Representation Preconditions Sample Efficiency at Constant Information

A Model-Free, Causally-Probed Structural Predictor of the Learning Gap Between Informationally Equivalent Serializations

Jérôme Vétillard
VP R&D + Engineering / Chief Product Officer, TweenMe by Qualees

---

Abstract

Two serializations of the same longitudinal cohort can carry provably identical information yet be learned with different ease. We study this under a hard constraint: the two representations are mutually reconstructible by a deterministic bijection, so neither contains more information than the other. Using ex nihilo synthetic generators with controlled ground truth, we define an *operational ontology* as the choice of head axis in the data tensor, that is, which axis is treated as the exchangeable unit one batches over. We report a bundle of mutually corroborating findings. (i) An asymmetric crossover: at equal information and equal budget, the representation whose head axis matches the prediction unit learns better, and this advantage survives holding the encoder fixed, which isolates the representation from the encoder and decomposes the raw crossover into a representation share and an encoder share. (ii) A null control confirms the pipeline fabricates no gap absent an asymmetric mechanism. (iii) The crossover is reproduced without training, by a k-nearest-neighbor probe, so it is a property of the induced metric. (iv) A model-free structural quantity, neighborhood purity, predicts the sign and magnitude of the performance gap across a family of tasks (Pearson $r = 0.996$; cluster bootstrap and permutation reported). (v) An information-preserving intervention on the metric moves performance together with purity at fixed label and fixed information ($r = 0.942$), establishing causality *at the level of the metric*. The bundle reproduces across four synthetic worlds differing in observation modality, including a graph domain learned by message passing; we are explicit that these worlds share a two-scale latent structure, so their independence is of modality, not of the generative skeleton. We add a direct data-to-performance measurement and a validation on two real longitudinal panels, where the predictor holds ($r = 0.98$ and $r = 0.996$). We show honestly that neighborhood purity is a sufficient but not a unique mediator: a distance-based separability predicts equally well. We also chart the theory's own boundary: purity predicts learners dominated by the induced neighborhood (a kernel machine, a tree ensemble) and, as pre-registered to fail, does not predict a global linear learner on an XOR-structured label; and the predictor degrades gracefully, not catastrophically, when the bijection is only approximate. We frame the result as an interaction between representation, learnability, and sample efficiency, and declare its boundaries.

1 Introduction

1.1 Representation as inductive bias

Learning is generalization from resemblances. A model that has never seen an exact case must decide which known case to treat it as similar to; every generalization presupposes a notion of proximity. Without a bias, all hypotheses consistent with the data remain tied (Mitchell, 1980), and no learner is better than another averaged over all problems (Wolpert, 1996; Wolpert and Macready, 1997). This tradition speaks of the bias of the *model*: the locality of a convolution, the sequentiality of a recurrence, the entities and relations of a graph network (Battaglia et al., 2018). A second bias, distinct and less often named, is induced by the

representation. De Finetti formalized it (de Finetti, 1937): to declare observations exchangeable is to posit a latent model, and choosing the exchangeable unit fixes the object whose repetition one assumes.

The two biases compose rather than substitute. A representation preconditions the architectural bias, in the sense of numerical optimization where a preconditioner does not change the solution of a system but the ease of reaching it. The chain is not Architecture $\rightarrow$ Function but Architecture $\rightarrow$ Representation $\rightarrow$ Function: the same encoder, trained on two serializations of one dataset, can learn two different functions, because the representation has already fixed the proximities within which the architectural bias operates.

1.2 The tensor representation

Concretely, an operational ontology is a choice of tensor layout: which axis is the exchangeable axis, traversed in i.i.d. batches, and how the remaining axes nest. A cohort can be arranged as $[\text{patient}, \text{time}, \text{variables}]$ or $[\text{time}, \text{patient}, \text{variables}]$: two tensors of the same bytes, reversible into one another, but whose head axis differs, and with it what the machine holds as near. A reader from knowledge representation will object that this is a tensor layout, not an ontology in the sense of entities, relations, and types. The objection is correct as to vocabulary and beside the point as to mechanism: the decision that data engineering calls the ontology, the choice of exchangeable unit, is computationally a choice of layout, and it is that choice, not a vocabulary, that we show acts on learning. Throughout, "representation" is operationalized as the choice of grouping axis, and no claim is made beyond it.

1.3 What the literature does not test

Representation learning studies how to *learn* features (Bengio et al., 2013), not how a fixed layout at constant information changes learnability. The literature on aggregation and ecological inference, from Simpson (1951) to Robinson (1950), warns that the level of aggregation changes what one reads, but for inference, not for sample efficiency. On the model side, the neural tangent kernel (Jacot et al., 2018), spectral bias (Rahaman et al., 2019), and double descent (Belkin et al., 2019; Nakkiran et al., 2019) characterize what a given architecture finds easy, holding the data representation fixed. Within the scope of our review of these literatures, we did not find prior work that isolates the representation while holding information provably constant by an explicit bijection; we state this as a bounded search, not an exhaustive novelty claim.

1.4 Contribution and scope

We contribute a model-free structural predictor of the learning gap between informationally equivalent serializations, a causal ablation that upgrades it from correlational to causal at the level of the metric, an independence demonstration across four synthetic worlds and two real panels, and an honest bounding of the mediator. The demonstration is synthetic and real-panel, with controlled or observed ground truth; we hold the boundary throughout.

2 Related work

Representation and metric learning. Representation learning seeks features that ease downstream tasks (Bengio et al., 2013); metric learning optimizes a distance for a task. Both *learn* the representation or metric from data. We instead fix two representations a priori, constrain them to identical information, and study the resulting learnability gap, which isolates the effect of the layout choice rather than of a learned transform.

Inductive bias, invariance, and geometry. Relational inductive biases (Battaglia et al., 2018) and geometric deep learning (Bronstein et al., 2021) formalize how architectural structure encodes assumptions. This is the bias of the model. Our object is the complementary and less studied bias of the data representation, and our fixed-encoder control (Section 3.4) separates the two.

The geometry of learning. The neural tangent kernel (Jacot et al., 2018), spectral bias (Rahaman et al., 2019), the information bottleneck (Tishby et al., 1999), and double descent (Belkin et al., 2019; Nakkiran et al., 2019) explain what a fixed architecture finds easy. They act on the model side; we show that the data representation is a lever on the same effective metric.

Exchangeability and aggregation. De Finetti's exchangeability (de Finetti, 1937) ties the choice of unit to a latent model. The Simpson and ecological-inference literature (Simpson, 1951; Robinson, 1950) shows that aggregation level changes inferential conclusions; we transport the warning from inference to sample efficiency.

Similarity and projectibility. The reading that a representation is a hypothesis about relevant similarities has antecedents in Edelman (1998), and further upstream in the projectibility of Goodman (1955) and the natural kinds of Quine (1969). Our addition is that this similarity, once measured, predicts and (at the metric level) causes sample efficiency.

3 Methods

3.1 Ex nihilo generators with controlled ground truth

The primary generator (world A) produces a synthetic longitudinal cohort with a discrete-time proportional-hazards survival mechanism. The hazard of patient i depends separately on two components of a latent driver,

$$h_i(t) = h_0(t)\, \exp\big(\beta_{\text{bet}}\, L_i + \beta_{\text{wit}}\, w_i(t) + \gamma^\top Z_i\big), \tag{1}$$

where L_i is the patient's structural level (between component), $w_i(t)$ the within deviation, Z_i static covariates, and h_0 a Weibull baseline. Opposite signs of $\beta_{\text{bet}}, \beta_{\text{wit}}$ give a Simpson-type regime. The learners see only noisy measured biomarkers at visits; ground truth is retained for validation only. Three further worlds share the test skeleton but differ in the generative process: world B, a continuous panel with Student-t innovations, bimodal levels, a mean-reverting oscillatory within process, and dropout; world C, a bipartite latent-factor matrix (users by items) with a per-item nuisance and native sparsity, no time; world D, a stochastic block model graph whose node signal splits into a graph-smooth (population) and an idiosyncratic (local) component. Full equations are in Appendix A. The four worlds differ on at least two axes each (mechanism, noise, dependency, level distribution, observation), but they share a two-scale latent decomposition, which we discuss as a limit.

3.2 Informational equivalence and three senses of information

The constraint is hard: the two serializations of a dataset are mutually reconstructible by a deterministic bijection, verified by a round-trip exact-equality test across configurations. This requires distinguishing three senses of the word *information*. The bijection guarantees identical *Shannon* information: the same bits, each a lossless function of the other. It does not guarantee identical *algorithmic accessibility*: the same bits can be arbitrarily harder to exploit under one layout. And it does not guarantee identical *statistical efficiency*: the target may be a minimal sufficient statistic in one representation and a deeply entangled function in the other.

Our result lives in the gap between the first sense and the other two. For world D the equivalence is looser than a strict transpose (two functions of a shared substrate), an honest cost of the geometric regime.

3.3 Tasks and a parameterized target

An *individual* task predicts an individual outcome from the individual's own history (world A: landmark survival concordance). A *transversal* task predicts a population-relative quantity, made cross-sectional by a wave effect, a random per-slice shift playing the role of a batch effect that only a model seeing the slice can remove. A knob w interpolates the label unit from individual ($w = 0$) to transversal ($w = 1$) by mixing standardized patient-context and slice-context components. Worlds C and D transpose the duality to item-vs-user and graph-vs-feature.

3.4 Encoders at equalized budget and the fixed-encoder control

Each ontology is given the encoder adapted to its structure at equalized parameter budget (causal GRU and Transformer, person-period pooled MLP, set encoder). Because pairing each ontology with a different encoder confounds representation and encoder, we introduce a fixed-encoder control: one set encoder applied two ways differing only in the grouping axis over which it pools (by slice vs by patient). In world D the fixed encoder is a one-hop message-passing layer whose pooled context is the graph neighborhood or a geometry-blind global set. Hyperparameters are in Appendix B.

3.5 Structural alignment, model-free

A representation induces a metric, hence a neighborhood graph. *Neighborhood purity* is, for a representation and label, the average fraction of the k nearest neighbors sharing the point's label; it is a property of (metric, label), computed with no trained model. As a competitor we compute a distance-based Fisher-type *separability*.

3.6 Robustness and validity controls

We report multi-seed estimates with bootstrap intervals and cross indicators against three learners (logistic regression, linear SVM, random forest); the random forest, lacking any built-in neighborhood, guards against a self-validating theory. All sequence models are strictly causal (anti-leakage); a landmark fixes the risk horizon; a null generator must predict a zero gap. We pre-register the predicted sign, an amplitude band, and a refutation threshold before each independence and depth experiment (Appendix C).

3.7 Causal ablation

To move from prediction to cause we intervene on the metric at fixed label and fixed information, by an invertible diagonal rescaling of the discriminative block, $\text{rep}_s = [\, s\, z_{\text{aligned}},\ (1/s)\, z_{\text{misaligned}},\ \text{covariates}\,]$. An invertible map preserves information but changes the Euclidean metric, hence purity and the learnability of a metric-sensitive learner. We measure, without standardization, purity and the AUC of a radial-basis-function kernel machine. Two scope conditions hold: the rescalings are a subfamily of all metric deformations, and the intervention acts on constructed coordinates, so it establishes causality along metric $\rightarrow$ neighborhood $\rightarrow$ performance, not end-to-end from the representation.

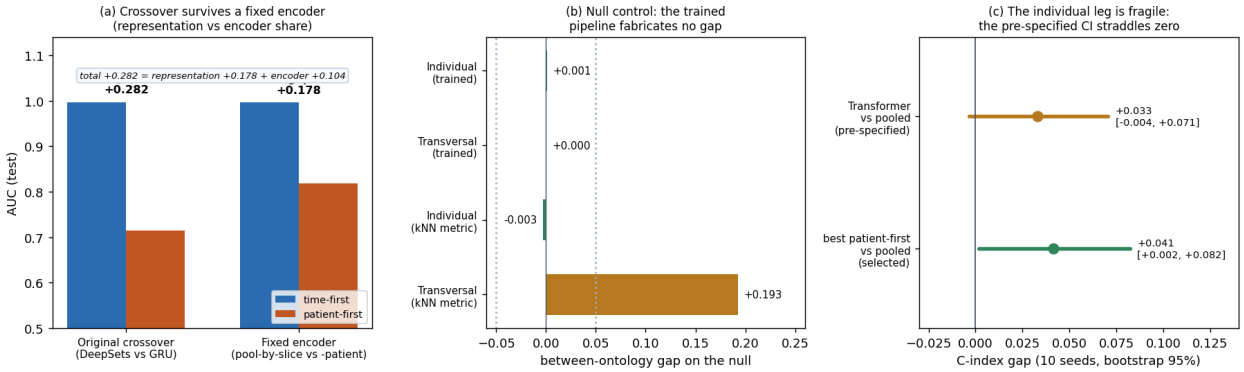

Figure 1: (a) The transversal crossover survives a fixed encoder, decomposing the raw 0.282 into representation and encoder shares. (b) On a null generator the trained pipeline produces no gap, though the model-free metric still favors time-first, a metric-versus-learnability dissociation. (c) The individual leg is fragile: the pre-specified bootstrap CI straddles zero.

4 Results

4.1 The asymmetric crossover and its honest decomposition

On two tasks on the same data, at equalized budget, the ranking reverses with the task: patient-first leads the individual task (concordance 0.785 vs 0.764), time-first leads the transversal task (AUC 0.997 vs 0.715). Two qualifications keep this honest (Figure 1). The crossover is *asymmetric*: the transversal advantage is franc (≈ 0.28), the individual advantage modest (≈ 0.02), and a ten-seed bootstrap of the pre-specified individual comparison gives $+0.033$ with 95% CI $[-0.004, 0.071]$, which straddles zero. Second, the fixed-encoder control decomposes the transversal gap: $0.282 = \underbrace{0.178}_{\text{representation}} + \underbrace{0.104}_{\text{encoder}}$. Just over a third of the headline crossover was an encoder artifact; the larger part is representational.

4.2 Null control and a metric-versus-learnability dissociation

On the representation-neutral generator the trained gap is $+0.001$ (individual) and $+0.000$ (transversal), far below the pre-registered threshold 0.05; removing the wave collapses the transversal advantage from $+0.28$ to zero. The model-free metric, however, still favors time-first by $+0.19$ on the wave-free transversal label while the trained models are at parity: a case where a structural alignment gap does not yield a performance gap because the disadvantaged model reconstructs the missing information. This is not a leak but the metric-versus-learnability distinction caught in the act.

4.3 Geometric reproduction without training

A k-nearest-neighbor probe, training no model, reproduces the crossover: on the transversal task the time-first metric reaches AUC 0.998 vs 0.740, while on the individual task the two metrics are at parity, consistent with the fragility of that leg. The crossover is thus a property of the metric, prior to any optimizer.

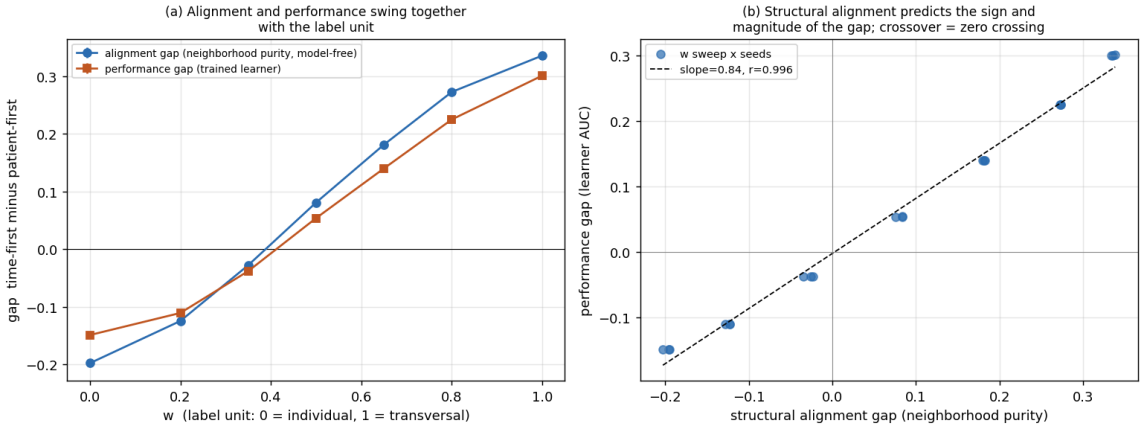

Figure 2: (a) Alignment and performance swing together with the label unit and cross zero at the same point. (b) The structural alignment gap predicts the sign and magnitude of the performance gap ($r = 0.996$, slope 0.84).

4.4 A model-free structural predictor

Parameterizing tasks by w, we measure the model-free alignment gap and the performance gap of a genuine learner (Figure 2). They swing together from negative to positive and the crossover becomes a zero crossing, with $r = 0.996$, slope 0.84. Because a correlation this high on a smooth sweep invites suspicion, we report its statistics: 21 points at 7 task levels; naive bootstrap $[0.995, 0.998]$; a cluster bootstrap resampling whole task levels $[0.994, 0.999]$; permutation $p < 0.001$. The honest caveat is not significance but interpretation, since alignment and performance are both driven by the task family, which is why the causal ablation is necessary.

4.5 Robustness

Across ten seeds, four structural indicators, and three learners, the correlation stays in $[0.987, 0.999]$; the random forest, with no notion of neighborhood, is itself predicted at 0.999 by a neighborhood measure, ruling out an optimization artifact and a self-validating theory (Appendix D).

4.6 Experimental independence across four worlds

We replicated the reduced bundle (fixed-encoder crossover, kNN alignment, alignment-to-gap sign) on three further generators, pre-registered before each run (Figure 3, Table 1). The fixed-encoder crossover gap is $0.178, 0.207, 0.108, 0.104$ across A–D; the alignment gap $0.258, 0.411, 0.141, 0.097$; the alignment-to-gap correlation between 0.996 and 0.999, with cluster-bootstrap intervals above 0.98 and permutation $p < 0.001$ in every world. The passage from world C to world D leaves the exchangeable-tensor modality for a graph with internal geometry learned by message passing. We claim generality of the sign, not the magnitude. Two limits must be stated with the result: the four worlds share a two-scale latent skeleton, so their independence is of *modality*, not statistical, and a generator breaking that structure remains to be tested; and they run through one implementation, so a common pipeline artifact is not excluded.

4.7 Direct data-to-performance measurement

We sweep the training cohort size under the fixed encoder on the transversal task (Figure 4). The time-first representation reaches AUC > 0.90 by $N \leq 200$ and stays near ceiling; the patient-first representation

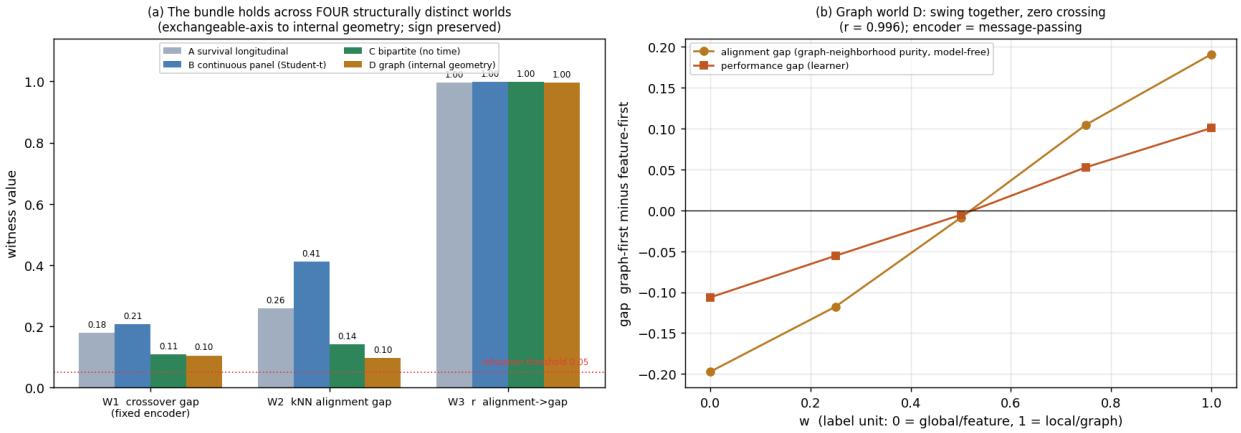

Figure 3: (a) The three witnesses across four worlds; the sign is preserved everywhere, the magnitude is world-specific. (b) The alignment-to-gap sweep in the graph world D, learned by a message-passing encoder ($r = 0.996$).

Table 1: Three witnesses and correlation statistics across four worlds. W1: fixed-encoder crossover gap. W2: kNN alignment gap. W3: alignment-to-gap Pearson r with cluster-bootstrap 95% interval and permutation p.

World	W1	W2	W3 (r)	cluster 95%	perm p
A survival longitudinal	0.178	0.258	0.996	[0.994, 0.999]	< 0.001
B continuous panel (Student-t)	0.207	0.411	0.999	[0.997, 1.000]	< 0.001
C bipartite (no time)	0.108	0.141	0.998	[0.993, 0.999]	< 0.001
D graph (internal geometry)	0.104	0.097	0.996	[0.980, 0.998]	< 0.001

plateaus below 0.85 and does not close the gap over two orders of magnitude, never reaching 0.85. This is a precise and, in one respect, stronger statement than a sample-efficiency difference. Where both serializations can converge, the effect appears as a rate difference; here, because the misaligned grouping cannot debias the wave, it appears as a representational ceiling. We keep the vocabulary exact, and we treat these as two distinct claims rather than one. The sample-efficiency claim, that the better-aligned representation reaches a common threshold with fewer samples, holds where both serializations converge. The stronger representational-ceiling claim, that the misaligned representation does not reach the aligned one's performance at all, holds where debiasing is impossible. They share a cause, the representation fixing what is easy to learn, but they are not the same statement, and a task exhibits one or the other depending on whether the misaligned grouping can in principle recover the label. We verified this dependence directly: on a bipartite task whose label is a genuine group aggregate but carries no inaccessible nuisance, the misaligned grouping does not plateau and both representations converge to within 0.001; the ceiling is therefore specific to tasks where the misaligned grouping cannot recover the label, and elsewhere the effect is the milder rate difference.

4.8 From prediction to cause, and the mechanism

Intervening on the metric by invertible rescaling at fixed label and fixed information (Figure 6a), purity and kernel-machine performance rise together, $r = 0.942$, above the pre-registered threshold 0.7. The two extremes are the two ontologies, so the choice of ontology is a point on a continuum of metric. The predictor is thereby causal at the level of the metric; we restate that this establishes metric $\rightarrow$ neighborhood $\rightarrow$ performance, not the full chain from the representation, whose first link rests on construction and on the model-free reproduction above. The link is also grounded: on the ablation data the learner error is affine

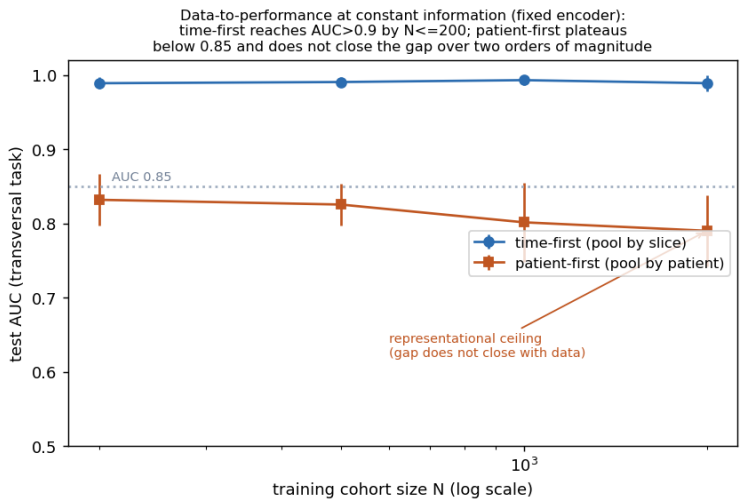

Figure 4: Data-to-performance at constant information, fixed encoder, transversal task. Time-first reaches AUC > 0.9 by $N \leq 200$; patient-first plateaus below 0.85 and does not close the gap: a representational ceiling.

in impurity (Figure 6b), $R^2 = 0.887$, positive slope, as predicted by a smoothness bias, though the slope (≈ 0.21) reflects our use of $1 -$ AUC rather than misclassification and should not be over-interpreted.

4.9 Minimality, honestly refuted

Neighborhood purity alone explains $R^2 = 0.992$ of the performance gap, and adding separability improves it by only 0.006: purity is a sufficient statistic (Figure 6c). But it is not unique. Separability alone reaches 0.997, marginally better, and the partial correlations lean against purity (perf$\sim$sep given purity $= 0.86$; perf$\sim$purity given sep $= 0.48$). Purity and separability are collinear reflections of the structure of the classes in the induced metric; the data do not crown purity. Broadening the competition (Figure 5) to four label-aware structural measures (purity, separability, mutual information, a geometric margin) and two label-agnostic geometric descriptors (effective rank, local intrinsic dimension) sharpens the diagnosis: the four label-aware measures all predict the gap (correlation between 0.99 and 1.00), while the two label-agnostic descriptors predict nothing ($r = 0.00$), because they do not distinguish the two serializations, which are re-centerings of the same points. The mediator is therefore label-aware class structure in the metric; purely geometric descriptors, blind to the label, do not suffice. We claim a causal, structural, metric-level mediator, of which purity is a sufficient and convenient estimator, not the unique fundamental variable.

4.10 Validation on real longitudinal panels

Finally we test the model-free predictor on two real measured panels whose two serializations are a strict transpose (hence informationally equivalent): *dietox* (72 pigs, 12 visits, weights) and *BtheB* (100 patients, depression BDI over 5 visits, clinical). Sweeping the label unit and measuring the model-free alignment gap against the learner performance gap (Figure 7), the predictor holds: $r = 0.983$ on the agricultural panel and $r = 0.996$ on the clinical panel, both with a zero crossing. The effect therefore survives real measurement noise on genuinely observed trajectories. As a surface-realism check, we calibrate world A against the real PBC cohort (Appendix E): event rate 0.385, bilirubin median 1.4 (IQR 0.8–3.4), median follow-up 1730 days. We are careful about what this validates: the real panels confirm the downstream, correlational half of the chain, that neighborhood purity predicts the learning gap on observed trajectories, but not the causal half,

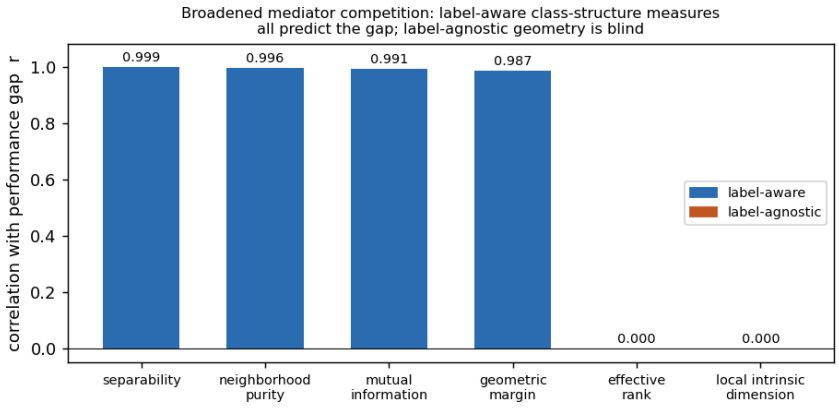

Figure 5: Broadened mediator competition across the task family. Four label-aware structural measures all predict the performance gap ($r \in [0.99, 1.00]$); two label-agnostic geometric descriptors are blind to it ($r = 0.00$). The active quantity is label-aware class structure in the metric, of which neighborhood purity is one sufficient estimator.

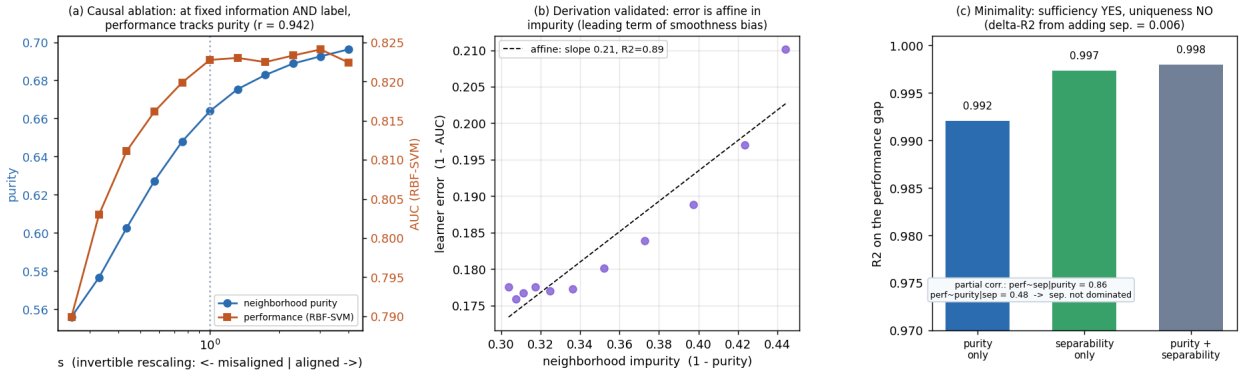

Figure 6: (a) Causal ablation: under an information-preserving, label-fixing rescaling, performance tracks purity ($r = 0.942$). (b) Error is affine in impurity ($R^2 = 0.887$). (c) Minimality: purity is sufficient but not unique; separability predicts equally well.

since the representation-to-metric intervention of Section 4.8 remains entirely synthetic. Real-data evidence for the causal link is left to future work.

4.11 The predictor's failure regime and robustness to quasi-bijection

A mature predictor should name where it fails and tolerate imperfection; we pre-registered both tests (Figure 8). **Failure regime.** On a task swept from linear-separable to XOR-structured on a fixed representation, neighborhood purity stays moderate (≈ 0.81 at the XOR pole) and a smoothness-biased RBF kernel machine tracks it ($r = 0.99$ across the sweep; AUC 0.99 at XOR), but a global linear learner collapses to chance (0.50) despite the moderate purity. The predictor therefore characterizes learning dominated by the induced neighborhood; a globally-biased learner on a locally-pure but globally-nonlinear label escapes it. Stated positively, the predictor is valid for learners whose generalization is governed by local label homogeneity in the representation's metric, that is, smoothness-biased or Lipschitz-neighborhood learners (nearest neighbors, kernel machines, and empirically tree ensembles); a formal characterization of this validity domain, in terms of a hypothesis class admitting a Lipschitz neighborhood decomposition, is left to future work. Our pre-registration guessed the failure would fall on the kernel learner; the opposite occurred, the local kernel learner

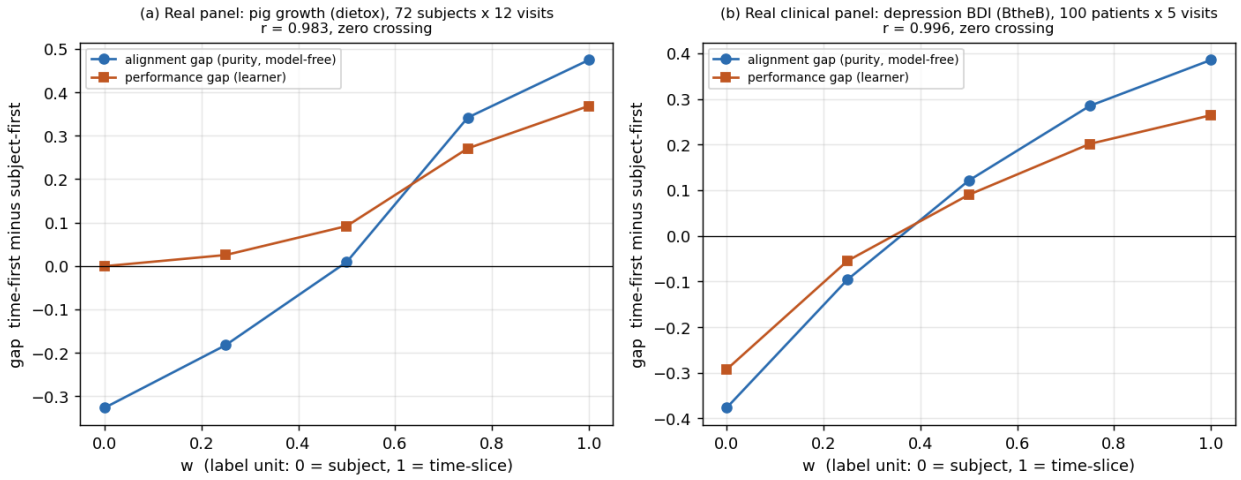

Figure 7: The model-free structural predictor on two real longitudinal panels. (a) Pig growth (dietox), $r = 0.983$. (b) Clinical depression scores (BtheB), $r = 0.996$. Both cross zero.

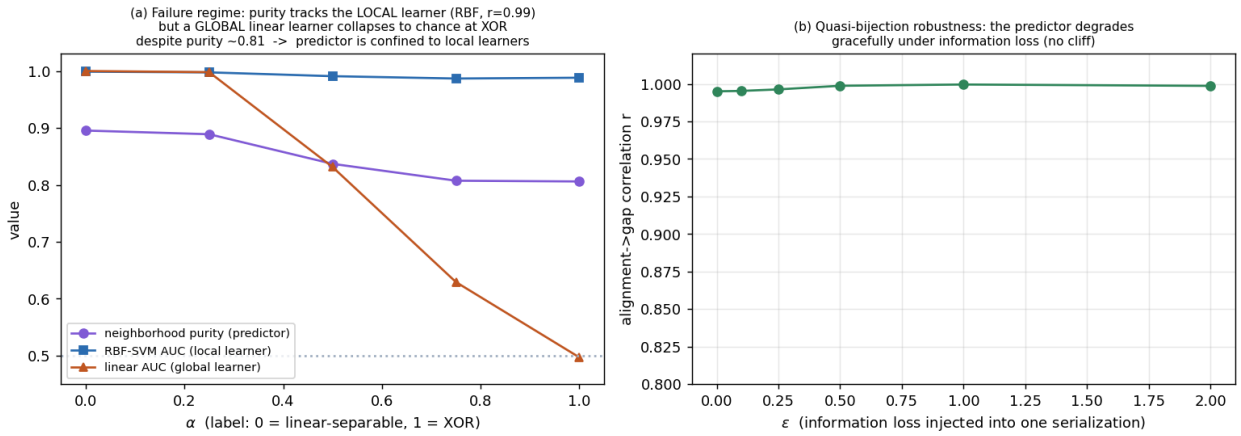

Figure 8: (a) Failure regime: purity and a local RBF learner stay high while a global linear learner collapses to chance at the XOR pole; the predictor is confined to neighborhood-dominated learners. (b) Robustness: under injected information loss (approximate bijection), the alignment-to-gap correlation degrades gracefully, no cliff.

follows purity and the global linear learner escapes it, and we report the inversion as pre-registration intends.

Robustness to quasi-bijection. Injecting information loss into one serialization, so the two representations are only approximately reconstructible, the alignment-to-gap correlation degrades gracefully rather than off a cliff: it stays above 0.99 up to substantial corruption, because the predictor is empirical and measures the purity of the representation actually supplied. Exact bijection is a clean sufficient condition, not a fragile requirement.

5 Discussion

Mechanism. A representation induces a *model-dependent family* of metrics: a convolution, a Transformer, and a message-passing network do not see exactly the same one, and the fixed-encoder control holds the model constant so that the metric contrast is attributable to the representation. Strictly, the operative metric is

co-induced by the representation and the hypothesis class, not a property of the representation alone; our claim is therefore the conditional one, that at a fixed hypothesis class the representation moves the metric, which is exactly what the fixed-encoder control operationalizes. That metric fixes what counts as near, which fixes a neighborhood whose label homogeneity governs what a smoothness-biased learner finds easy. The ablation establishes this as a mechanism at the metric level; the minimality analysis bounds it: the active quantity is the class structure in the metric, and purity is one sufficient measure of it.

Representation as a hypothesis about relevant similarities. We offer, as a reading, that a representation is a hypothesis about relevant similarities (Edelman, 1998; Goodman, 1955; Quine, 1969); our addition is that this similarity, once measured, predicts and (at the metric level) causes sample efficiency.

Sample efficiency vs optimization vs generalization. Our protocol separates the structural part from raw content but not fully optimization ease, learning quality, and generalization; the direct measurement of Section 4.7 shows the transversal effect is a ceiling, not merely a rate.

Sample efficiency vs optimization. We oppose representation and optimization for clarity, but they interact: the ablation uses a kernel machine with no gradient descent, which partly isolates the effect from optimization dynamics, yet a representation also reshapes the loss landscape a first-order optimizer traverses, an interaction we do not study.

Limits. The demonstration is synthetic plus two real panels; the four synthetic worlds share a two-scale skeleton and one implementation; the ablation covers a subfamily of metric changes and is local to the metric; minimality is refuted, not established; the formal derivation is validated in form, not written as a bound. The predictor's validity is confined to learning dominated by the induced neighborhood (Section 4.11), so it covers much of modern machine learning but not learners whose bias is global or symbolic. Because the two poolings induce different input and minibatch distributions even at a fixed encoder, a residual distributional confound cannot be fully excluded. The task family is a constructed interpolation of a label unit, in the synthetic worlds and in the real panels alike; whether truly independent, non-interpolated tasks reproduce the same continuity is untested. Vision with spatial locality and language with sequential order are not tested in their own right, and a graph does not capture the translation-invariant geometry of a grid.

An engineering heuristic. Identify the unit of decision; build the ontology whose elementary objects coincide with it; verify it preserves the relevant dependencies; only then adapt the architecture. The ontology is not an unconditional gain but a bet on alignment, and this bet can be lost: the most instructive case is the one in which the richer representation learns the least, because its richness is orthogonal to the question.

6 Conclusion

At constant, provably equal information, the representation fixes the proximities a learner uses, and therefore its sample efficiency. We give a model-free structural quantity that predicts, in sign and magnitude, the learning gap between equivalent serializations; we show by an information-preserving intervention that it causally mediates the gap at the level of the metric; we show the phenomenon reproduces across four synthetic worlds and two real panels; and we show, against our own initial expectation, that the mediator is structural but not uniquely neighborhood purity. What is demonstrated is a causal, structural predictor of the learning gap; what remains is a written derivation and a demonstration beyond the two-scale, tabular-and-graph regime.

Broader Impact Statement

This work is methodological and uses synthetic generators and public, de-identified research datasets (dietox, BtheB, PBC). It suggests that at constant information the choice of data representation materially changes learnability, which can guide more sample-efficient and less compute-intensive modeling in longitudinal and relational domains, including healthcare. The same lever, used carelessly, can bias which patterns a model finds easy; practitioners should treat the ontology choice as a modeling decision with consequences for equity and error distribution, not as a neutral preprocessing step. No human-subjects data were collected; the clinical dataset used is a long-published, de-identified teaching dataset.

References

P. W. Battaglia, J. B. Hamrick, V. Bapst, A. Sanchez-Gonzalez, V. Zambaldi, M. Malinowski, et al. Relational inductive biases, deep learning, and graph networks. *arXiv preprint arXiv:1806.01261*, 2018.

M. Belkin, D. Hsu, S. Ma, and S. Mandal. Reconciling modern machine-learning practice and the classical bias-variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019. doi: 10.1073/pnas.1903070116.

Y. Bengio, A. Courville, and P. Vincent. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828, 2013. doi: 10.1109/TPAMI.2013.50.

M. M. Bronstein, J. Bruna, T. Cohen, and P. Veličković. Geometric deep learning: Grids, groups, graphs, geodesics, and gauges. *arXiv preprint arXiv:2104.13478*, 2021.

B. de Finetti. La prévision: ses lois logiques, ses sources subjectives. *Annales de l'Institut Henri Poincaré*, 7 (1):1–68, 1937.

S. Edelman. Representation is representation of similarities. *Behavioral and Brain Sciences*, 21(4):449–467, 1998.

N. Goodman. *Fact, Fiction, and Forecast*. Harvard University Press, Cambridge, MA, 1955.

A. Jacot, F. Gabriel, and C. Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 31, 2018.

T. M. Mitchell. The need for biases in learning generalizations. Technical Report CBM-TR-117, Rutgers University, 1980.

P. Nakkiran, G. Kaplun, Y. Bansal, T. Yang, B. Barak, and I. Sutskever. Deep double descent: Where bigger models and more data hurt. *arXiv preprint arXiv:1912.02292*, 2019.

W. V. Quine. Natural kinds. In *Ontological Relativity and Other Essays*, pages 114–138. Columbia University Press, New York, 1969.

N. Rahaman, A. Baratin, D. Arpit, F. Draxler, M. Lin, F. A. Hamprecht, et al. On the spectral bias of neural networks. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, volume 97 of *PMLR*, pages 5301–5310, 2019.

W. S. Robinson. Ecological correlations and the behavior of individuals. *American Sociological Review*, 15 (3):351–357, 1950.

E. H. Simpson. The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society, Series B*, 13(2):238–241, 1951.

N. Tishby, F. C. Pereira, and W. Bialek. The information bottleneck method. In *Proceedings of the 37th Allerton Conference on Communication, Control, and Computing*, pages 368–377, 1999.

D. H. Wolpert. The lack of a priori distinctions between learning algorithms. *Neural Computation*, 8(7): 1341–1390, 1996. doi: 10.1162/neco.1996.8.7.1341.

D. H. Wolpert and W. G. Macready. No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation*, 1(1):67–82, 1997. doi: 10.1109/4235.585893.

A Generator equations

World A (survival). Random effects $(L_i, \text{slope}_i) \sim \mathcal{N}(0, \Sigma)$; within process $w_i(t)$ is a linear trend plus a Gaussian AR(1); observed marker = latent + Gaussian measurement noise; hazard as in Eq. (1) with a Weibull h_0; event times by inverse cumulative hazard; administrative and random censoring.

World B (continuous panel). Level L_i from a two-component Gaussian mixture (bimodal); within process by mean reversion $w_{i,b} = \phi\, w_{i,b-1} + \epsilon$, $\epsilon \sim t_\nu$, plus a patient-specific sinusoid $A_i \sin(\omega_i t + \varphi_i)$; observed $= L_i + w + \text{osc} + t_4$ measurement noise; random visit dropout. No survival; the marker itself carries the two-scale signal.

World C (bipartite). User factors P_u, item factors $Q_i \sim \mathcal{N}(0, I_d)$; value $V_{ui} = \langle P_u, Q_i \rangle + a_u$ (item level set to zero so the per-item nuisance is a pure wave); observed $O_{ui} = V_{ui} + \delta_i +$ noise with a per-item shift δ_i; entries observed with density ρ (sparsity). Item-first pools the column (removes δ_i), user-first the row.

World D (graph). Stochastic block model with intra/inter edge probabilities; node signal $o_i = s_i + g_i +$ noise with a block-constant smooth part s_i and an i.i.d. local part g_i; graph-first context is the adjacency-neighbor mean (one message-passing step), feature-first the global mean. The local label is $g_i > 0$; the global label is s_i above its median.

B Encoders and hyperparameters

Sequence encoders: GRU (hidden 24–32) and a single-layer causally masked Transformer ($d_{\text{model}} = 20$, 2 heads). Person-period pooled MLP (hidden ≈ 58). Set encoder: ϕ, ρ MLPs with a mean-pooled context, hidden 32. Graph encoder: one-hop mean message passing with the same ϕ, ρ form, hidden 32. Optimizer Adam, learning rate 10^{-2}; epochs 100–400 by experiment; parameter budgets equalized within each comparison and logged. kNN and purity use $k = 20$–25. The RBF kernel machine uses a fixed $\gamma = 0.3$ and no standardization in the ablation.

C Pre-registration summary

Each independence and depth experiment was pre-registered with a UTC timestamp before execution, fixing the predicted sign, an amplitude band, and a refutation threshold. Timestamps: rang-0 controls 2026-08-08T18:44Z; world B 2026-08-09T13:46Z; world C 2026-08-09T14:18Z; world D 2026-08-09T14:45Z;

depth (ablation, minimality, derivation) 2026-08-09T15:01Z; edge (failure regime, quasi-bijection) 2026-08-09T17:56Z. All predictions and thresholds are released in the reproducibility package. We note one honest outcome: the failure-regime pre-registration guessed the kernel learner would escape the predictor, whereas the global linear learner did; the inversion is reported in Section 4.11.

D Robustness and correlation tables

The robustness matrix crosses four structural indicators (neighborhood purity at $k \in \{10, 25, 50\}$ and separability) with three learners (logistic regression, linear SVM, random forest) over ten seeds; the alignment-to-gap correlation stays in $[0.987, 0.999]$, with the random forest predicted at 0.999. Minimality: $R^2_{\text{purity}} = 0.992$, $R^2_{\text{sep}} = 0.997$, $R^2_{\text{both}} = 0.998$; partial correlations perf$\sim$sep|purity $= 0.864$, perf$\sim$purity|sep $= 0.480$. Broadened competition (post-hoc robustness, not pre-registered): correlation of each structural gap with the performance gap across the task family: separability 0.999, neighborhood purity 0.996, mutual information 0.991, geometric margin 0.987 (all label-aware), versus effective rank 0.000 and local intrinsic dimension 0.000 (both label-agnostic). The wave-free ceiling check (Section 4.7) and this broadened competition are post-hoc robustness analyses and were not part of the pre-registration.

E Real-data validation details

Panels are loaded from public R datasets (dietox, geepack; BtheB, HSAUR; pbc, survival). Each panel is arranged as a subject-by-visit matrix by last-observation-carried-forward; the two serializations are the matrix and its transpose (strict bijection). The label-unit sweep mixes standardized slice-context and subject-context components; the alignment gap is neighborhood purity (model-free), the performance gap is a logistic learner, over five random splits per w. Results: dietox $r = 0.983$; BtheB $r = 0.996$; both cross zero. PBC calibration targets (baseline cohort, $n = 418$): event rate (transplant or death) 0.385, bilirubin median 1.4 mg/dL (IQR 0.8–3.4), median follow-up 1730 days, median age 51.

F Reproducibility checklist

Code, seeds, pre-registrations, and a one-command reproduction script are released. All randomness is seeded; figures regenerate from saved result files. Synthetic generators are ex nihilo with released parameters; real panels are public. Compute is modest (CPU-only; the heaviest experiment, the per-subject set-encoder learning curve, runs in tens of minutes on a single core). No proprietary data or models are used.