \rightarrow 0.1$; (B) fixed soft, $\tau = 1.0$ throughout; (C) fixed sharp, $\tau = 0.1$ throughout; (D) a reversed schedule, $\tau : 0.1 \rightarrow 1.0$, as a direction control.

Table 3 shows a result more precise than, and partly in tension with, a simple "soft start prevents local minima" story. Fixed-sharp (C) matches the annealed default (A) exactly at all three sample sizes (ARI $= 1.000$, 6/6 seeds each) – a soft start is not demonstrably necessary on this benchmark. What does matter is direction: the reversed schedule (D) is the only condition that degrades meaningfully, and specifically at small n (ARI 0.500 at $n = 5$, with some individual seeds failing completely),

Table 3: Temperature-schedule ablation (M_{obs} field, symbolic reference, 6 seeds; ARI vs. true partition, mean over seeds).

Condition	$n = 5$	$n = 10$	$n = 40$
A: annealed $1.0 \to 0.1$ (paper default)	1.000	1.000	1.000
C: fixed sharp ($\tau = 0.1$ throughout)	1.000	1.000	1.000
B: fixed soft ($\tau = 1.0$ throughout)	0.833	1.000	0.667
D: reversed ($0.1 \to 1.0$)	0.500	0.667	1.000

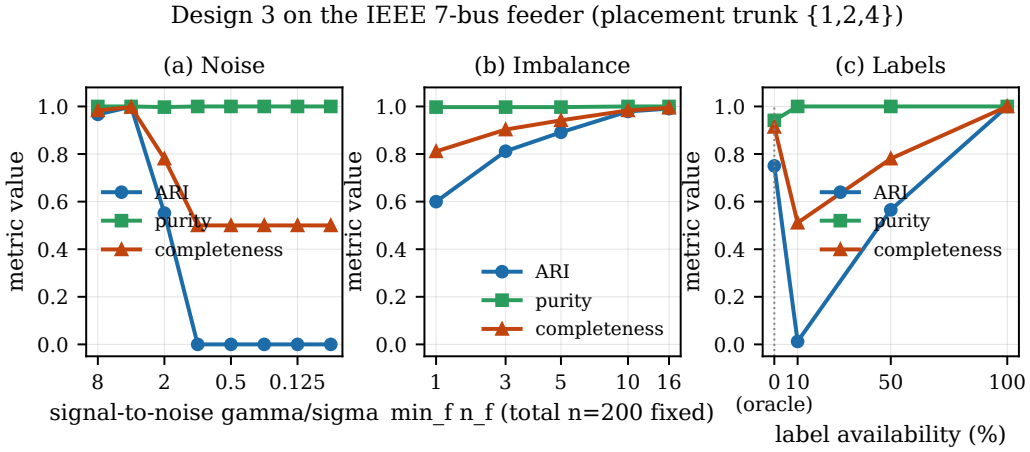

Figure 4: Design 3 (imbalance, label availability, and observation noise) on the IEEE 7-bus feeder benchmark, at the `trunk{1,2,4}` placement. All three panels reproduce the purity-holds/completeness-breaks pattern and the oracle-beats-bad-supervision reversal found on the two-tank benchmark (Figure 1), on a structurally independent system. 30 random seeds per condition.

recovering by $n = 40$. Fixed-soft (B) is intermediate and non-monotonic in n. The evidence therefore supports a narrower, more precise claim than the schedule's existence being what matters: *training should not end soft; how it gets to a sharp temperature (annealed or fixed-sharp from the start) is not distinguishable here.* We note this is a small, well-separated synthetic benchmark (5 true fibres, population margin $\gamma = 0.42$, a few-hundred-step MLP) – precisely the regime where local minima are least likely to bite in the first place, so this ablation does not rule out annealing mattering more on harder problems (more classes, weaker margins, deeper networks); it shows only that it is not doing detectable work here.

H.13 Additional Figures: Cross-Benchmark and Real-Data Replication

Two further figures referenced from Section 8: the cross-benchmark Design 3 confirmation on the IEEE 7-bus feeder, and the learned-realization comparison including the task-structure counterexample (both testing H1, H3, H4, and H6).

Figure 4 repeats the three Design 3 sweeps on the IEEE 7-bus feeder benchmark, at the `trunk{1,2,4}` placement (chosen, as for the tank benchmark, because it induces a genuine merged fibre rather than an all-singleton partition a degenerate estimator could match by construction). All three factors reproduce the same qualitative pattern found on the two-tank benchmark – including the same purity-holds/completeness-breaks signature and the same reversal in which the ℓ=0% oracle outperforms the ℓ=10% semi-supervised estimator – on a structurally unrelated physical system (a radial power feeder rather than a coupled-tank process) using a completely different fault taxonomy (line, injection, and sensor faults rather than pump, valve, and topological faults). This cross- benchmark agreement is evidence that the fragmentation mechanism identified in Section 8 is a property of the statistical estimator, not an artifact of the two-tank benchmark's particular fault structure.

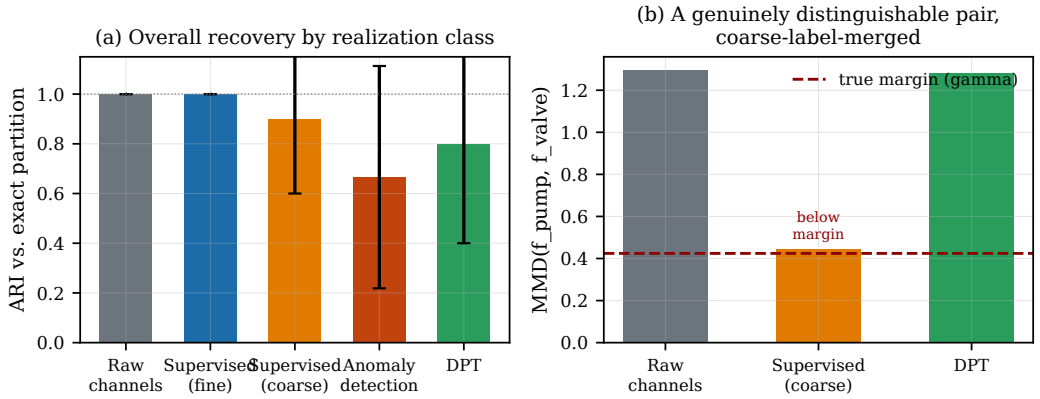

Figure 5: Learned realization classes on the two-tank benchmark (field `M_obs`, embedding dimension 2, held- out test evaluation, 10 random seeds). (a) Overall ARI against the exact partition. (b) The pairwise MMD between `f_pump` and `f_valve` – genuinely distinguishable in raw channels, but sharing a coarse label for the supervised classifier – against the true population margin γ. The coarse-label representation pushes this pair essentially onto the confusability threshold; DPT, which never sees fault-specific labels, does not.

Figure 5 evaluates the four realization classes of Section 8 directly, using a small MLP ϕ_θ (two-layer, embedding dimension 2) trained with standard backpropagation, scored on a held-out test split never seen during training so that a classifier cannot appear to separate genuinely confusable faults by memorizing finite-sample training noise. The DPT realization is trained via the partition bottleneck loss of Section 7 using only the reference partition matrix, no fault-specific labels.

Panel (a) reports overall recovery. The supervised (fine) realization matches the raw-channel baseline; the coarse- label supervised, anomaly-detection, and DPT realizations all show reduced and more variable recovery under this narrow embedding. Panel (b) isolates the mechanism the task-structure counterexample of Appendix F argues for: `f_pump` and `f_valve` are genuinely distinguishable in the raw input (their effect vectors differ), but are given the *same* coarse label ("actuator") for the supervised classifier. A representation trained only to predict that coarse label has no incentive to preserve the fine distinction – and, at this embedding width, does not: the pairwise MMD between the two faults' embeddings falls from 1.29 under raw channels to 0.44, landing essentially exactly on the true population margin $\gamma = 0.42$, meaning classification sufficiency for the coarse task has come at the cost of task-structure sufficiency for the fine one. DPT, trained directly against the true partition rather than a coarse label, keeps the pair well separated (1.28). This effect required deliberately narrowing the embedding: at a wider embedding (dimension 8, not shown) the network had enough spare capacity to preserve the fine distinction without being asked to, and the counterexample did not appear – itself a relevant observation, since it means this particular failure mode is conditional on representation capacity rather than an inevitable consequence of coarse-label training.

DPT's own overall ARI in panel (a) (0.80) is not perfect at this embedding width. We attribute this to the optimization budget of a small network trained for a fixed, short schedule rather than to a conceptual limitation of the objective: DPT correctly protects the specific pair the counterexample turns on (panel b) while its aggregate partition recovery has more variance than the raw-channel or fine-supervised baselines. A longer training schedule and a systematic sweep over embedding width are natural next steps and are not yet included here.

H.13.1 CWRU Real-Data Replication: Setup Detail

Supplementary detail for the CWRU experiment reported in Section 8 (Figure 6), beyond what is stated there.

The drive-end accelerometer channel is used at a fixed motor load (to avoid load acting as a confound alongside fault identity), windowed into non-overlapping 1024-sample segments, giving 117-

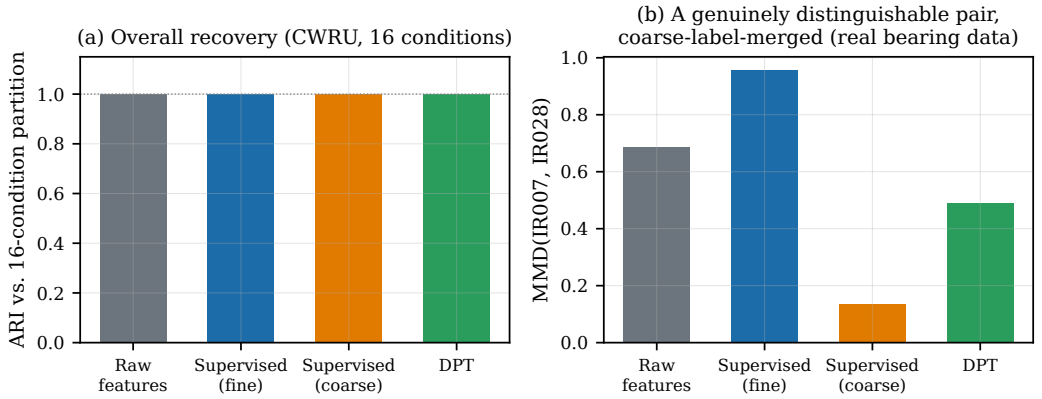

Figure 6: Real-data replication on the CWRU bearing dataset (16 fault conditions, drive-end sensor, fixed load, 1D-CNN ϕ_θ, embedding dimension 4). (a) Overall ARI against the 16-condition ground truth. (b) The pairwise MMD between `IR007` and `IR028` – two inner-race faults of different severity, genuinely distinguishable in raw features but sharing a coarse (location-only) label for the supervised classifier.

238 windows per condition depending on recording length – comparable in scale to the synthetic benchmarks' per-fault sample counts. Each window is standardized (zero mean, unit variance) independently, since raw amplitude scale differs across recordings for reasons (sensor mounting, gain) unrelated to fault identity. A held-out 30% test split per fault condition is used for all evaluation, consistent with the train/test discipline used for the synthetic benchmarks' learned realizations.

The 1D CNN architecture is deliberately small (two convolution layers, kernel sizes 32 and 8 with strides 8 and 4, 8 and 16 channels respectively, global average pooling, linear projection to the embedding), chosen for CPU-only training via the autograd library rather than for accuracy. The five raw-feature statistics used for the pre-experiment distinguishability check and the raw-feature baseline are mean absolute value, standard deviation, RMS, peak amplitude, and kurtosis – standard time-domain vibration-analysis features, not features selected to favor any particular outcome.

I Full Limitations Discussion

Full discussion of the framework's limitations and threats to validity.

The results of this paper establish a realization-theoretic and statistical framework for partition recovery. As with all theoretical frameworks, the conclusions depend on a collection of assumptions and scope restrictions. This section identifies the principal limitations and threats to validity.

I.1 Approximate Rather than Exact Partitions

A central challenge of machine-learning partition discovery is that latent representations typically inhabit continuous, high-dimensional spaces. Consequently, exact equality of latent signature distributions, $\Phi(f_i) = \Phi(f_j)$, is generally unobservable.

The framework therefore replaces exact equality by ε-confusability defined using a signature metric.

The resulting partitions depend upon:

- the choice of metric;
- the selected threshold ε;
- the quality of distribution estimation.

Different choices may induce different partition structures.

Accordingly, empirical partitions should be interpreted as estimated partitions rather than exact representations of population diagnosability.

I.2 Scalability of the Pairwise Matrix ($O(m^2)$ Bottleneck)

DPT relies on a pairwise matrix P_{ij} computed across all fault modes F (Section 8). If the number of latent modes m becomes very large (e.g., thousands of fine-grained classes in large-scale classification or open-world recognition), computing and optimizing the full $m \times m$ pairwise distance matrix scales as $\mathcal{O}(m^2)$. To address scaling to large m, we can apply the standard mitigation (subsampling pairs per step, as contrastive/metric-learning methods do).

I.3 Dependence on the Partition Margin

The consistency and sample-complexity results rely on the existence of a positive partition margin $\gamma > \varepsilon$. This assumption excludes situations in which distinct fault signatures become arbitrarily close.

In practical systems, particularly under severe noise, multiple faults may produce nearly identical latent distributions. When $\gamma \approx \varepsilon$, small estimation errors may significantly alter the resulting partition.

Consequently, partition recovery may require substantially larger datasets than predicted by the asymptotic theory.

It is worth distinguishing two mechanisms by which noise degrades partition recovery, since they call for different remedies. *Estimation noise* slows the convergence $\widehat{\Phi}_n(f) \to \Phi(f)$ captured by assumption (A3) above, for a fixed population margin γ; this is cured by more data, and is what the sample-complexity bounds quantify. *Margin-shrinking noise*, by contrast, is noise propagated through the observation process and the representation ϕ that reduces the population separation γ itself — the latent signatures $\Phi(f_i)$ and $\Phi(f_j)$ genuinely move closer together, independently of sample size. No amount of additional data cures this; it requires better sensors, a better-conditioned representation, or an acceptance that the affected fault pair is not diagnosable under the given construction structure Λ. The classical detectability/isolability degradation studied in observer-based fault detection and isolation is, in this framework, an instance of margin-shrinking noise. The experimental design in Section 8 separates these two regimes by holding sensor noise fixed while varying training-sample size (to probe estimation noise) and, separately, sweeping sensor noise level directly (to probe margin shrinkage).

I.4 Finite Partition Sets

Throughout the analysis the partition set F is assumed finite. Many practical systems satisfy this assumption because partition analysis is ultimately performed over enumerated sets, e.g., fault modes. However, some applications admit continuous partition parameterizations – examples in diagnosis include fault magnitudes, degradation levels, and drifting sensor biases.

The *definitions* in Sections 3–2 already extend verbatim to continuous F: $\Pi_\varepsilon^\star$ and γ are defined via the confusability relation $d(\Phi(f_i), \Phi(f_j)) \le \varepsilon$ and its transitive closure, which requires no finiteness assumption. What breaks is the *sample-complexity theory* (Theorem 6, Corollary 2), which implicitly sums or union-bounds over the m fault classes (the $\log(m/\delta)$ term). We sketch a resolution-limited extension below, rather than a full continuous theory; it answers "how would you adapt this" with a concrete construction, not a claim that continuous F reduces to the finite case without loss.

Construction. Let (F, ρ) be a compact metric space and $\Phi : F \to \mathcal{S}$ (signature space, metric d) be L-Lipschitz: $d(\Phi(f), \Phi(f')) \le L\,\rho(f, f')$ for all $f, f' \in F$. Fix a resolution $r > 0$ and let $F_r = \{f_1, \dots, f_{N(r)}\}$ be an r-net of F (every $f \in F$ satisfies $\rho(f, f_i) \le r$ for some $f_i \in F_r$), $N(r)$ the r-covering number of (F, ρ). Let γ_r denote the partition margin (Definition 2) of the *finite* discretized problem on F_r.

Proposition 4 (ε-Net Reduction for Continuous Fault Spaces)**.** *Suppose $\gamma_r > 4Lr$. Then:*

(a) ***Soundness.*** *For any $\varepsilon \in (2Lr,\ \gamma_r - 2Lr)$, the confusability structure of F at threshold ε agrees with the pullback of F_r's confusability structure under nearest-net assignment: if*

$f, f' \in F$ have distinct nearest net points $f_i \neq f_j$ lying in different blocks of $\Pi^{\star}_{\varepsilon_r}(F_r)$ (any $\varepsilon_r < \gamma_r$), then $d(\Phi(f), \Phi(f')) > \varepsilon$.

(b) ***Sample complexity.*** *Applying Corollary 2 to the finite problem on F_r, with effective margin $\gamma_r^{\text{eff}} := \gamma_r - 4Lr > 0$: exact recovery of the r-resolution partition with probability $\geq 1 - \delta$ occurs whenever* $\min_{f_i \in F_r} n_{f_i} = \Omega\left(\dfrac{C(H) + \log(N(r)/\delta)}{(\gamma_r^{\text{eff}})^2}\right)$.

Proof sketch. (a): let f, f' have nearest net points f_i, f_j. Lipschitz continuity gives $d(\Phi(f), \Phi(f_i)) \leq Lr$ and $d(\Phi(f'), \Phi(f_j)) \leq Lr$; by the triangle inequality, $d(\Phi(f), \Phi(f')) \geq d(\Phi(f_i), \Phi(f_j)) - 2Lr \geq \gamma_r - 2Lr > \varepsilon$ whenever $\varepsilon < \gamma_r - 2Lr$, and the stated interval is nonempty exactly when $\gamma_r > 4Lr$. (b): the same threshold-sandwich argument used for Theorem 5/Corollary 2 applies verbatim to the finite problem on F_r ($N(r)$ in place of m), with the perturbation bound from (a) absorbed into the effective margin $\gamma_r^{\text{eff}} = \gamma_r - 4Lr$ in place of $\gamma - \varepsilon$. A full formalization (matching the level of detail of Appendix G's other proofs) is left to future work; this sketch establishes the construction and the form of the bound, not a fully worked concentration argument. □

Limiting case. If F is already finite and r is taken below the minimum pairwise ρ-distance, $N(r) = m$, $Lr \to 0$, and Proposition 4 reduces exactly to Theorem 6/Corollary 2.

Explicit rate for Euclidean fault manifolds. If F is a compact subset of $\mathbb{R}^d$ (e.g. a bounded box of degradation parameters or sensor-bias values) with the standard metric, the covering number satisfies the standard volumetric bound $N(r) = O\big((\operatorname{diam}(F)/r)^d\big)$, giving

$$\min_{f_i \in F_r} n_{f_i} = \Omega\left(\frac{C(H) + d\log(\operatorname{diam}(F)/r) + \log(1/\delta)}{(\gamma_r - 4Lr)^2}\right),$$

an explicit dependence on the fault manifold's intrinsic dimension d and the chosen resolution r, rather than an abstract existence claim.

What this does and does not give. This is a *resolution-limited* guarantee: it recovers the partition at coarseness r (fault parameters within r of each other are not distinguished), not exact recovery of the continuum's true partition structure – a genuinely weaker statement than the finite-F exact- recovery theorems elsewhere in the paper, not a drop-in equivalent. Two assumptions carry the real weight and can fail exactly where continuous extensions are most wanted: *compactness* of F (a reasonable assumption for physically bounded degradation or drift parameters, but not for unbounded trajectories without an additional boundedness argument), and *Lipschitz continuity* of Φ (which can fail at bifurcations or discontinuous fault-mode transitions in nonlinear systems, where nearby parameter values produce qualitatively different behavior – precisely the regime a practitioner would most want covered). Weakening to local or Hölder continuity, or excising a lower-dimensional singular set, is the natural next step and is not attempted here.

Structurally, this construction is the same move as the equivariant orbit-space reduction discussed in relation to symmetry groups elsewhere in this line of work: both replace a large space with a canonical, coarser summary object (net representatives vs. group orbits) and recover statistical complexity in terms of that object rather than the original space. The two are not identical – metric covering is an approximate, resolution-dependent reduction, while a group orbit is exact – but the pattern (compress via an equivalence- generating map, pay for it in a controlled approximation term) is the same one.

I.5 Representation Realizability

The partition-consistency theorem assumes the existence of a realizable representation class capable of expressing the population partition.

If the hypothesis class H cannot represent the true partition, recovery becomes impossible regardless of sample size. The present theory therefore addresses estimation error but does not explicitly characterize approximation error.

Future work should incorporate representation misspecification into the analysis.

I.6 Capacity Measures

The sample-complexity analysis is stated abstractly in terms of a capacity measure $C(H)$. This formulation accommodates VC dimension, Rademacher complexity, covering numbers, and related capacity notions.

However, the theory does not derive explicit bounds for particular deep-learning architectures. This gap is not specific to the present framework: classical VC-dimension and Rademacher-complexity bounds are well known to be vacuous for the overparameterized networks used throughout modern deep learning, which routinely generalize despite having enough capacity to fit randomly-labeled data (Zhang et al., 2016). No existing line of work has fully closed this gap even for ordinary supervised classification – norm-based and spectrally-normalized margin bounds (Bartlett et al., 2017), PAC-Bayes bounds optimized directly against the trained network (Dziugaite & Roy, 2017), and compression-based bounds (Arora et al., 2018) each give non-vacuous bounds for specific architectures and settings, but none currently instantiates $C(H)$ for the deep, task-structure-sufficient representations ϕ_θ studied here.

In the absence of an instantiated $C(H)$, the Design-Choice Ablations of Section 8.3 and the Temperature-Schedule Ablation of Appendix H.12.3 serve as the practical complement: both show the qualitative predictions of the abstract theory – sample-size scaling (recovery improving with n at the rate the theory predicts up to constants), margin behavior (the $\lambda_{\text{sep}} = 0$ ablation's embedding margin $\hat{\gamma}$ tracking exactly what the theory says a properly-trained representation should do to the raw population margin γ), and bias-versus-variance signatures (the wrong-quotient ablation's plateau, Section 8.3, distinguishing an irreducible approximation-error floor from shrinking estimation error) – holding empirically on a real trained architecture, even without a closed-form $C(H)$ for that architecture.

In modern neural networks, meaningful capacity measures remain an active area of research.

Consequently, the sample-complexity results should be interpreted as structural scaling laws, empirically supported on the architectures tested here, rather than architecture-specific guarantees derived from first principles.

I.7 Task-Structure Sufficiency versus Accuracy

The framework deliberately focuses on partition preservation rather than classification accuracy. These quantities are related but not identical. A representation may achieve high fault-classification accuracy while failing to preserve the underlying fault partition.

Conversely, a representation may preserve the partition but yield suboptimal classification performance under a particular classifier. The theory therefore characterizes structural diagnosability rather than end-to-end predictive performance.

I.8 Benchmark Dependence

The experimental methodology emphasizes the multi-tank system and IEEE 7-bus feeder. These benchmarks are particularly attractive because their confusable fibre structures are either known or can be derived analytically. However, conclusions drawn from these systems may not generalize immediately to all scientific (e.g., diagnostic) domains. Additional validation on larger process systems, industrial benchmark datasets, cyber-physical systems, and real operational fleets would strengthen the empirical claims.

I.9 Choice of Signature Metric

Partition estimation depends upon the metric d used to compare latent signature distributions. Different metrics emphasize different geometric properties. For example, Wasserstein distance is sensitive to distribution geometry, whereas kernel-based metrics such as MMD capture distributional discrepancies through feature embeddings.

Although the theory is metric-agnostic, empirical results may vary with the selected metric. Future work should investigate robustness across multiple distance measures.

I.10 Strong and Weak Fault Models

The empirical programme compares strong-fault and weak-fault realizations. The strong/weak distinction provides a principled method for varying fault knowledge, but it represents only one dimension of realization-construction structure.

Other forms of knowledge variation are possible, including:

- structural abstraction,
- parameter uncertainty,
- incomplete component descriptions,
- probabilistic fault semantics.

The proposed experiments therefore test only a subset of possible realization-construction trade-offs.

I.11 Unlabeled and Weakly Supervised Data: Full Discussion

The Empirical Signature definition of Section 3 presupposes that each observation's fault identity is known, so that the samples contributing to $\widehat{\Phi}_n(f)$ can be grouped by f before the signature is formed. This is a genuinely different, and stronger, requirement than uniform consistency of the estimator (A3): it is a labeling requirement, not a sample-size requirement, and it is not satisfied when fault identity is unobserved.

The fully unlabeled setting (Factor F at $\ell = 0\%$) therefore calls for a two-stage estimator rather than a direct application of Section 3's machinery: (i) partition the unlabeled latent samples into clusters using a distribution-clustering procedure operating on Z directly, without reference to F; (ii) assign each resulting cluster to a fault identity by matching against a small labeled reference set or the reference partition $P^\star$. Stage (i) raises an identifiability question that has no counterpart in the labeled theory of this paper: whether the population partition $\Pi^\star_\varepsilon$ is even recoverable as the limit of unsupervised clusters, prior to and independently of whether those clusters can subsequently be labeled correctly. This is a distinct estimation problem, with its own consistency conditions on the clustering procedure (e.g. separation and density conditions considerably stronger than the pairwise margin γ of Section 2), and a supervised-vs-unsupervised gap can open even when γ is large, since clustering consistency depends on properties of Φ beyond pairwise separation (e.g. the shape and density of each $\Phi(f)$, not only the distances between them).

A full identifiability and consistency theory for the general unlabeled setting is left to future work: nothing here addresses how $k = |\Pi^\star_\varepsilon|$ would be estimated from data, nor clustering procedures beyond k-means, nor non-spherical or heavy-tailed $\Phi(f)$. What follows is a narrower sufficient condition under which stage (i) is already solved by a classical result.

Assumption 4 (Cluster Identifiability)**.** *$k = |\Pi^\star_\varepsilon|$ is known, and the population signatures $\{\Phi(f)\}$, at the fibre level, form k well-separated, finite-variance clusters for which the population k-means objective's global minimizer induces exactly $\Pi^\star_\varepsilon$.*

Proposition 5 (Two-Stage Estimator Consistency)**.** *Under Assumption 4, if stage (i) is solved by k-means fit to a global optimum, the empirical clustering converges almost surely to $\Pi^\star_\varepsilon$ as $n \to \infty$ (Pollard, 1981); given one labeled reference point per fibre, stage (ii) is eventually correct almost surely.*

The qualifier *global, not merely local* optimum in Proposition 5 is load-bearing, not a formality, and is the source of a mistake worth recording. Pollard's theorem is a statement about the population k-means objective's global minimizer; it says nothing about whether Lloyd's algorithm, run once from a single random initialization, finds that minimizer. An earlier version of the oracle baseline used in Section 8 did exactly this – a single initialization, no restarts – and its recovery rate empirically *plateaued* short of exact recovery even as n grew into the hundreds, which is the signature of a bad local optimum, not of a violated identifiability assumption, and would have been easy to mistake for the latter. Re-running the same estimator with ten random restarts, keeping the lowest-cost result, changed the outcome from a persistent plateau to exact recovery from $n = 20$ onward (Table 4), matching Proposition 5 exactly. We flag this explicitly because it is the kind of failure that

Table 4: Oracle ($\ell = 0\%$) two-stage estimator recovery vs. n, field `M_obs`, k-means with 10 restarts, 30 seeds per n. Consistent with Proposition 5 under Assumption 4.

n per fault	10	20	40	80	160	320	640
ARI	0.947	1.000	0.989	1.000	1.000	0.989	1.000
exact match	0.90	1.00	0.97	1.00	1.00	0.97	1.00

looks identical, from the outside, to a genuine statistical inconsistency, and the two call for entirely different fixes.

Assumption 4 is not free, and does not always hold: on the IEEE 7-bus feeder's `trunk{1,2,4}` placement (Section 8), whose true fibres vary considerably in size and are not spherical in the same sense, the properly-optimized oracle recovers the true partition only partially ($\mathrm{ARI} \approx 0.79$) even at $n = 40$ – illustrating concretely, on a benchmark already in this paper rather than a constructed counterexample, what happens when the assumption's separation-and-shape condition is not met. Section H.4's $\ell = 0\%$ results throughout should be read against this: they hold under Assumption 4, are empirically consistent with it on the two-tank benchmark, and are only partially borne out on the feeder benchmark where it is more strongly violated.

I.12 Causality and Interpretation

The framework characterizes relationships between realization-construction resources and partition recovery. It does not establish a causal theory of fault generation, physical mechanisms, or representation formation. The induced partition should therefore be interpreted as a task-structure object rather than a causal model of the underlying system.

I.13 Scope of the Present Contribution

The principal contribution of this paper is not a new fault classifier, representation-learning algorithm, or neural architecture. Instead, the paper provides:

1. a statistical theory of partitions;
2. consistency guarantees for partition recovery;
3. a notion of task-structure sufficiency;
4. an information-theoretic characterization of task-structure capability;
5. experimentally testable predictions relating sensors, fault knowledge, data availability, and latent representations.

The framework should therefore be viewed as a foundation for task-structure representation learning.

J Full Related Work Discussion

We now extend the condensed related-work summary of Section 2. This work lies at the intersection of realization theory, statistical learning theory, representation learning, information bottlenecks, and geometric deep learning, with application to scientific domains like fault diagnosis. This section positions the proposed framework relative to these literatures.

J.1 Partition Learning and Structured Outputs

The perspective of framing learning as partition recovery connects with broader machine learning literature on learning with structured outputs, metric learning, and clustering-based representation learning (Caron et al., 2020; Taskar et al., 2005). Rather than predicting flat categorical labels, structured prediction and metric learning aim to respect underlying topological or combinatorial constraints. In the context of deep representation learning, clustering and partition-level objectives have gained traction for discovering natural groupings without rigid supervision (Caron et al., 2020).

However, existing partition-learning approaches are typically heuristic clustering objectives (such as k-means variants or modularity maximization). The framework developed here provides a rigorous statistical foundation for partition recovery, establishing consistency and sample-complexity guarantees grounded in partition margins and hypothesis-space capacity.

With regard to *OOD / open-set recognition*, knowing the true partition structure is essential for open-set recognition: merged fibres correspond to classes that a model will systematically confuse under distribution shift.

J.2 Statistical Learning Theory

The statistical component of the present framework is inspired by classical learning theory. Uniform convergence, consistency, sample-complexity analysis, and capacity control form the foundation of modern statistical learning (Vapnik, 1998; Shalev-Shwartz & Ben-David, 2014).

Traditional learning theory studies convergence of prediction error, classification risk, or regression risk. The present work instead studies convergence of a combinatorial object: the partition.

The partition consistency theorem established in this paper may be viewed as a partition-recovery analogue of statistical consistency. Similarly, the partition recovery bounds developed here play a role analogous to classical sample-complexity results, with partition margin replacing the role of a classification margin.

J.3 Representation Learning and Sufficient Representations

Representation learning seeks latent variables that capture task-relevant structure while removing irrelevant variability. Modern approaches include autoencoders, variational autoencoders, contrastive learning methods, and deep representation-learning frameworks.

A central question in this literature concerns what information must be preserved by a learned representation. Classical statistical sufficiency characterizes representations that preserve complete information about a parameter of interest.

The notion of task-structure sufficiency introduced here is related but distinct. Instead of preserving parameter information, a task sufficient representation preserves the partition of the fault space. This yields a realization-theoretic notion of sufficiency defined directly through fault distinguishability.

J.4 Information Bottlenecks

The information bottleneck framework introduced by Tishby, Pereira, and Bialek formulates representation learning as a tradeoff between compression and retention of relevant information (Tishby et al., 1999; 2000). Given input X, target Y, and representation Z, the objective seeks representations preserving information about Y while compressing X.

The partition information bottleneck introduced in this paper differs in two important respects.

First, the relevant object is not a label variable but a partition.

Second, preservation is measured through fault distinguishability rather than predictive information.

Consequently, task-structure sufficiency is generally stronger than standard predictive sufficiency. A representation may preserve classification performance while failing to preserve the underlying confusable-fibre structure of the fault space.

J.5 Geometric Deep Learning and Equivariance

Recent work in geometric deep learning has emphasized the role of symmetry, invariance, and equivariance in representation learning (Bronstein et al., 2021; 2017). The central premise is that useful representations should respect the geometric structure and symmetries present in the data.

This perspective aligns naturally with realization theory. Confusable fibres may be interpreted as orbit-like objects generated by task-relevant invariances. The latent signature distributions intro-

duced in this paper may therefore be viewed as geometric objects whose separation determines diagnosability.

The partition margin introduced in the present framework plays a role analogous to geometric separation measures used in representation learning and manifold learning.

J.6 Information-Theoretic Views of Deep Representations

Information-theoretic analyses of deep learning have attempted to explain representation formation through compression, mutual information, and bottleneck principles (Tishby et al., 1999; 2000).

The present work contributes an alternative interpretation. Rather than regarding deep representations primarily as compressed predictors, we regard them as estimators of partitions.

Consequently, partition information is defined at the level of fault partitions rather than latent variables. This shift permits direct comparison between symbolic and learned realizations, something that standard representation-learning theory does not naturally provide.

J.7 Behavioral Systems and Realizations

The realization-theoretic perspective adopted in this paper is closely related to Willems' behavioral approach to dynamical systems (Willems, 2007). In the behavioral framework, the primary object is the behavior itself rather than a particular state-space representation. Systems are characterized through the trajectories they permit, and multiple realizations may correspond to the same behavior.

Our realization-theoretic perspective follows Willems' behavioral approach to dynamical systems (Willems, 2007), but where Willems studies behavioral equivalence, we study *task-structure* equivalence: two realizations differing internally are task-structure equivalent iff they induce the same confusable-fibre partition. This paper shares vocabulary with a companion Diagnostic System Specification (DSS) framework (a category-theoretic account with a Representation Theorem establishing a canonical minimal sufficient realization S_M), but the two are complementary, not overlapping: DSS is algebraic and exact, characterizing S_M from a behavioural specification with no sampling or estimation; this paper is statistical, asking when a partition is recoverable from finite samples of a *learned* representation with no symbolic specification assumed. Section 7.1 uses DSS's S_M, where available, as one source of the reference partition $P^\star$ – the $n \to \infty$, symbolic-model limit this paper's finite-sample theory estimates toward – but the statistical machinery applies without it (via the bootstrap of Section 7.1).

Our work shares this emphasis on representations of behavior but differs in objective. Willems studies behavioral equivalence and realization of dynamical systems. By contrast, we study task-structure equivalence and partition recovery. Two realizations may differ significantly in their internal structure while remaining task-structure equivalent if they induce the same confusable-fibre partition.

From this perspective, task-structure capability becomes a property of an induced partition rather than of a specific representation formalism.

J.8 Foundation models for Science

As foundation models are increasingly applied to physical and biological systems, the mismatch between predictive pre-training objectives and downstream structural tasks becomes critical. Transformer and self-supervised architectures now are prevalent in physics (Raissi, 2018), fault diagnosis and time-series representation learning (Wen et al., 2023; Nie et al., 2023; Franceschi et al., 2019; Yue et al., 2022), but are evaluated through classification accuracy or reconstruction/contrastive loss, treating representations as continuous embeddings rather than discrete partition structures; we provide a partition-theoretic lens for what such representations preserve about fault distinguishability. Partition and clustering-based learning (Caron et al., 2020; Taskar et al., 2005) typically uses heuristic objectives (k-means variants, modularity); we give a statistical foundation with consistency and sample-complexity guarantees grounded in a partition margin.

J.9 Deep Learning for Time-Series Representations

Recent years have witnessed a rapid shift toward deep learning paradigms in time-series modeling, e.g., financial forecasting, fault diagnosis and industrial anomaly detection (Lim & Zohren, 2021). Transformer-based architectures have become prominent for capturing long-range temporal dependencies and complex multi-variable interactions in non-stationary sensor streams (Wen et al., 2023; Nie et al., 2023). Concurrently, self-supervised pre-training frameworks designed specifically for time-series and multi-sensor data have addressed the critical bottleneck of scarce labeled failure modes in industrial deployments (Franceschi et al., 2019; Yue et al., 2022). While these models achieve high predictive performance, they are typically evaluated through classification accuracy, reconstruction error, or contrastive losses, treating representations as continuous embeddings rather than discrete partition structures. Our work complements this literature by providing a formal partition-theoretic lens to understand what these deep representations actually preserve regarding fault distinguishability.

J.10 Diagnosis

Relationship to Existing Diagnosis Frameworks Existing diagnosis frameworks typically compare methods through prediction accuracy, fault-isolation rate, computational cost, or model assumptions. The present work proposes a different comparison criterion. Two diagnostic systems are considered equivalent when they recover the same confusable-fibre partition.

Consistency-based diagnosis. This approach originates in Reiter's classical formulation of diagnosis from first principles (Reiter, 1987). A system description together with observations induces a set of conflicts from which diagnoses are derived.

Strong-fault and weak-fault models subsequently became an important distinction in model-based diagnosis (Feldman et al., 2010). Strong-fault models explicitly characterize fault-mode behavior, while weak-fault models assume only nominal behavior and represent faults through abnormality predicates.

Consistency-based diagnosis (Reiter, 1987) and the strong-/weak-fault model distinction are reinterpreted here as different construction structures Λ inducing different partitions, permitting direct comparison with learned realizations through induced fibres rather than implementation mechanism. Observer-based FDI (Gertler, 1998; Ding, 2014) residual spaces are one realization class among many in this framework; we extend the idea to learned latent representations.

The present work does not replace model-based diagnosis. Instead, it is a data-driven framework that reinterprets strong- and weak-fault models as different construction structures giving rise to different partitions. This permits direct comparison of symbolic and learned realizations through their induced confusable fibres, rather than through their implementation mechanisms.

DSS Realization-Theoretic Framework. This paper has an overlap with (Provan, 2026), which develops a model-based diagnosis Representation Theorem establishing a minimal sufficient realization S_M, canonical and terminal among realizations supporting a given behavioural model M, together with associated diagnosability recovery theorems. It is worth being precise about how the two frameworks differ, since the overlap in terminology could otherwise suggest overlap in content.

This framework (termed DSS) is algebraic and exact: given a behavioural specification, it characterizes the canonical realization and the partition it induces without reference to sampling or estimation. Our paper is statistical and estimation-theoretic: it does not assume a behavioural specification is available, and instead asks when a partition can be recovered from finitely many samples of a *learned* representation, with no guarantee that the learned representation coincides with, or even approximates, any symbolic realization. None of the concentration-based consistency, sample-complexity, or differentiable-training machinery developed here has a counterpart in the DSS framework, and conversely none of DSS's universal-property/canonicity apparatus is used here.

The two frameworks are better understood as complementary than as overlapping. Section 7.1 uses the DSS minimal sufficient realization S_M, where available for a given benchmark and sensor configuration, as one source of the exact reference partition $P^\star$ against which a learned representation's task-structure sufficiency is evaluated in Section 8; in this sense the DSS partition can be read as the $n \to \infty$, $\varepsilon \to 0$, symbolic-model limit that the present paper's finite-sample theory is estimat-

ing toward, when such a limit is available and known. Where no such symbolic characterization exists — for instance because no tractable behavioural model of the system has been constructed, or because diagnosability must be assessed directly from a black-box learned representation — the present paper's statistical machinery applies without requiring one.

Fault Detection and Isolation Observer-based fault detection and isolation methods construct residual generators from dynamical models and use residual signatures for detection and isolation. Classical treatments include the work of Gertler and Ding (Gertler, 1998; Ding, 2014).

Within the realization-theoretic framework, residual spaces constitute one realization class among many. Residual signatures induce partitions of fault space through detectability and isolability relations. The current paper extends this idea by treating learned latent representations as alternative realizations whose task-structure value is determined by the fault partitions they induce.