VOCABULARY-RELATIVE TRACEABILITY: REPRESENTABILITY BEFORE COMPLEXITY

Anonymous authors
Paper under double-blind review

ABSTRACT

Exact faithfulness in interpretable machine learning requires a declared explanation vocabulary that can express the frozen model decision. We test this requirement before minimizing explanation complexity. For binary decisions, the traceability budget K is the minimum number of distinct named boundaries in an almost-surely exact Boolean transcript; $K = \infty$ if none exists. Under the stated regularity contract, an essential-boundary characterization yields three exhaustive obstructions: information loss, unavailable curved structure, and a missing flat direction. On finite samples, we compute maximum agreement over deterministic interface readouts, which also bounds every readout of every subinterface. A hard ECG bottleneck exposes a naming gap; PTB-XL supplies a separate post-hoc one-boundary witness resolving an eight-record sample obstruction. On public CUB concept-bottleneck checkpoints, an ordinal policy fixed before held-out endpoint outputs were opened refines only the displayed interface. Its primary confirmation-sample agreement ceiling rises from 98.83% to 100%, removing 34 unavoidable disagreements and establishing exact sample factorization with a finite transcript. Nearly half the records retain shared codes; two fixed checkpoints also reach zero conflict. With 16 bins per concept, exact sample factorization holds for one checkpoint but fails for two: resolution alone does not determine sufficiency. The audit determines when exact explanation requires interface augmentation rather than further readout optimization.

1 INTRODUCTION

Learned predictors compute with internal states, while explanations expose selected features, named concepts, measurements or rules. Attribution methods assign feature importance, concept bottlenecks predict through intermediate concepts, and rule or tree models use predicates (Sundararajan et al., 2017; Lundberg & Lee, 2017; Koh et al., 2020; Craven & Shavlik, 1995; Letham et al., 2015; Angelino et al., 2018).

Decision sufficiency of a learned representation does not guarantee an exact transcript in a declared vocabulary. A concept bottleneck makes the distinction concrete: its frozen head determines a decision from the exact concept state it consumes (Koh et al., 2020), while a binary display or named subset may discard distinctions the head uses. An internal-feature display can likewise omit relations among the features it exposes. We examine this distinction in a frozen language model by restoring the activation order of displayed feature IDs.

Representation and vocabulary are different failure points. A representation fiber is the set of inputs mapped to the same value of F. In B1, the frozen decision does not factor through F; on a presented sample, opposite decisions in one observed fiber witness sample-level failure. In B2 and B3, a representation-space decision exists, but no finite Boolean combination of the named thresholds in $\mathcal{W}$ agrees with it almost everywhere. Under the regularity contract, the required boundary patch is respectively curved or flat with an unnamed normal. A failed explanation search may therefore reflect weak search or an incapable interface. Only the former yields to further optimization.

Post-hoc measurement interfaces pose the same question. An electrocardiogram (ECG) records cardiac electrical activity through multiple leads; clinical interpretation summarizes it through named rates, intervals, axes, amplitudes and threshold criteria. This separation makes ECG useful for auditing a waveform model through a restricted measurement interface. Agreement certifies fidelity

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.

to the frozen model decision on the evaluated inputs; clinical correctness requires separate validation in the intended population (Estes, 2013; Attia et al., 2021).

This yields two questions in a fixed order. *Representability* asks whether any faithful transcript exists in the declared vocabulary. *Complexity* asks, conditional on existence, for the least number of named boundaries required. We call these Gate I and Gate II. Their ordering is not cosmetic: a large finite transcript and a nonexistent transcript require different responses. The audit fits between freezing a predictor and optimizing a faithful explanation over a chosen vocabulary. For a finite interface, its unavoidable disagreement gives an agreement ceiling that no deterministic readout over that interface can exceed.

We develop this order at three levels. First, a vocabulary-relative traceability budget separates exact existence from the minimum number of named boundaries. Second, an essential-boundary theorem and a regular obstruction classification distinguish information loss (B1), nonlinear boundary mismatch (B2) and a missing flat direction (B3). Third, each obstruction restricts the admissible augmentation class, which makes a bounded selection-and-certification test possible.

We examine learned and post-hoc interfaces with distinct evidence roles. Public CUB checkpoints first expose information lost when continuous concept states are displayed as binary or low-resolution attributes. On a previously unaudited Joint1 endpoint, an ordinal policy fixed before evaluation closes Gate I on a held-out confirmation half and supplies a finite sample transcript. A hard ECG bottleneck shows the distinct cost of omitting unnamed units. A GPT-2 sparse-feature display tests partial recovery when the same selected IDs retain their activation order. PTB-XL, a public collection of ten-second, twelve-lead ECG records with report-derived labels (Wagner et al., 2020), provides the fixed-artifact example: eight records share one declared nine-scalar vector while receiving both decisions, and a typed rate atom restores sample factorability on that fiber post hoc. Sleep-EDF sleep staging and ILPD hepatology exercise selected components under other vocabularies.

The quantified object differs from neighboring work. Concept completeness measures how much of a model's prediction behavior a concept set captures (Yeh et al., 2020); Gate I tests exact expression of one frozen decision over a separately declared vocabulary. Aggregation and tree-compression results instead approximate a base class or minimize an equivalent tree with inherited splits (Bressan et al., 2024; Vidal & Schiffer, 2020), while sufficient-subset work studies local or global feature subsets (Bassan et al., 2024; 2025). Mechanistic and verified explanations bind internal subnetworks or other explanation objects (Hsu et al., 2025; Adolfi et al., 2025; Méloux et al., 2025; Hadad et al., 2026; Wu et al., 2025; Saha & Bandyopadhyay, 2026). Our certificate is indexed by the frozen R, F, named atom vocabulary $\mathcal{W}$, measure and equality contract. Sparse-autoencoder benchmarks evaluate reconstruction, probing and intervention (Karvonen et al., 2025; Gao et al., 2025); our display audit instead computes the exact sample optimum over all deterministic readouts.

Figure 1 summarizes the two-gate order and the three regular obstructions.

2 A VOCABULARY AUDIT HAS TWO GATES

Let $(\mathcal{X}, \mathcal{A}, \mu)$ be a probability space, fix a measurable feature map $F : \mathcal{X} \to \mathbb{R}^m$, and write $\nu = F_{\#}\mu$. For $f \in \mathbb{R}^m$, the fiber $F^{-1}(\{f\})$ contains all inputs with the same representation value. A vocabulary $\mathcal{W} \subset \mathbb{R}^m \setminus \{0\}$ is a finite nonempty set of named directions. Pairing $w \in \mathcal{W}$ with $t \in \mathbb{R}$ defines the named half-space $H(w,t) = \{f : \langle w, f\rangle \geq t\}$; equivalently, $\mathbf{1}_{H(w,t)}$ is the corresponding yes–no vocabulary atom. Admissible transcripts are finite Boolean combinations of these atoms. Two decisions are treated as the same when they differ only on a μ-null set.

Write $\mathcal{HS}_{\mathcal{W}} = \{H(w,t) : w \in \mathcal{W},\ t \in \mathbb{R}\}$ and let $\mathrm{Bool}(\mathcal{HS}_{\mathcal{W}})$ be the algebra of its finite Boolean combinations, including $\varnothing$ and $\mathbb{R}^m$.

Definition 1 (Traceability budget)**.** For a binary decision $R : \mathcal{X} \to \{0,1\}$, let $K_{F,\mathcal{W}}([R])$ be the least N such that a region $D \in \mathrm{Bool}(\mathcal{HS}_{\mathcal{W}})$, generated by at most N distinct named hyperplanes, satisfies $R = \mathbf{1}_D \circ F$ μ-almost surely. Set $K_{F,\mathcal{W}}([R]) = \infty$ when no such region exists.

The budget counts distinct named boundaries, not gates in a Boolean circuit. Gate I asks whether $R = \mathbf{1}_D \circ F$ almost surely for some $D \in \mathrm{Bool}(\mathcal{HS}_{\mathcal{W}})$: a factorization through F by a vocabulary-admissible readout. Failure to factor through F at all is the B1 case. Under the obstruction regularity conditions stated below and in Appendix E, B2 and B3 instead occur when a representation-space

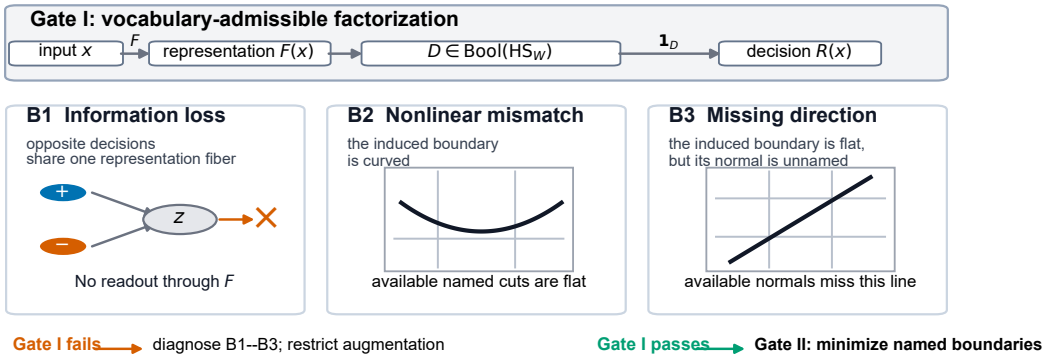


Figure 1: Vocabulary-admissible factorization and the three regular Gate-I obstructions. Gate I asks whether $R = \mathbf{1}_D \circ F$ almost surely for a $D \in \mathrm{Bool}(\mathcal{HS}_{\mathcal{W}})$. B1 prevents any readout through F; in B2 and B3, a representation-space decision exists but its regular boundary is respectively curved or flat with an unnamed normal. Gate-I success precedes Gate-II minimization. Failure restricts the admissible augmentation class; the controlled study receives B1, B2 or B3 from the generator rather than classifying it automatically.

decision exists but lies outside the Boolean language generated by the named atoms. Gate II asks for the minimum boundary count only after Gate I passes. On a presented sample, if $F(x) = F(x')$ but $R(x) \neq R(x')$, every readout over F is constant on the pair, regardless of its length. A population conclusion additionally requires the conflicting fiber to carry positive probability, as in the controlled B1 law below.

On a presented sample, this B1 obstruction has an exact test. Call records a tied group when they share the same declared feature vector.

Proposition 2 (Tied-group factorization)**.** *Let G be a tied group, let s be a score, and set $R_\tau = \mathbf{1}\{s \geq \tau\}$. The group is split exactly when $\tau \in (\min_G s, \max_G s]$. Hence a finite sample fails to factor if and only if τ lies in the union of these intervals over its tied groups.*

The same Gate-I certificate extends to a finite label set without extending the boundary theory for K. For a finite atom family A, write $a_A(x_i)$ for the joint atom readout, let Π_A be its partition of a presented sample, and let n_C^a count decision label a in cell C. Define

$$U_{\mathrm{mc}}(A) = \sum_{C \in \Pi_A} \left(|C| - \max_a n_C^a \right). \tag{1}$$

For a sample of size n, define the best deterministic agreement through the interface by

$$\mathrm{Acc}_n^\star(A) = \max_g \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\{g(a_A(x_i)) = R_i\} = 1 - \frac{U_{\mathrm{mc}}(A)}{n}. \tag{2}$$

Proposition 3 (Finite-label interface certificate)**.** *$U_{\mathrm{mc}}(A)$ is the minimum sample disagreement of any deterministic readout of A, and it is zero exactly when the observed decision factors through the interface. If $A \subseteq B$, then $U_{\mathrm{mc}}(B) \leq U_{\mathrm{mc}}(A)$ and $\mathrm{Acc}_n^\star(A) \leq \mathrm{Acc}_n^\star(B)$.*

The proof is cellwise and appears in Appendix M. For binary labels, $U_{\mathrm{mc}}(A) = \sum_C \min\{n_C^+, n_C^-\}$. Thus $U_{\mathrm{mc}} > 0$ certifies failure of exact factorization, while U_{mc}/n is the smallest unavoidable disagreement. If $S \subseteq A$, then $\mathrm{Acc}_n^\star(S) \leq \mathrm{Acc}_n^\star(A)$: the full-interface value is a certified agreement ceiling for every subfamily.

This identity gives the exact optimum on the presented sample. Certifying population agreement of at least $1 - \varepsilon$ with the frozen decision requires explicit sampling assumptions and a separate generalization argument. Conversely, when Gate I passes on a labeled sample, Gate II is a separation problem: choose the fewest named cuts so that no induced cell contains opposite labels.

Proposition 4 (Geometric read-off)**.** *Under the full-dimensional support contract in Appendix E, the exact budget of a supported fitted tree or Boolean circuit is the number of its named hyperplanes that carry a relatively open bichromatic boundary facet.*

Proposition 5 (Sample-presented computational hardness)**.** *On a labeled rational sample, deciding whether at most k named cuts suffice is NP-complete by cost-preserving equivalence with minimum set cover.*

Proposition 6 (Sample floors under a sampling law)**.** *If a finite sample has common marginal μ (independence is not required), then its minimum separating-cut count m_n satisfies $m_n \leq K_{F,\mathcal{W}}([R])$ almost surely.*

The two computational regimes explain the certificates used below. Visible geometry permits read-off. Sample separation requires lower and upper certificates because exact optimization can be hard. Proposition 6 is asymmetric: a sample cover never supplies a population ceiling. Its proof is in Appendix A; the complete complexity development is in Appendix E.

3 Three mechanisms of failed representability

For measurable $D \subset \mathbb{R}^m$, the support-based essential boundary records where both decisions persist at every scale and ignores null edits. Formally, $\partial^e_\nu D = \operatorname{supp}(\nu|_D) \cap \operatorname{supp}(\nu|_{D^c})$.

Theorem 7 (Essential-boundary characterization)**.** *Let $\nu \sim \lambda_m$ and let $\mathcal{A}$ be a finite named hyperplane arrangement. A measurable decision region D is almost everywhere a union of full-dimensional cells of $\mathcal{A}$ if and only if $\partial^e_\nu D \subseteq \mathcal{A}$. Consequently, when $[R] = [\mathbf{1}_D \circ F]$, $K_{F,\mathcal{W}}([R])$ is the least size of a finite arrangement of hyperplanes named by $\mathcal{W}$ whose union contains $\partial^e_\nu D$, and is infinite if no such arrangement exists.*

An infinite budget is not a single diagnosis. Suppose named hyperplanes have zero ν-mass, $\nu \sim \lambda_m$, and any factorable rule has a closed definable essential boundary with a finite C^2 Whitney stratification whose hypersurface strata are dense. Then $K = \infty$ exactly when B1, B2 or B3 occurs (Theorem 12; Table 1). B2 and B3 are exclusive on one stratum but may occur on different strata; rules outside this contract remain unclassified.

Table 1: Three mechanisms of failed representability and the intervention class each permits one to examine.

	What fails	Why more old cuts do not help	Admissible augmentation class
B1	R does not factor through F	Different decisions lie in one representation fiber	Add or refine information that separates the fiber
B2	A regular boundary patch is curved	A finite union of named hyperplanes cannot contain the patch	Add an appropriate named nonlinear scalar
B3	A regular boundary patch is flat but unnamed	Every available cut has the wrong normal	Add the missing direction or a scalar carrying it

Interface augmentation can change the representation, the vocabulary or both while keeping the decision fixed. B1 requires information absent from the base representation; new predicates of that representation cannot supply it. In prespecified B2 and B3 instances, the missing structure is nonlinear or directional, respectively. The diagnosis restricts candidate augmentation, not the scientific meaning of a new scalar.

Proposition 8 (Controlled theorem instances)**.** *For each prespecified controlled world, the base vocabulary has infinite budget. The B1 extension admits a two-boundary XOR transcript, so $K \leq 2$. Conditional on every fixed prespecified B2 permutation, the named product extension admits a five-boundary transcript under the almost-sure equality contract, so $K \leq 5$. Conditional on every fixed prespecified B3 direction, the added direction gives $K = 1$.*

The direct proofs are in Appendix B. The B1 and B2 conclusions are upper bounds because no matching lower bound is required. The B2 statement is a real-valued population result; its frozen float64 evaluator is an implementation predicate, not its proof.

4 WHAT THE AUDIT FINDS IN TRAINED SYSTEMS

We audit learned interfaces in vision, ECG and language, then examine a post-hoc measurement interface for an ECG black box. All predictors are frozen before audit. The evidence layers are kept separate because they answer different questions (Table 2).

Table 2: Evidence layers and their claim targets.

Evidence layer	Claim target
Direct interface evidence (released CUB checkpoints; fixed ECG bottleneck)	Resolution diagnosis, held-out fixed-policy augmentation and named-subset factorization
Direct display evidence (GPT-2 SAE; reused corpus)	Same-ID rank augmentation, partial sample recovery and occupancy-conditioned comparison
Fixed-artifact audit (PTB-XL)	Gate-I localization, Gate-II brackets and the fixed-eight witness
Controlled B1/B2/B3 worlds	Closure of a class-matched finite-library certification chain
Executability (Sleep-EDF; ILPD)	Instantiation under other targets, vocabularies and grouping rules
Negative and construction controls	Endpoint replay, answer-removed/wrong-class arms, exclusions and solver checks

4.1 CONCEPT BOTTLENECKS EXPOSE AND CLOSE INTERFACE GAPS

Resolution. A bottleneck guarantees factorization through the exact concept state consumed by its head, not through a coarsened display. On three released joint-sigmoid checkpoints of one recipe on CUB-200-2011 (Koh et al., 2020; Wah et al., 2011), binary and eighth-grid displays both retain positive U_{mc}; a validation-selected one-atom augmentation also fails to transport exactly. That complete negative census, including construction and near-duplicate controls, appears in Appendix M.

We next tested the diagnosis as an action rule on a previously unaudited Joint1 endpoint. The frozen head consumes 112 raw concept logits z; $p = \mathrm{sigmoid}(z)$ defines only the ordinal display. Refinement changes the displayed information while keeping the learned concept extractor, prediction head and target decisions fixed. The official $5{,}794$-image test split was divided outcome-blind into equal discovery and confirmation halves. Before opening Joint1 outputs, we fixed the q32 grid, Seed 1 as primary and Seeds 2–3 as non-rescuing replications. For $q \in \{2, 4, 8, 16, 32\}$, $\mathcal{I}_q$ contains the ordinal cuts $\mathcal{I}_q = \{\mathbf{1}[p_j \geq r/q] : j = 1, \ldots, 112,\ r = 1, \ldots, q-1\}$. Table 3 reports the confirmation endpoints; the complete threshold family and stopping rules are in Appendix M.4.

Table 3: Held-out fixed-policy repair on $2{,}897$ CUB confirmation images. Lower U_{mc} is better; zero is exact sample factorization. Seed 1 is the sole primary endpoint. The final column is the fraction of confirmation records in non-singleton q32 fibers.

Checkpoint	U_2	U_4	U_8	U_{16}	U_{32}	q32 non-singleton mass
Seed 1 (primary)	34	7	3	2	0	47.53%
Seed 2	25	5	1	1	0	45.53%
Seed 3	26	1	1	0	0	45.32%

On the primary confirmation half, the binary display caps deterministic model-replay agreement at $2{,}863/2{,}897 = 98.83\%$. The fixed q32 display removes all 34 unavoidable disagreements. This 1.17-percentage-point increase in the certified replay ceiling changes sample Gate I from failure to exact factorization while leaving $1{,}377/2{,}897 = 47.53\%$ of the records in non-singleton fibers. For the primary checkpoint, q16 appeared sufficient on discovery ($U_{16} = 0$) but left two unavoidable disagreements on confirmation; q32 is the first among the frozen grids with zero conflict on both halves. Writing $K_{n,\mathrm{mc}}$ for the number of threshold atoms in an exhibited sample-level multiclass transcript, the q32 construction gives $K_{n,\mathrm{mc}} \leq 112(32-1) = 3{,}472$; this is an upper bound, not

the minimum. Both fixed replications reach the same zero-conflict endpoint. A separate verifier reconstructed the decisions, partitions, metrics and transcripts. Appendix M.4 reports the full sequence, the earlier q16 negative stage on a different released recipe and the unsuccessful compact local secondary analysis. The learned-interface result is therefore a held-out fixed-policy repair of resolution, not a claim of compact concept discovery.

Naming. The ECG bottleneck supplies a complementary case. Its masked linear head consumes 22 hard concepts: 19 named clinical rules and three unnamed morphology anchors. On 2,198 patient-disjoint records, the full layer factors at all 66 disease-threshold cells. The named subset factors at $53/66$; all 13 failures occur for WPW, IRBBB and CLBBB, exactly the endpoints whose head mask uses an unnamed anchor. Removing the anchors merges 167 tied groups into 116 and raises the largest from 255 to 547 records. CUB exposes a resolution gap; this model exposes a naming gap (Appendix N).

4.2 Retaining feature order partially restores a language-model display

The same interface question applies within a language model. We used released sparse autoencoders (SAEs) to encode GPT-2 residual states on WikiText-103 (Radford et al., 2019; Merity et al., 2016; Bloom, 2024). At each position, the base display retained the unordered IDs of the three largest positive SAE activations. The augmented display retained their order, with the model and selected IDs unchanged. The target was GPT-2's own next-token argmax. Layer 8 was the primary analysis; the other three fixed layers and exact interface definitions appear in Appendix O.

On 120,000 positions, retaining rank reduced unavoidable disagreement from 17,610 to 12,791 (Table 4). On the same 37,107 positions in non-singleton ranked codes, rank still removed 712 disagreements, so the improvement was not entirely due to singleton separation.

Table 4: Partial recovery through the same feature IDs on GPT-2, layer 8, $k = 3$. Each entry gives U_{mc} and the certified best sample agreement. The second row excludes every singleton of the ranked display from both interfaces. The corpus is reused; these are sample optima, not fitted-readout accuracies.

Support	Positions	Unordered IDs	Ranked IDs
Full	120,000	17,610 / 85.33%	12,791 / 89.34%
Ranked non-singletons	37,107	13,503 / 63.61%	12,791 / 65.53%

Rank adds information and refines the partition, which can improve the sample optimum even under random label assignments. A within-parent reference preserving decision counts and child-cell sizes reduced conflict by 4,648.67 on average, compared with the observed 4,819. The excess was 170.33 disagreements, or 0.142 percentage points of the full sample. Thus rank partly restores this deliberately restricted display, while most of its total gain is also attainable under the occupancy reference. Appendix O.3 gives the fixed 499-assignment comparison and its descriptive scope.

4.3 A fixed-artifact PTB-XL audit and sample-factorization witness

We froze a twelve-lead ResNet1D trained on PTB-XL folds 1–9 and audited patient-disjoint fold 10 in a vocabulary of nine named ECG scalars. The LVH endpoint is an SCP-ECG report label. The estimand is exact factorization of the presented fold under the frozen scores and declared vocabulary.

Gate I. Fold 10 contains one tied group: eight records whose audited base representation encodes all nine named scalars as zero, although their wider feature records are not. At the top-440 cut, one record lies above the threshold and seven below. Proposition 2 therefore rules out every transcript over the declared vocabulary, at any budget. Across the eleven decisions, 30 of 66 operating points fail factorization for the main score artifact. Across six fixed score artifacts the count is 21–30 (median 28); every positive cell localizes to the same eight-record fiber. The six artifacts form the complete frozen census. Under the native numerical profile, excluding those eight records reduces the count from $30/66$ to $0/66$; closing the exclusion over patient identity also gives $0/66$. This is exact localization of the presented-sample obstruction (Appendix H.4).

Resolution control. On the same frozen decisions, quantile vocabularies failed at $66/66$ operating points for $q = 2$, $47/66$ for $q = 6$, and $30/66$ for every tested q from 8 through the native representation. The complete ten-resolution sweep and fixed-count reference are in Appendix K. The sweep quantifies resolution sensitivity for this frozen decision–vocabulary pair.

Write $U(W)$ for the binary specialization of Equation 1 under the atom family induced by vocabulary W. The all-zero fiber has $U(W_0) = 1$. In an explicitly post-hoc analysis restricted to these eight records, augmenting W_0 with the typed atom "median lead-II RR $< 400\,\mathrm{ms}$" gives $U(W_1) = 0$ and an exact one-atom transcript, $m_8(W_1) = 1$. A coarser $600\,\mathrm{ms}$ atom leaves $U = 1$ (Table 5).

Table 5: Post-hoc sample-factorization witness on the fixed eight-record fiber. A dash means that Gate II is not entered.

Vocabulary on the fiber	U	Gate I	m_8
W_0	1	Fail	–
$W_0 \cup \{\mathrm{median}(RR_{\mathrm{II}}) < 400\,\mathrm{ms}\}$	0	Pass	1
$W_0 \cup \{\mathrm{median}(RR_{\mathrm{II}}) < 600\,\mathrm{ms}\}$	1	Fail	–

Thus, on the presented eight-record fiber, augmentation closes both gates: representability changes from failure to success, and Gate II returns a one-atom exact transcript. The typed atom is a post-hoc sample-factorization witness. The controlled study below tests whether the same logic can be selected and certified under prespecified conditions when the obstruction class and candidate library are fixed in advance.

Same-record robustness checks preserve the truth vector of the $400\,\mathrm{ms}$ atom. The candidate and threshold were nevertheless chosen after inspecting the eight decisions and pilot measurements. The separated record carries the dataset's PSVT annotation, and the rate atom is not an LVH criterion (Appendix L).

Figure 2 places the Gate-I collision beside a separate factorable Gate-II example.

Gate II. At the separate factorable 0.5 operating point, an LP dual certifies at least eight named cuts and an eighteen-cut family separates every bichromatic pair, giving $m_n \in [8, 18]$. A known one-atom decision, Sokolow–Lyon ≥ 3.5, returns $[1, 1]$. Three of ten firing-rate-matched permutations are infeasible; the seven factorable permutation brackets overlap the network bracket.

At the raw LVH top-22 decision, let A_0 denote the nine-scalar base vocabulary and let $A_2 = A_0 \cup \{\text{Sokolow–Lyon}, \text{Cornell}\}$. An exact frozen-protocol search tightened the extended-vocabulary bracket from $[4, 8]$ to $[5, 8]$. Together with the base bracket, this gives $5 \leq m_n(A_2) \leq m_n(A_0) \leq 8$; equality and strict benefit from augmentation remain unresolved (Appendix I).

Negative and localization controls. Zero-anchored profiles and six frozen artifacts test persistence, while the closed exclusions localize the failure. A separate fiducial-measurement dictionary returned unavailable status for all 72 coordinates and no separating atom; the later rate atom uses a different raw-signal measurement. Deletion, adaptive one-bit, placebo and exact-solver controls are reported in Appendices J, G and H.

Cross-domain executability. Sleep-EDF and ILPD instantiate selected audit and compiler components under different targets, vocabularies and grouping rules, with residual local-rewrite gaps of $0/44$ and $0/8$. These are cross-dataset execution checks (Appendices F and G).

5 Obstruction-conditioned interface augmentation

Given a supplied obstruction class, we test closure of a restricted chain: selection from a frozen class-matched library, theorem-instance verification and evaluation on an independent re-audit sample. Selection uses discovery labels and candidate values; planted identity, population truth and the re-audit sample remain sealed. Algorithm 1 and the evidence contract are in Appendix C.

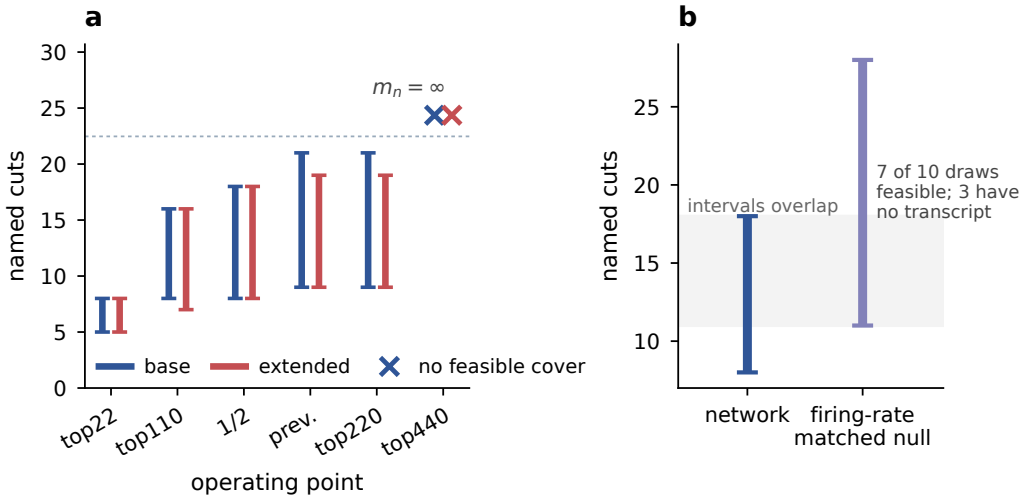


Figure 2: Two gates on the presented PTB-XL fold. **a**, certified lower and upper bounds on named cuts in the base and extended vocabularies. At the top-440 operating point, one tied representation fiber receives both decisions, so the sample-presented minimum is infinite. **b**, the separate factorable 0.5 decision against firing-rate-matched permutations; the brackets provide a matched reference.

5.1 The loop closes in controlled worlds

Each family contains 100 prespecified worlds and three paired arms: a positive class-matched library, an answer-removed library and a wrong-class library. B1 uses XOR with one missing bit, B2 a curved product boundary and B3 a non-axis-aligned half-space. Selection uses discovery data and evaluation uses a separately generated re-audit sample; the three estimands are not pooled.

Table 6: Controlled class-conditioned augmentation. Budget transitions are per-world analytic certificates; Y^{sci} is empirical positive-arm closure with marginal two-sided Wilson 95% intervals. AR/WC are accepted answer-removed/wrong-class certificates.

Family	Analytic certificate	$Y^{\mathrm{sci}}/100$ (95% CI)	Gate	Accepted AR/WC
B1	$\infty \to K \leq 2$	100/100 [96.3, 100.0]%	pass	0/0
B2	$\infty \to K \leq 5$	100/100 [96.3, 100.0]%	pass	0/0
B3	$\infty \to K = 1$	100/100 [96.3, 100.0]%	pass	0/0

The positive arm closes all 100 worlds in each family. In B2 and B3, the method, mutual-information baseline and generic univariate baseline recover every world; only the random baseline is weaker. These results establish closure of the supplied-class, fixed-library chain; the B2/B3 recovery rates do not distinguish informed selectors. Table 6 gives the exact counts; Appendix D gives the intervals, negative arms and paired contrasts.

6 What the certificates establish

A fitted deterministic readout provides one empirical agreement value. The finite-interface certificate instead bounds the agreement of every such readout of every subfamily on the presented sample, regardless of the optimizer or transcript length. For exact reproduction of the frozen decision, this separates an inadequate interface from an unsuccessful search: the former requires changing the interface, whereas the latter may be resolved within the existing vocabulary. A fixed-readout comparison illustrates the distinction between sample representability and learned fidelity (Appendix M.5).

Sparse rules, decision sets and optimal trees trade fit against structural complexity (Letham et al., 2015; Lakkaraju et al., 2016; Angelino et al., 2018). Once Gate I passes, our sample problem belongs

to the classical minimum-test-set/set-cover family (Moret & Shapiro, 1985; Garey & Johnson, 1979; Chlebus & Nguyen, 1998). The contribution is the preceding vocabulary audit and obstruction-conditioned evidence chain, not a new generic set-cover algorithm.

The evidence tiers are complementary rather than exchangeable. On the public Joint1 primary endpoint, binary coarsening forces $34/2{,}897$ disagreements; the fixed q32 policy removes them while preserving substantial shared-fiber mass. At q16, the three Joint1 checkpoints already differ in Gate-I status, while a different released bottleneck recipe retains positive conflict at all three fixed checkpoints. Interface adequacy is therefore indexed by the decision and interface, not by bin count alone. The fixed ECG artifacts use operating-point units: $13/66$ cells fail after unnamed bottleneck units are removed and $30/66$ PTB-XL operating points fail under the nine-scalar vocabulary. These quantities are not pooled. The controlled study tests bounded chain closure when the obstruction class and candidate library are known.

Each certificate is indexed by its decision, representation, vocabulary and equality contract; the fixed-eight rate atom is a post-hoc existence witness. An infinite budget rules out exact reproduction within the atom class, whereas a finite bracket establishes adequacy but may leave minimization open. Population and clinical conclusions require separate evidence.

7 Conclusion

Interpretability has a declared-interface adequacy problem before it has an optimization problem. The traceability budget first asks whether an exact transcript exists and only then counts its named boundaries. For finite interfaces, U_{mc}/n measures the least unavoidable disagreement and $1 - U_{\mathrm{mc}}/n$ the certified best agreement. On a public learned bottleneck, an execution-frozen ordinal interface converted a held-out binary obstruction into exact q32 sample factorization with a finite transcript. In a frozen language model, retaining feature order partly restored a restricted display; an occupancy-preserving reference quantified the finite excess over random refinement. The regular B1/B2/B3 classification and the remaining evidence tiers distinguish information loss from failures of the available boundary language. When Gate I fails, harder search over the same atoms cannot repair the interface. The appropriate order is existence, obstruction, constrained augmentation and, only after exact representability is established, explanation complexity.

AI use statement

The authors used OpenAI Codex and Anthropic Claude to assist with literature search, code review and debugging, and final language polish. The authors independently checked the cited sources, code changes and reported claims, and take responsibility for the final content.

Ethics statement

Only de-identified public or previously distributed secondary datasets were analyzed. No clinical care decision and no prospective participant interaction was involved. Users of these data must follow the licenses and access conditions of the source repositories. Sex and gender were not used as stratification variables in these methodological demonstrations, and the resulting absence of subgroup evidence limits clinical generalization. The reported transcripts were not evaluated for clinical use.

Reproducibility statement

Each quantitative claim is mapped to a stored artifact. Replayable formal claims additionally name a driver. Appendix C specifies the controlled augmentation study and its replay-bound analysis; Appendix F records the real-data protocols; Appendices K and L report the vocabulary-resolution audit and the explicitly post-hoc fixed-eight augmentation; Appendix I records the frozen one-use exact four-cut search; Appendix M, especially Section M.4, records the visual concept-interface diagnosis and execution-frozen ordinal repair; Appendix N records the ECG concept-interface audit; Appendix O records the language-model display analysis and its saved-data replay; Appendix H.4 reports the focused PTB-XL persistence census; Appendix E contains the full formal development; and Appendix H lists commands and artifact hashes. Statistics and decision branches described as prespecified were fixed before execution; the fixed-eight augmentation is labeled post hoc throughout.

REFERENCES

Federico Adolfi, Martina G. Vilas, and Todd Wareham. The computational complexity of circuit discovery for inner interpretability. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/1e45dc45f1649f880130169b2271129e-Abstract-Conference.html.

Elaine Angelino, Nicholas Larus-Stone, Daniel Alabi, Margo Seltzer, and Cynthia Rudin. Learning certifiably optimal rule lists for categorical data. *Journal of Machine Learning Research*, 18(234): 1–78, 2018. URL https://www.jmlr.org/papers/v18/17-716.html.

Martin Anthony and Peter L. Bartlett. *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, 1999. doi: 10.1017/CBO9780511624216.

Zachi I. Attia, David M. Harmon, Elijah R. Behr, and Paul A. Friedman. Application of artificial intelligence to the electrocardiogram. *European Heart Journal*, 42(46):4717–4730, 2021. doi: 10.1093/eurheartj/ehab649.

Shahaf Bassan, Guy Amir, and Guy Katz. Local vs. global interpretability: A computational complexity perspective. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 3133–3167. PMLR, 2024. URL https://proceedings.mlr.press/v235/bassan24a.html.

Shahaf Bassan, Guy Amir, Meirav Zehavi, and Guy Katz. What makes an ensemble (un)interpretable? In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 3142–3201. PMLR, 2025. URL https://proceedings.mlr.press/v267/bassan25a.html.

Saugata Basu, Richard Pollack, and Marie-Françoise Roy. *Algorithms in Real Algebraic Geometry*. Springer, second edition, 2006. doi: 10.1007/3-540-33099-2.

Henry C. Bazett. An analysis of the time-relations of electrocardiograms. *Heart*, 7:353–370, 1920. URL https://ndlsearch.ndl.go.jp/books/R100000136-I1571980074984147968.

Joseph Bloom. GPT2-Small-SAEs-Reformatted. Hugging Face model release, 2024. URL https://huggingface.co/jbloom/GPT2-Small-SAEs-Reformatted. Accessed September 9, 2026.

Marco Bressan, Nicolò Cesa-Bianchi, Emmanuel Esposito, Yishay Mansour, Shay Moran, and Maximilian Thiessen. A theory of interpretable approximations. In *Proceedings of the Conference on Learning Theory*, volume 247 of *Proceedings of Machine Learning Research*, pp. 648–668. PMLR, 2024. URL https://proceedings.mlr.press/v247/bressan24a.html.

Paul N. Casale, Richard B. Devereux, Paul Kligfield, Richard R. Eisenberg, David H. Miller, Banvir S. Chaudhary, and Malcolm C. Phillips. Electrocardiographic detection of left ventricular hypertrophy: Development and prospective validation of improved criteria. *Journal of the American College of Cardiology*, 6(3):572–580, 1985. doi: 10.1016/S0735-1097(85)80115-7.

Bogdan S. Chlebus and Sinh Hoa Nguyen. On finding optimal discretizations for two attributes. In *Rough Sets and Current Trends in Computing*, volume 1424 of *Lecture Notes in Artificial Intelligence*, pp. 537–544. Springer, 1998. doi: 10.1007/3-540-69115-4_74.

Vašek Chvátal. A greedy heuristic for the set-covering problem. *Mathematics of Operations Research*, 4(3):233–235, 1979. doi: 10.1287/moor.4.3.233.

Mark W. Craven and Jude W. Shavlik. Extracting tree-structured representations of trained networks. In *Advances in Neural Information Processing Systems*, volume 8, pp. 24–30. MIT Press, 1995. URL https://papers.nips.cc/paper_files/paper/1995/hash/45f31d16b1058d586fc3be7207b58053-Abstract.html.

Fernando De Ritis, Mario Coltorti, and Giuseppe Giusti. An enzymic test for the diagnosis of viral hepatitis: The transaminase serum activities. *Clinica Chimica Acta*, 2:70–74, 1957. doi: 10.1016/0009-8981(57)90027-X.

Irit Dinur and David Steurer. Analytical approach to parallel repetition. In *Proceedings of the 46th ACM Symposium on Theory of Computing*, pp. 624–633, 2014. doi: 10.1145/2591796.2591884.

Rodney G. Downey and Michael R. Fellows. *Fundamentals of Parameterized Complexity*. Springer, 2013. doi: 10.1007/978-1-4471-5559-1.

N. A. Mark Estes, III. Computerized interpretation of ECGs: Supplement not a substitute. *Circulation: Arrhythmia and Electrophysiology*, 6(1):2–4, 2013. doi: 10.1161/CIRCEP.111.000097.

Uriel Feige. A threshold of $\ln n$ for approximating set cover. *Journal of the ACM*, 45(4):634–652, 1998. doi: 10.1145/285055.285059.

Delbert R. Fulkerson and Oliver A. Gross. Incidence matrices and interval graphs. *Pacific Journal of Mathematics*, 15(3):835–855, 1965. doi: 10.2140/pjm.1965.15.835.

Leo Gao, Tom Dupré la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. Scaling and evaluating sparse autoencoders. In *International Conference on Learning Representations*, pp. 26721–26754, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/42ef3308c230942d223c411adf182c88-Abstract-Conference.html.

Michael R. Garey and David S. Johnson. *Computers and Intractability: A Guide to the Theory of NP-Completeness*. W. H. Freeman, 1979. ISBN 9780716710455.

Itamar Hadad, Guy Katz, and Shahaf Bassan. Formal mechanistic interpretability: Automated circuit discovery with provable guarantees. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/b5afe13494c825089b1e3944fdaba212-Abstract-Conference.html.

Alan J. Hoffman and Joseph B. Kruskal. Integral boundary points of convex polyhedra. In Harold W. Kuhn and Albert W. Tucker (eds.), *Linear Inequalities and Related Systems*, volume 38 of *Annals of Mathematics Studies*, pp. 223–246. Princeton University Press, 1956. doi: 10.1515/9781400881987-014.

Aliyah R. Hsu, Georgia Zhou, Yeshwanth Cherapanamjeri, Yaxuan Huang, Anobel Y. Odisho, Peter R. Carroll, and Bin Yu. Efficient automated circuit discovery in transformers using contextual decomposition. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/916ee60e315531d6b3954af8a8dc3437-Abstract-Conference.html.

David S. Johnson. Approximation algorithms for combinatorial problems. *Journal of Computer and System Sciences*, 9(3):256–278, 1974. doi: 10.1016/S0022-0000(74)80044-9.

Richard M. Karp. Reducibility among combinatorial problems. In Raymond E. Miller and James W. Thatcher (eds.), *Complexity of Computer Computations*, pp. 85–103. Plenum Press, 1972. doi: 10.1007/978-1-4684-2001-2_9.

Adam Karvonen, Can Rager, Johnny Lin, Curt Tigges, Joseph Isaac Bloom, David Chanin, Yeu-Tong Lau, Eoin Farrell, Callum Stuart Mcdougall, Kola Ayonrinde, Demian Till, Matthew Wearden, Arthur Conmy, Samuel Marks, and Neel Nanda. SAEBench: A comprehensive benchmark for sparse autoencoders in language model interpretability. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 29223–29264. PMLR, 2025. URL https://proceedings.mlr.press/v267/karvonen25a.html.

Bastiaan Kemp, Aeilko Zwinderman, Bert Tuk, Hilbert Kamphuisen, and Josefien Oberyé. The Sleep-EDF database [expanded]. PhysioNet, version 1.0.0, 2018. URL https://physionet.org/content/sleep-edfx/1.0.0/.

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 5338–5348. PMLR, 2020. URL https://proceedings.mlr.press/v119/koh20a.html.

Himabindu Lakkaraju, Stephen H. Bach, and Jure Leskovec. Interpretable decision sets: A joint framework for description and prediction. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1675–1684. ACM, 2016. doi: 10.1145/2939672.2939874.

Benjamin Letham, Cynthia Rudin, Tyler H. McCormick, and David Madigan. Interpretable classifiers using rules and Bayesian analysis: Building a better stroke prediction model. *Annals of Applied Statistics*, 9(3):1350–1371, 2015. doi: 10.1214/15-AOAS848.

László Lovász. On the ratio of optimal integral and fractional covers. *Discrete Mathematics*, 13(4): 383–390, 1975. doi: 10.1016/0012-365X(75)90058-8.

Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, volume 30, pp. 4765–4774. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html.

Maxime Méloux, Silviu Maniu, François Portet, and Maxime Peyrard. Everything, everywhere, all at once: Is mechanistic interpretability identifiable? In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/51157713f1f1f572a05647e1aef1b1ed-Abstract-Conference.html.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models, 2016. URL https://arxiv.org/abs/1609.07843.

Bernard M. E. Moret and Henry D. Shapiro. On minimizing a set of tests. *SIAM Journal on Scientific and Statistical Computing*, 6(4):983–1003, 1985. doi: 10.1137/0906067.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. Technical report, OpenAI, 2019. URL https://cdn.openai.com/better-language-models/language-models.pdf.

B. Ramana and N. Venkateswarlu. ILPD (Indian liver patient dataset). UCI Machine Learning Repository, 2022. URL https://doi.org/10.24432/C5D02C.

Sayan Saha and Sanghamitra Bandyopadhyay. Certified evaluation of model-level explanations for graph neural networks. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/c354d2a64a71776538d6dbbc38285216-Abstract-Conference.html.

Maurice Sokolow and Thomas P. Lyon. The ventricular complex in left ventricular hypertrophy as obtained by unipolar precordial and limb leads. *American Heart Journal*, 37:161–186, 1949. doi: 10.1016/0002-8703(49)90562-1.

Richard P. Stanley. An introduction to hyperplane arrangements. In Ezra Miller, Bernd Sturmfels, and Victor Reiner (eds.), *Geometric Combinatorics*, volume 13 of *IAS/Park City Mathematics Series*, pp. 389–496. American Mathematical Society, 2007. doi: 10.1090/pcms/013/08.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 3319–3328. PMLR, 2017. URL https://proceedings.mlr.press/v70/sundararajan17a.html.

Lou van den Dries. *Tame Topology and O-Minimal Structures*. Cambridge University Press, 1998. doi: 10.1017/CBO9780511525919.

Thibaut Vidal and Maximilian Schiffer. Born-again tree ensembles. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 9743–9753. PMLR, 2020. URL https://proceedings.mlr.press/v119/vidal20a.html.

Patrick Wagner, Nils Strodthoff, Ralf-Dieter Bousseljot, Dieter Kreiseler, Fatima I. Lunze, Wojciech Samek, and Tobias Schaeffter. PTB-XL, a large publicly available electrocardiography dataset. *Scientific Data*, 7:154, 2020. doi: 10.1038/s41597-020-0495-6.

Patrick Wagner, Nils Strodthoff, Ralf-Dieter Bousseljot, Wojciech Samek, and Tobias Schaeffter. PTB-XL, a large publicly available electrocardiography dataset (version 1.0.3). PhysioNet, 2022. URL https://physionet.org/content/ptb-xl/1.0.3/.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The Caltech-UCSD birds-200-2011 dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011. URL https://www.vision.caltech.edu/datasets/cub_200_2011/.

Wilson Wu, Louis Jaburi, Jacob Drori, and Jason Gross. Towards a unified and verified understanding of group-operation networks. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/9a583178bd18b2e55095c3a53f8fe88a-Abstract-Conference.html.

Chih-Kuan Yeh, Been Kim, Sercan O. Arik, Chun-Liang Li, Tomas Pfister, and Pradeep Ravikumar. On completeness-aware concept-based explanations in deep neural networks. In *Advances in Neural Information Processing Systems*, volume 33, pp. 20554–20565. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/hash/ecb287ff763c169694f682af52c1f309-Abstract.html.

A Main-to-appendix map and transferred proof

A.1 How the compact and full statements correspond

The main text states the load-bearing results with only the assumptions needed to follow the audit. The formal development in Appendix E supplies their complete contracts. Definition 9 is the full version of Definition 1; its historical label "transcription complexity" denotes the same quantity called the traceability budget here. Specifically, Theorem 11 is the full version of Theorem 7; Proposition 13 is the full version of Proposition 4; Proposition 14 strengthens Proposition 5 with the approximation, parameterized-complexity and empirical-frontier consequences. The full budgeted-frontier and distribution-free estimation results are retained in Appendix E as supporting scope results rather than a second main narrative. These pairs are compact and full-contract statements of the same results, not independent contributions.

Appendix K reports the complete vocabulary-resolution sweep, its exact fixed-positive-count reference and the deviations that restrict inference. Appendix L reports the fixed-eight rate atom, the exact $m_8 = 1$ argument and same-record robustness checks. These sections support the bounded PTB-XL statements in Section 4.3; neither supplies a population or clinical conclusion.

Appendix M proves the multiclass sample certificate and reports the complete visual concept-interface census. Appendix N reports the complementary hard-bottleneck named-subset audit.

A.2 Proof of Proposition 6

Proof. If $K = \infty$ the inequality is immediate. Otherwise take a population witness built from K named weak-threshold atoms and a Boolean readout. The witness agrees with R μ-almost surely. Because the sample is finite and every instance has marginal μ, the agreement holds for every sampled instance almost surely. Two instances with the same K atom readings receive the same Boolean output and hence cannot have opposite decisions. Those K cuts are feasible for the sample problem, so $m_n \leq K$. □

A.3 Scope of the imported audit commands

The mathematical development below is retained from v4. This derivative preserves the formal arguments while aligning notation, semantic cross-references and dataset-citation metadata; these differences are recorded in the release ledger. Appendix H records commands for the complete

research workspace, including drivers and large inputs that are not redistributed in the anonymous release package. Those commands are provenance records, not claims of standalone package replay. The anonymous package has one supported entry point, `python -B verify_v14r5.py`; it checks the frozen scientific payloads, headline ledgers, source identities, anonymous PDF and package closure without retraining or reconstructing omitted scores.

B CONTROLLED AUGMENTATION THEOREM INSTANCES

This section gives direct experiment-specific proofs because the generic full-dimensional theorem is not, by itself, the correct route for the discrete B1 law or the singular extended feature maps in B2 and B3. All population statements condition on the sealed world parameter. Machine replay is kept separate from these real-valued arguments.

B.1 B1: INFORMATION LOSS

Let $U, V, Z_1, \ldots, Z_9$ be independent Bernoulli$(1/2)$ variables and let $R = U \operatorname{xor} V$. The base representation is $F_0 = (U)$ with its coordinate direction. For either value of U, both labels have conditional probability $1/2$. Every transcript over F_0 is constant on the fiber, so $K_{F_0,\mathcal{W}_0}([R]) = \infty$. Extend the representation to $F_1 = (U, V)$. With $A = \mathbf{1}\{U \geq 1/2\}$ and $B = \mathbf{1}\{V \geq 1/2\}$, the rule $(A \wedge \neg B) \vee (\neg A \wedge B)$ is exact and uses two boundaries. Thus $K_{F_1,\mathcal{W}_1}([R]) \leq 2$. No one-boundary lower bound is claimed.

B.2 B2: NONLINEAR BOUNDARY MISMATCH

Fix a prespecified permutation $p = (p_0, p_1, p_2, p_3)$ and draw $X \sim \mathrm{Unif}(-1, 1)^4$. Write $\mathrm{sgn}_+(x) = 1$ for $x \geq 0$ and -1 otherwise, and define

$$R_p(X) = \mathbf{1}\left\{ \frac{\mathrm{sgn}_+(X_{p_0}) + \mathrm{sgn}_+(X_{p_1})}{4} + 3X_{p_2}X_{p_3} \geq 0 \right\}.$$

Any finite transcript over coordinate thresholds is constant on each open cell of a finite axis-aligned partition. In the positive–positive sign stratum, the target boundary contains the regular hyperbolic patch $X_{p_2}X_{p_3} = -1/6$. A point on this patch can be chosen away from every finite coordinate cut. A sufficiently small ball around it has positive-volume subsets on both sides of the target boundary within one partition cell, contradicting constancy. Hence the coordinate vocabulary has $K = \infty$.

Now name $Z_p = X_{p_2}X_{p_3}$. Let $A_i = \mathbf{1}\{X_{p_i} \geq 0\}$ for $i = 0, 1$. The Boolean rule

$$(A_0 \wedge A_1 \wedge \mathbf{1}\{Z_p \geq -1/6\}) \vee ((A_0 \operatorname{xor} A_1) \wedge \mathbf{1}\{Z_p \geq 0\}) \vee (\neg A_0 \wedge \neg A_1 \wedge \mathbf{1}\{Z_p \geq 1/6\})$$

equals R_p and uses five boundaries. Therefore $K \leq 5$ for every fixed prespecified p under the almost-sure equality contract. The sign-boundary set $\{X_{p_0} = 0\} \cup \{X_{p_1} = 0\}$ is null. The frozen float64 predicate reuses one stored product and the three exact binary64 thresholds; that replay convention does not prove this analytic statement.

B.3 B3: A MISSING DIRECTION

Fix a non-axis-aligned canonical primitive integer direction $a \in \mathbb{Z}^4$ with at least two nonzero entries. Let $X \sim N(0, I_4)$ and $R_a(X) = \mathbf{1}\{a^\top X \geq 0\}$. A finite coordinate-threshold transcript induces an axis-aligned partition. The target hyperplane is not a coordinate hyperplane and is not contained in the finite union of coordinate cut hyperplanes. Therefore a target-boundary point can be chosen in the interior of a partition cell; Gaussian density is positive on both sides of that cell. Thus the coordinate vocabulary has $K = \infty$. Adding $Z_a = a^\top X$ makes $\mathbf{1}\{Z_a \geq 0\}$ an exact one-boundary transcript. Since $a^\top X$ is nondegenerate, the target is not constant, and the extended budget is exactly $K = 1$.

C CONTROLLED AUGMENTATION PROTOCOL, ALGORITHM AND EVIDENCE CONTRACT

The controlled study contains three known-class experiments, 100 prespecified worlds per experiment, and the paired arms *positive*, *answer removed*, and *wrong class*. The complete matrix therefore

contains 900 cells. Discovery and re-audit samples are conditionally independent given the fixed world parameter. All arms within a world share the prespecified world and sample seed schedule. No world is replaced or excluded.

Selection receives the experiment identifier, bound seeds, formula commitments, static costs, discovery labels, discovery candidate values and discovery base values. It does not receive planted candidate identity, population truth, population certificates, or the independent re-audit sample. The generation-seal digest is an opaque integrity commitment and is not a ranking input. The finite-library verifier is experiment-specific and must return the exact prespecified PASS status for every row before any recovery probability is estimated.

Algorithm 1 (Prespecified class-conditioned selection, status-bound scoring and certification)**. Input.** Experiment class $e \in \{\mathrm{B1}, \mathrm{B2}, \mathrm{B3}\}$; discovery base representations $B \in \mathbb{R}^{n \times p}$, labels y and candidate columns $\{v_j\}_{j=1}^{m}$; opaque candidate identifiers, unique formula commitments, static costs and a tie seed; sealed re-audit data; prespecified finite verifier, theorem-instance binder and recovery equivalence.

1. Validate the exact selection-view schema, generation seal and formula-list alignment. Set $p_j = \mathrm{SHA256}(\texttt{str(seed)} \parallel \texttt{"|"} \parallel h_j)$. The planted identity, semantic map, population truth and re-audit data remain sealed.
2. Compute the prespecified discovery objective for every candidate. For B1, retain j exactly when no identical row of (B, v_j) carries both labels, with objective $(0, c_j)$. For B2, in each prespecified stratum $s \in \{-2, 0, 2\}$ search every midpoint between adjacent distinct values, the two outer `nextafter` thresholds and both threshold orientations; order local fits by $(\mathrm{error}, \texttt{threshold.hex}, \mathrm{orientation})$. Sum the stratum errors to E_j and set $K_j = 2+$ the number of distinct selected threshold hexadecimals; retain j exactly when $E_j = 0$ and $K_j \leq 5$, with objective (E_j, K_j, c_j). For B3, retain every candidate and set E_j to the number of mismatches between $\mathbf{1}\{v_j \geq 0\}$ and y, with objective (E_j, c_j).
3. Select the lexicographic minimum of $(\mathrm{objective}_j, p_j, \mathrm{candidate_id}_j)$ among retained candidates; abstain if none is retained. Compute the random, mutual-information and generic-univariate baselines separately; they cannot alter selection. Seal the proposal, objective and any discovery-fitted rules.
4. An independent finite verifier validates opaque identifiers, unique formula commitments, array and cost alignment, and the bound seals, then recomputes the class-specific optimum and fitted rules. Record the exact prespecified PASS status; a mismatch or non-PASS suppresses confirmatory estimation.
5. After selection is sealed, scoring opens the semantic map, tests the prespecified recovery equivalence and binds the theorem instance. It then runs the class-specific re-audit rule without reselection: the discovery lookup for B1; the prespecified stored-product predicate for a recovered B2 candidate (otherwise its sealed discovery-threshold rules); and the semantically bound direction for B3. Scoring may construct diagnostic rows after a verifier non-PASS, but such rows cannot enter the confirmatory estimand.
6. Analysis returns $Y^{\mathrm{sci}} = 1$ exactly when recovery holds, the finite status is the exact prespecified PASS, theorem binding is VERIFIED, and re-audit error is zero. Ordinary failures return zero; a finite-verifier non-PASS suppresses estimation.

Output. Abstention or a committed candidate and transcript, verifier and theorem statuses, re-audit error, and Y^{sci}. With hash-based row coding, selection costs $O(np{+}mn)$ for B1/B3 and $O(np{+}mn^2)$ for B2; the complete prespecified bundle is $O(np + mn^2)$ because the generic threshold baseline is also evaluated.

For a valid implementation, the primary endpoint $Y^{\mathrm{sci}}_{e,w}$ equals one exactly when the positive arm recovers the planted candidate or prespecified equivalence class, independent theorem-instance binding returns the prespecified certificate, and certified re-audit error is zero. Candidate nonrecovery, theorem-instance rejection, and nonzero re-audit error are scientific zeros in the fixed denominator. A finite-verifier non-PASS, status mismatch, or other integrity failure suppresses confirmatory estimation. The wording lock is:

```
FINITE_VERIFIER_NONPASS_SUPPRESSES_ESTIMATION;
ORDINARY_CANDIDATE_NONRECOVERY_REMAINS_IN_DENOMINATOR.
```

Primary output consists of three separate counts, rates and marginal two-sided Wilson 95% intervals. We also report the plug-in binomial Monte Carlo standard error and the worst-case $n = 100$ bound of 0.05 under the frozen generator/PRNG idealization. The operational gate is at least 95 successes in each experiment; it is not a confidence claim that $\pi_e \geq 0.95$. Negative-arm dispositions form a four-way exact partition and receive no probability intervals. Secondary comparisons use paired risk differences, $10{,}000$-replicate percentile-bootstrap intervals, exact two-sided sign tests excluding ties, and Holm adjustment over exactly nine prespecified comparisons.

Any formal run is governed by a one-use authorization, a write-once evidence store, a complete close seal, an independent scientific replay, and a replay-bound analysis envelope. Exact file identities, commands, replay status and hashes are listed in the release ledger rather than repeated in the scientific narrative.

D CONTROLLED AUGMENTATION RESULTS

The replay-bound envelope passed its integrity gate. Table 7 reports the registered primary summaries; Table 8 reports all nine secondary contrasts.

Table 7: Complete primary controlled-study summaries. MCSE is Monte Carlo variability under the frozen generator and PRNG idealization, not a population standard-error claim.

Family	Count	Rate	Wilson 95% CI	plug-in MCSE
B1	100/100	100.0%	[96.3, 100.0]%	0.0000
B2	100/100	100.0%	[96.3, 100.0]%	0.0000
B3	100/100	100.0%	[96.3, 100.0]%	0.0000

Table 8: Registered paired baseline comparisons. Risk differences and bootstrap intervals are in percentage points; exact sign-test p-values exclude ties, and Holm adjustment covers exactly these nine rows.

Family	Baseline	Method/base	RD (95% CI)	Ties	p	Holm p
B1	MI	100/7	93.0 [88.0, 97.0]	7	2.02×10^{-28}	1.41×10^{-27}
B1	univariate	100/8	92.0 [86.0, 97.0]	8	4.04×10^{-28}	2.02×10^{-27}
B1	random	100/6	94.0 [89.0, 98.0]	6	1.01×10^{-28}	8.08×10^{-28}
B2	MI	100/100	0.0 [0.0, 0.0]	100	1.000	1.000
B2	univariate	100/100	0.0 [0.0, 0.0]	100	1.000	1.000
B2	random	100/7	93.0 [88.0, 98.0]	7	2.02×10^{-28}	1.41×10^{-27}
B3	MI	100/100	0.0 [0.0, 0.0]	100	1.000	1.000
B3	univariate	100/100	0.0 [0.0, 0.0]	100	1.000	1.000
B3	random	100/1	99.0 [97.0, 100.0]	1	3.16×10^{-30}	2.84×10^{-29}

E THE FORMAL DEVELOPMENT IN FULL

The main text states the load-bearing results but omits several standing assumptions and supporting propositions. This section gives the full formal contract and the imported proof development in the order used by the argument. Elementary main-text propositions proved in place are not duplicated here.

E.1 FORMAL SETTING

Let $(\mathcal{X}, \mathcal{A}, \mu)$ be a standard Borel probability space and let $X \sim \mu$ denote the instance. Let $F : \mathcal{X} \to \mathbb{R}^m$ be measurable, and write $\nu = F_{\#}\mu$. A finite vocabulary is a nonempty set $\mathcal{W} \subset \mathbb{R}^m \setminus \{0\}$ of named directions. For $w \in \mathcal{W}$ and $t \in \mathbb{R}$, define

$$H(w,t) = \{f : \langle w, f\rangle \geq t\}, \qquad L(w,t) = \{f : \langle w, f\rangle = t\}.$$

The name belongs to the direction and its documented feature construction; the threshold remains a fitted scalar. The word *atom* does double duty below and the two senses are kept apart. An atom of the vocabulary is one named half-space indicator, the primitive a budget counts. An atom of the measure is a point of positive ν-mass, and it is that second sense the following condition rules out. Throughout, named hyperplanes are nonatomic:

$$\nu(L(w,t)) = 0 \quad \text{for every } w \in \mathcal{W}, t \in \mathbb{R}. \tag{3}$$

Binary rules are identified when they differ only on a null set:

$$R \sim_\mu R' \iff \mu(R \neq R') = 0, \qquad \mathfrak{R}_\mu = L^0(\mu; \{0,1\})/\sim_\mu .$$

A rule $[R]$ factors through F if $R = d \circ F$ a.e. for a measurable d. Standard Borelness allows the a.e. Doob–Dynkin factorization used here. Because $\mu(\mathbf{1}_D \circ F \neq \mathbf{1}_{D'} \circ F) = \nu(D \triangle D')$, every a.e. statement about a pulled-back rule reduces to a ν-statement about its feature-space region.

Two related hypothesis classes must be kept distinct. Let $\mathcal{G}_k = \mathcal{G}_k(\mathcal{W})$ be the concrete functions obtained by Boolean circuits over at most k atoms $\mathbf{1}\{\langle w, f\rangle \geq t\}$, $w \in \mathcal{W}$ and $t \in \mathbb{R}$, with the weak-inequality convention fixing values on hyperplanes. Let $\mathcal{C}_k = \{[g] : g \in \mathcal{G}_k\}$ be their image in the ν-a.e. quotient. Population risk and transcription complexity use $\mathcal{C}_k$; empirical labels, growth functions, and VC dimension use $\mathcal{G}_k$. This separation prevents arbitrary changes on a finite, hence ν-null, sample from changing the empirical class. For comparisons between different feature maps, define the pullback class on the common instance space

$$\mathcal{H}_k(F, \mathcal{W}) = \{[g \circ F] \in \mathfrak{R}_\mu : g \in \mathcal{G}_k(\mathcal{W})\}.$$

Only these pullback classes, not quotient classes living on different feature spaces, are compared by set inclusion.

Definition 9 (Transcription complexity)**.** Let $\mathcal{HS}_\mathcal{W} = \{H(w,t) : w \in \mathcal{W}, t \in \mathbb{R}\}$, and let $\mathrm{Bool}(\mathcal{HS}_\mathcal{W})$ be the algebra of sets it generates: the finite Boolean combinations of named half-spaces, together with $\varnothing$ and $\mathbb{R}^m$. Define

$$K_{F,\mathcal{W}}([R]) = \min\big\{N : \exists D \in \mathrm{Bool}(\mathcal{HS}_\mathcal{W}) \text{ with } R = \mathbf{1}_D \circ F \ \mu\text{-a.e.},\ D \text{ in the algebra generated by named half-spaces with at most } N \text{ distinct boundaries } L(w,t)\big\},$$

with value ∞ when no witness exists. Repeated use of the same geometric hyperplane counts once. Connectives and cell labels do not count.

This definition is invariant under null-set changes by construction. In particular, when ν is absolutely continuous with respect to Lebesgue measure λ_m, the indicator of a sphere with no interior is the zero rule a.e. and has $K = 0$; its nonempty topological boundary cannot create a spurious obstruction.

E.2 The representability boundary

For measurable $D \subset \mathbb{R}^m$, define the support-based essential boundary

$$\partial^e_\nu D = \mathrm{supp}(\nu|_D) \cap \mathrm{supp}(\nu|_{D^c}). \tag{4}$$

Equivalently, $z \in \partial^e_\nu D$ exactly when every ball centered at z has positive ν-mass in both D and D^c. The set (4) depends only on the ν-a.e. class of D.

The main geometric statement uses the transparent condition $\nu \sim \lambda_m$ on $\mathbb{R}^m$. A relative version of the equivalence (i)–(iii) of Theorem 11 holds on an open domain Ω when boundaries, cells, and connectedness are defined relative to Ω. The minimum-cover identity requires in addition that every cell of the named arrangement meet Ω in a connected set, as it does for convex Ω; otherwise a rule can take different values on two components of one cell, leave no essential boundary inside Ω, and force K above the minimal cover.

For the relative form, let $\Omega \subset \mathbb{R}^m$ be nonempty, open, convex and full dimensional, assume $\nu(\Omega) = 1$ and $\nu \sim \lambda_m|_\Omega$, and take supports, closures and interiors relative to Ω. Define

$$\partial^e_{\nu,\Omega} D = \mathrm{supp}_\Omega(\nu|_D) \cap \mathrm{supp}_\Omega(\nu|_{\Omega \setminus D}).$$

Define $K_{\mathrm{atom},\Omega,\mathcal{W}}([D])$ by Definition 9, with equality and Boolean representation required only ν-a.e. on Ω. The ambient definitions are recovered when $\Omega = \mathbb{R}^m$. For a compiler box B, the theoretical domain is $\Omega = \mathrm{int}\, B$; the discarded box boundary is Lebesgue-null.

Lemma 10 (Constancy away from the essential boundary). *If $\nu \sim \lambda_m$, $U \subset \mathbb{R}^m$ is nonempty, open, and connected, and $U \cap \partial^e_\nu D = \varnothing$, then either $\nu(U \cap D) = 0$ or $\nu(U \cap D^c) = 0$.*

Proof. See Appendix E.5.1.

Theorem 11 (Finite-arrangement characterization). *Assume $\nu \sim \lambda_m$. Let $\mathcal{A} = \bigcup_{j=1}^N L(w_j, t_j)$ be a finite named arrangement. For measurable D, the following are equivalent:*

(i) D is, a.e., a union of full-dimensional cells of $\mathcal{A}$;

(ii) $\mathbf{1}_D$ is a.e. constant on each such cell;

(iii) $\partial^e_\nu D \subseteq \mathcal{A}$.

Consequently, for $[R] = [\mathbf{1}_D \circ F]$,

$$K_{F,\mathcal{W}}([R]) = \min\big\{N : \partial^e_\nu D \text{ is covered by an arrangement of } N \text{ named hyperplanes}\big\}.$$

Proof. See Appendix E.5.2.

To state a finite obstruction taxonomy, geometry must be regular enough to exclude isolated lower-dimensional artifacts. Here and below, definability is with respect to one fixed o-minimal expansion of the real field. Whenever a rule factors as $[R] = [\mathbf{1}_D \circ F]$, assume that $E = \partial^e_\nu D$ is a closed definable set with a finite C^2 Whitney stratification into connected strata (a definable set has finitely many connected components, so this refinement costs nothing) and

$$E = \overline{E^{(m-1)}}, \tag{5}$$

where $E^{(m-1)}$ is the union of its hypersurface strata. Under the other hypotheses and $\nu \sim \lambda_m$, (5) is equivalent to E having empty interior. Closure preserves definable dimension (van den Dries, 1998, Ch. 4), so (5) gives $\dim E \le m-1$ and empty interior; conversely, if some $z \in E$ had a ball meeting no hypersurface stratum, the strata inside that ball would have dimension at most $m-2$, their complement in the ball would stay connected, and Lemma 10 would place z outside $\partial^e_\nu D$. We retain the closure form because the proof of Theorem 12 uses it directly.

Theorem 12 (Obstruction classification). *Assume* (3) *and $\nu \sim \lambda_m$, and restrict to rules that either fail to factor through F or whose essential boundary satisfies the standing regularity hypothesis* (5)*. Such a rule is not finitely transcribable precisely when at least one of the following holds:*

B1 *it does not factor through F;*

B2 *the rule factors through F and a hypersurface stratum of E contains a relatively open patch with nonzero second fundamental form;*

B3 *the rule factors through F and a flat hypersurface stratum has a normal not parallel to any $w \in \mathcal{W}$.*

B2 and B3 are mutually exclusive per stratum, but can occur on different strata of one rule. Rules whose essential boundary is not definable fall outside this classification.

Proof. See Appendix E.5.3.

When the decision is presented as a circuit over named atoms rather than as an abstract region, the characterization becomes a recipe.

Proposition 13 (Constructive audit compiler). *Under the relative contract above, let a finite binary tree or Boolean circuit T use a fixed, geometrically canonicalized family of named half-spaces, and let $\mathcal{E}_\Omega(T)$ be the presented hyperplanes carrying a relatively open bichromatic facet of positive $\mathcal{H}^{m-1}$-measure of the induced hard-decision region in Ω. Then*

$$K_{\mathrm{atom},\Omega,\mathcal{W}}([T]) = |\mathcal{E}_\Omega(T)|,$$

and $\mathcal{E}_\Omega(T)$ is the unique minimum geometric hyperplane family. The rule is a.e. constant on each reachable full-dimensional sign cell of this family, so those cell labels construct an equivalent Boolean

readout using exactly the minimum atoms. For an axis-aligned tree on a box, an essential hyperplane is witnessed by two oppositely labeled leaf regions—each leaf box intersected with Ω—that abut with positive overlap in every other coordinate. Consequently the number $K_{\mathrm{split}}(T)$ of distinct presented split hyperplanes satisfies $K_{\mathrm{split}}(T) \geq K_{\mathrm{atom}}(T)$.

Proof. See Appendix E.5.4.

When Ω and $\mathcal{W}$ are fixed, $K_{\mathrm{atom}}(T)$ below abbreviates $K_{\mathrm{atom},\Omega,\mathcal{W}}([T])$.

For the box-supported axis-aligned class, the accompanying audit compiler treats stored IEEE-754 thresholds as exact binary rationals and the stored tree arrays as a geometric decision map on exact real inputs. It emits facet lower-bound witnesses, the full common refinement by every presented split, an equivalent DNF, exact support-volume coverage, and hashes binding the tree, support and vocabulary. The reference verifier recomputes all full-dimensional refinement cells and formula labels. Compiler and verifier share the same exact interval implementation, so this is a machine-recomputed reference certificate rather than a second independently implemented proof checker. It also does not certify the distinct numerical map obtained when `sklearn.predict` casts inputs to `float32`; implementation-level prediction invariance is tested separately.

Atomic and logical complexity are distinct. For a basis $\mathcal{B} = \{\neg, \wedge_2, \vee_2\}$, define

$$\mathcal{Q}_{\mathcal{B}}(T) = \{(k, g) : \text{an equivalent circuit uses } k \text{ distinct named atoms and } g \text{ basis gates}\},$$
$$G_K(T) = \min\{g : (K_{\mathrm{atom}}(T), g) \in \mathcal{Q}_{\mathcal{B}}(T)\}.$$

Proposition 13 identifies the first coordinate, not G_K, BDD size, or any short-circuit/adaptive query cost. The compiler reports the exact gate cost and worst-case nonadaptive atom evaluations of the emitted DNF DAG. For solved small instances the DNF is lexicographically minimum in number of terms and literals; its gate count is still only an upper bound on minimum general-circuit cost.

Presentation is what makes this easy. When the decision is given only through a finite labeled sample, the freedom to choose the region returns and the problem changes character.

Proposition 14 (Complexity of the sample-presented problem)**.** *Given points $x_1, \ldots, x_n \in \mathbb{Q}^m$ with labels in $\{0, 1\}$, a finite vocabulary $\mathcal{W} \subset \mathbb{Q}^m$ with $|\mathcal{W}| = M$, and an integer k, deciding whether some arrangement of at most k named hyperplanes leaves no cell containing two oppositely-labeled points is NP-complete, and there are polynomial cost-preserving reductions in both directions between that problem and minimum set cover. All rational data are presented in binary and the input size is the total bit length of the n labeled points, the M vocabulary directions and k; thresholds are not part of the input, since only the Mn observed projections $\langle w, x_i \rangle$ need be considered and each is a rational of bit length polynomial in that of the input, so a budget-k arrangement is a polynomial-size certificate and the claim is well posed in the Turing model. Consequently, writing n_0, n_1 for the class counts, claims (i)–(iii) below concern feasible, nontrivial instances with $n_0, n_1 \geq 1$; one-class instances have optimum zero:*

- *(i) for every $\varepsilon > 0$ it is NP-hard to approximate the optimum within $(1 - \varepsilon)\ln(n_0 n_1)$;*
- *(ii) parameterized by k the problem is $\mathrm{W}[2]$-hard, so no algorithm running in $f(k) \cdot \mathrm{poly}(n, M, m)$ exists unless $\mathrm{FPT} = \mathrm{W}[2]$;*
- *(iii) a two-stage linear-programming rounding returns a feasible family of size at most M times the optimum, and the greedy rule returns one of size at most $1 + \ln(n_0 n_1)$ times the optimum;*
- *(iv) without the preceding feasibility and class-count restriction, deciding whether $\widehat{A}_n(k) = 1$ is NP-complete, so computing the empirical frontier $\widehat{A}_n(k)$ is NP-hard already in the realisable case.*

Proof. See Appendix E.5.5.

At the scale of this paper the vocabulary bound is the binding one: $M = 9$ for the base ECG vocabulary against $1 + 2\ln n \approx 20.8$ at $n = 19{,}600$. Part (iv) is the formal content of the admission in Section F that greedy CART does not solve the combinatorial supremum: the quantity whose deviation Proposition 17 bounds is itself NP-hard to evaluate, so the statistical and the computational statements must be read together.

Remark 1 (The two regimes and what separates them). Propositions 13 and 14 are the same functional evaluated under two measures, and the standing non-atomicity assumption (3) is the dividing line. Under $\nu \sim \lambda_m$ the essential boundary names its own hyperplanes, the minimizer is unique and no search occurs. Under the empirical measure of a sample, (3) fails maximally, since every hyperplane through a sample point carries mass at least $1/n$, and the constraint degenerates to separating bichromatic pairs, and the minimization returns. Evaluation is easy; synthesis is hard.

For semialgebraic D and vocabulary atoms represented by rational polynomials, the containment conditions can be encoded in the first-order theory of real closed fields and are decidable in principle (Basu et al., 2006). This is an effectiveness statement for an explicit input language, not an efficiency claim.

E.3 Budgeted accuracy and estimation

Let $Y : \mathcal{X} \to \{0, 1\}$ be a measurable label on $(\mathcal{X}, \mathcal{A}, \mu)$, and let $\eta(f) = \Pr(Y = 1 \mid F = f)$. The feature Bayes rule and value are

$$h^*_F = \mathbf{1}\{\eta \geq 1/2\}, \qquad B^*_F = \mathbb{E} \max\{\eta(F), 1 - \eta(F)\}.$$

The quotient image $\mathcal{C}_k$ defined in Section E.1 lives in $\mathfrak{R}_\nu = L^0(\nu; \{0, 1\})/\sim_\nu$, and $K_{\mathrm{id},\mathcal{W}}$ is transcription complexity read on the base space $(\mathbb{R}^m, \mathcal{B}, \nu)$ with $F = \mathrm{id}$. With that reading, $\mathcal{C}_k = \{[g] : K_{\mathrm{id},\mathcal{W}}([g]) \leq k\}$: each atom contributes one hyperplane, which gives $\subseteq$; conversely Boolean complementation uses $H(w, t)^c$ without another atom; under (3) it agrees a.e. with the opposite closed orientation whenever that orientation is named, so a Boolean combination with at most k distinct boundaries is realized ν-a.e. by at most k atoms. Define $A_{\mathcal{W}}(k) = \sup_{g \in \mathcal{C}_k} \Pr(g(F) = Y)$. The exact regret identity is

$$B^*_F - \Pr(g(F) = Y) = \mathbb{E}[|2\eta(F) - 1|\mathbf{1}\{g(F) \neq h^*_F(F)\}]. \tag{6}$$

Theorem 15 (Accuracy frontier)**.** *Assume* (3) *and* $\nu(\eta = 1/2) = 0$ *(used for (iv)), let* $\mathcal{W}$ *be finite, and suppose it contains the coordinate directions (used only for the convergence claim (ii)). Then:*

(i) $A_{\mathcal{W}}(k)$ *is the supremum of cellwise Bayes accuracy over partitions generated by at most* k *named hyperplanes;*

(ii) $A_{\mathcal{W}}(k)$ *is nondecreasing and converges to* B^*_F*;*

(iii) the supremum is attained for every finite k*;*

(iv) $A_{\mathcal{W}}(k) = B^*_F$ *if and only if* $K_{\mathrm{id},\mathcal{W}}([h^*_F]) \leq k$*.*

Thus a unique Bayes rule outside a finite budget class induces a strictly positive finite-budget gap.

Proof. See Appendix E.5.6.

The strict-gap conclusion depends on attainment. Without compactness, every candidate may be strictly suboptimal while a sequence approaches the Bayes value. More generally,

$$B^*_F - A_{\mathcal{W}}(k) = \inf_{g \in \mathcal{C}_k} \mathbb{E}[|2\eta - 1|\mathbf{1}\{g \neq h^*_F\}],$$

so strictness means that this infimum, not each individual regret, is positive.

For log loss, if $P = \phi(F)$ is a named partition, the information deficit separates as

$$H(Y \mid P) - H(Y \mid X) = I(Y; X \mid F) + I(Y; F \mid P). \tag{7}$$

The first term is measurement insufficiency; only the second is residual information lost by the partition. We use (7) as a decomposition, not as an identity with Shannon's classical rate–distortion function.

The capacity of a budget is governed by how many distinct atom readings a k-atom system can produce. For $\theta = ((w_1, t_1), \ldots, (w_k, t_k)) \in (\mathcal{W} \times \mathbb{R})^k$ write $\sigma_\theta(f) = (\mathbf{1}\{\langle w_i, f\rangle \geq t_i\})_{i \leq k}$ and

$$N_{\mathcal{W}}(k) = \max_{\theta \in (\mathcal{W} \times \mathbb{R})^k} \big|\{\sigma_\theta(f) : f \in \mathbb{R}^m\}\big|, \tag{8}$$

the largest number of distinct sign vectors the vocabulary admits at budget k. Two abbreviations are used below: $C_r(k) = \sum_{j=0}^{\min(r,k)} \binom{k}{j}$, the classical count of full-dimensional cells of an arrangement of k hyperplanes in general position in $\mathbb{R}^r$ (Stanley, 2007), and $G_r(k) = \max_{k_1 + \cdots + k_r = k} \prod_j (k_j + 1)$.

Proposition 16 (Reachable sign vectors). *Let $r = \operatorname{rank}(\operatorname{span} \mathcal{W})$. Then*

(i) $N_{\mathcal{W}}(k) \le C_r(k)$ for every vocabulary and every arrangement, degenerate ones included, and under the weak, the strict, or any mixed threshold convention;

(ii) $N_{\mathcal{W}}(k) = 2^k$ exactly when $\mathcal{W}$ contains k linearly independent directions; in particular $N_{\mathcal{W}}(k) = 2^k = C_r(k)$ for every $k \le r$ when $\mathcal{W}$ has full rank;

(iii) if $\mathcal{W}$ consists of r linearly independent directions then, writing $k = qr + \rho$ with $0 \le \rho < r$, $N_{\mathcal{W}}(k) = G_r(k) = (q+1)^{r-\rho}(q+2)^{\rho}$;

(iv) if $\mathcal{W}$ merely contains r independent directions then $G_r(k) \le N_{\mathcal{W}}(k) \le C_r(k)$, both of order k^r, and $C_r(k)/G_r(k) \to r^r/r!$ as $k \to \infty$.

Proof. See Appendix E.5.7.

Remark 2 (Sign vectors are not always cells). Under the weak-inequality convention a point lying on one named hyperplane inherits the sign vector of the closed side, so it adds no reading. When several tight directions are pointed, that is, $0 \notin \operatorname{conv}\{w_i\}$ over the tight indices, as holds whenever $0 \notin \operatorname{conv}(\mathcal{W})$ and hence for every coordinate vocabulary used here: the readings are in bijection with the full-dimensional cells of the arrangement. Without pointedness they are not: on $\mathbb{R}$ the atoms $\mathbf{1}\{x \ge 0\}$ and $\mathbf{1}\{-x \ge 0\}$ admit three readings and only two cells, the third being realized at the single point $x = 0$. Part (i) is stated so as to cover that case as well.

Proposition 17 (Growth and uniform estimation). *Let $M = |\mathcal{W}|$. The growth function of the concrete circuit class $\mathcal{G}_k$ satisfies*

$$\Pi_{\mathcal{G}_k}(n) \le [M(n+1)]^k \, 2^{\min\{n,\, N_{\mathcal{W}}(k)\}}.$$

Hence, for a universal constant C, on n i.i.d. draws from μ, with probability at least $1 - \delta$,

$$\sup_{g \in \mathcal{G}_k} |\operatorname{Acc}(g) - \widehat{\operatorname{Acc}}_n(g)| \le C \sqrt{\frac{N_{\mathcal{W}}(k) + k \log(M(2n+1)) + \log(1/\delta)}{n}}.$$

With $\widehat{A}_n(k) = \sup_{g \in \mathcal{G}_k} \widehat{\operatorname{Acc}}_n(g)$, the same bound controls $|A_{\mathcal{W}}(k) - \widehat{A}_n(k)|$, because $\operatorname{Acc}(g) = \Pr(g(F) = Y)$ is invariant under ν-a.e. modification and $\mathcal{C}_k$ is the image of $\mathcal{G}_k$, so $A_{\mathcal{W}}(k) = \sup_{g \in \mathcal{G}_k} \operatorname{Acc}(g)$. The capacity term is two-sided:

$$N_{\mathcal{W}}(k) \;\le\; \operatorname{VCdim}(\mathcal{G}_k) \;\le\; \max\{d \in \mathbb{N} : \, d \le N_{\mathcal{W}}(k) + k \log_2(M(d+1))\}.$$

Proof. See Appendix E.5.8.

For this construction weak sign readings confined to arrangement boundaries do not count. For a fixed nonempty open domain Ω, define the maximum number of full-dimensional readings

$$N^{\circ}_{\mathcal{W},\Omega}(k) = \max_{\theta} \left| \{\sigma : \operatorname{int}_{\Omega}(\sigma_\theta^{-1}(\sigma)) \ne \varnothing\} \right|.$$

Proposition 18 (The plug-in frontier is biased upward). *Let $d = N^{\circ}_{\mathcal{W},\Omega}(k)$ be attained by full cells $C_1, \dots, C_d$ in the open domain Ω. On each C_c, choose a normalized density that is strictly positive everywhere, mix the d cell laws with equal weights, and let Y be a fair coin independent of F. The resulting $\nu(\Omega) = 1$ and $\nu \sim \lambda_m|_\Omega$, each full reading has probability $1/d$, and $A_{\mathcal{W}}(k) = B^*_F = 1/2$. Write S_m for a sum of m independent Rademacher variables, $S_0 = 0$, and $\gamma(m) = \mathbb{E}|S_m|$. The empirical optimizer may adapt its thresholds, but the fixed maximizing atom system with cellwise empirical-majority labels is feasible; hence the bias obeys the finite-sample lower bound*

$$\begin{aligned} \mathbb{E}\, \widehat{A}_n(k) - A_{\mathcal{W}}(k) \;&\ge\; \beta(d,n) := \frac{d}{2n}\, \mathbb{E}_{N \sim \operatorname{Bin}(n,1/d)}\big[\gamma(N)\big], \\ \gamma(m) &= \frac{m}{2^{m-1}} \binom{m-1}{\lfloor (m-1)/2 \rfloor} \quad (m \ge 1), \qquad \gamma(0) = 0. \end{aligned} \tag{9}$$

No asymptotics enter: $\beta(d,n)$ is a finite sum, and $\gamma(m) \le m$ makes $\beta(d,n) \le 1/2$ for every d and n, as it must be since $\widehat{A}_n(k) \le 1$. The two ends of the range of d/n are two faces of one picture rather than competing bounds:

(i) sparse end. *If $n/d \to 0$ then $\beta(d,n) \to 1/2$; and if $n = o(\sqrt{d})$ then with probability tending to one no two sample points share a cell, so every labeling of the sample is realized and $\widehat{A}_n(k) = 1$.*

(ii) dense end. *If $d/n \to 0$ then $\beta(d,n) = (1+o(1))\,\frac{1}{2}\sqrt{2d/(\pi n)}$, the constant being $\frac{1}{2}\sqrt{2/\pi} = 0.3989\ldots \approx 0.399$.*

The closed form in (ii) holds only in that regime and may not be extrapolated: it exceeds the ceiling $1/2$ as soon as $d/n > \pi/2$, which β never does. For this constructed fair-coin law, $\widehat{A}_n(k) \geq 1/2 = A_{\mathcal{W}}(k)$ pointwise, so $\mathbb{E}|A_{\mathcal{W}}(k) - \widehat{A}_n(k)| = \mathbb{E}\,\widehat{A}_n(k) - A_{\mathcal{W}}(k)$. Therefore a mean absolute deviation of at most ε requires $\beta(d,n) \leq \varepsilon$; in the regime of (ii) that reads $n \geq (1+o(1))\,d/(2\pi\varepsilon^2)$, or $0.159\,d/\varepsilon^2$, while for $\varepsilon \geq 1/2$ it is no constraint at all. This constrains the expectation, not any single realized deviation.

Proof. See Appendix E.5.9.

Read in the other direction, the same two results answer a question a knowledge engineer can ask before collecting anything.

Corollary 19 (A distribution-free sample-size obstruction)**.** *Fix a coordinate vocabulary of rank r, a budget k, and $0 < \varepsilon < 1/2$. Every reading in a maximal coordinate arrangement can be made full dimensional in the open domain, so $N^{\circ}_{\mathcal{W},\Omega}(k) = G_r(k)$ by Proposition 16(iii). Write $n^*(r,k,\varepsilon)$ for the least $n \in \mathbb{N}_{\geq 1}$ with $\beta(G_r(k),n) \leq \varepsilon$. The map $n \mapsto \beta(d,n)$ is non-increasing, so n^* is well defined, and*

$$n < n^*(r,k,\varepsilon) \implies \textit{no distribution-free guarantee}$$
$$\mathbb{E}\left|A_{\mathcal{W}}(k) - \widehat{A}_n(k)\right| \leq \varepsilon \ \textit{ can hold},$$

because the law of Proposition 18 already violates it. Two consequences follow from the shape of $G_r(k)$. It never exceeds 2^k, with equality exactly when $k \leq r$, so the dense-end approximation is at most $0.159 \cdot 2^k/\varepsilon^2$. Increasing coordinate rank can increase $G_r(k)$, but the exact obstruction must be inverted separately; no strict monotonicity of the integer-valued n^* in r is claimed.

The condition is necessary, not sufficient. β is a lower bound from one construction rather than a minimax bound over laws, so a study with $n \geq n^*$ has not thereby shown its frontier estimate accurate to ε; it has only escaped the obstruction this construction raises. The table is the coordinate case. For a general vocabulary, boundary-only weak readings must be removed before applying this obstruction; substituting the all-readings count can overstate d.

Proof. See Appendix E.5.10.

Figure 3 evaluates the exact obstruction on the cohort sizes and vocabulary ranks used in this paper.

The capacity term can therefore not be replaced by anything smaller: no analysis of this class can improve $N_{\mathcal{W}}(k)$ by more than the additive direction-selection term. For the nine-scalar ECG vocabulary at $k = 8$ the bracket is $256 \leq \mathrm{VCdim}(\mathcal{G}_8) \leq 348$, and for the four-coordinate controlled vocabulary at $k = 64$ it is $83{,}521 \leq \mathrm{VCdim}(\mathcal{G}_{64}) \leq 84{,}696$. What remains one-sided is the passage from uniform convergence to the scalar $A_{\mathcal{W}}(k)$: we do not claim a minimax rate for estimating that quantity by an arbitrary procedure. Budget, dimension, and vocabulary size jointly control capacity; k alone does not.

E.4 Dictionary-selection identifiability

Selection of a named quantity is a separate problem from representation of a decision. Let $\mathcal{D} \subset \mathcal{Q}$ be a finite, prespecified dictionary with at least two elements, let $q \in \mathcal{Q}$ be a target, and let $V(a)$ be an explicit estimand, such as the AUC of a fixed univariate learner. It is not the value of an unrestricted $\sigma(a)$-measurable predictor. When $q \in \mathcal{D}$, define

$$\Delta_{\mathcal{D}}(q) = V(q) - \max_{a \in \mathcal{D}\setminus\{q\}} V(a).$$

Proposition 20 (Identifiability and robust certificate)**.** *If $q \notin \mathcal{D}$, exact dictionary selection is impossible. If $q \in \mathcal{D}$, it is the unique population maximizer if and only if $\Delta_{\mathcal{D}}(q) > 0$. Further, q is the unique maximizer of every $\widetilde{V}$ with $\sup_{a\in\mathcal{D}} |\widetilde{V}(a) - V(a)| \leq \varepsilon$ if and only if $\Delta_{\mathcal{D}}(q) > 2\varepsilon$.*

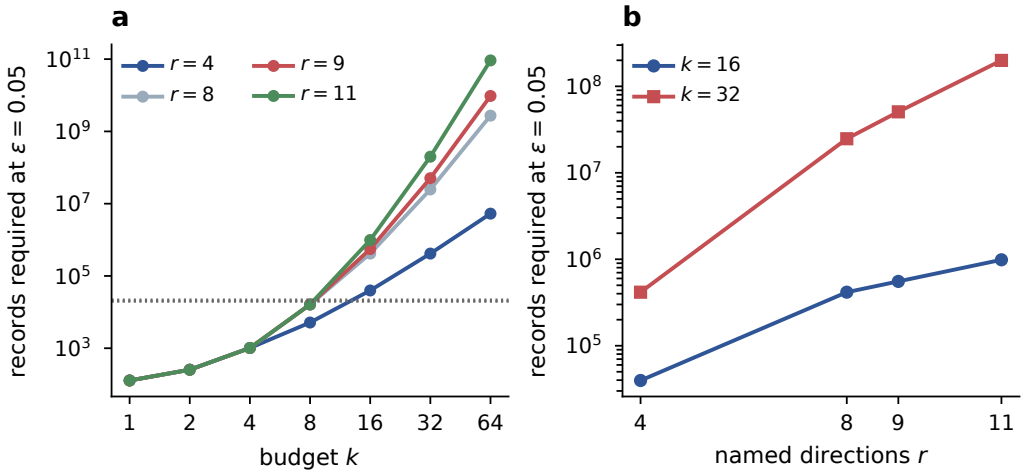


Figure 3: A counterfactual distribution-free sample-size obstruction. **a**, The least n at which the bound of Proposition 18 stops excluding a distribution-free guarantee of accuracy $\varepsilon = 0.05$, against the budget, for coordinate vocabularies of rank r. The dotted line marks the three cohorts analyzed here, which hold 19,600, 20,000 and 22,199 fitting records and coincide at this scale. Every curve crosses it between $k = 8$ and $k = 16$, so on the budget ladder of Section F this particular obstruction excludes a distribution-free frontier guarantee beyond $k = 8$. **b**, The same requirement against the number of named directions, at two budgets: a larger coordinate rank can enlarge the reading count and this obstruction, but the exact integer inversion can plateau. For $k \leq r$ the requirement does not depend on r, which is why the smallest-budget curves repeat across ranks. The condition is necessary and not sufficient, so a study above a curve has escaped this obstruction rather than established accuracy.

Proof. See Appendix E.5.11.

Failure of this worst-case certificate is not an information-theoretic impossibility. A necessary sample-size result would require an observation model and a lower-bound argument. Likewise, selection from a fixed dictionary does not recover causal or clinically valid knowledge.

E.5 Proofs of the results in the main text

E.5.1 Proof of Lemma 10 (Constancy away from the essential boundary)

Proof. Set $S_1 = U \cap \operatorname{supp}(\nu|_D)$ and $S_0 = U \cap \operatorname{supp}(\nu|_{D^c})$. If some $z \in U$ lay in neither, its two witness balls would meet in a ball $B \ni z$ with $\nu(B) = \nu(B \cap D) + \nu(B \cap D^c) = 0$, contradicting $\operatorname{supp} \nu = \mathbb{R}^m$; hence $S_0 \cup S_1 = U$; the boundary assumption gives $S_0 \cap S_1 = \varnothing$. Both sets are closed in U and complements of one another, hence open in U. Connectedness makes one empty, say S_1. Each $z \in U$ then carries a ball on which $\nu(\cdot \cap D)$ vanishes, and U is second countable, hence Lindelöf, so countably many such balls cover U and $\nu(U \cap D) = 0$. Only $\operatorname{supp} \nu = \mathbb{R}^m$ enters here, not the full equivalence $\nu \sim \lambda_m$. □

E.5.2 Proof of Theorem 11 (Finite-arrangement characterization)

Proof. The hyperplane union has zero ν-mass, and its complement has finitely many connected open cells. Thus (i) and (ii) are equivalent. A cellwise-constant representative has no essential boundary outside the arrangement, which proves (ii)⇒(iii). Conversely, each cell avoids the essential boundary, so Lemma 10 proves a.e. constancy and (iii)⇒(ii). Cell indicators are Boolean combinations of the defining half-spaces outside the null arrangement. Minimizing over arrangements gives the final identity. □

E.5.3 Proof of Theorem 12 (Obstruction classification)

Proof. B1 violates factorization. A curved C^2 patch cannot be covered by finitely many hyperplanes: if $L_1, \dots, L_N$ covered it, some L_j would meet it in a set of dimension $m-1$, and definability makes such a full-dimensional intersection contain a relative open set, which would force zero curvature there. A flat patch with an unnamed normal likewise cannot be covered by named hyperplanes. Conversely, suppose B2 and B3 fail. On a connected hypersurface stratum the second fundamental form is continuous, by C^2 regularity, so failure of B2 on every relatively open patch makes it vanish identically; the stratum is then flat, failure of B3 names its normal, and the stratum lies in a named hyperplane. There are finitely many strata. Equation (5) and closedness of the resulting arrangement cover all of E, and Theorem 11 applies. □

E.5.4 Proof of Proposition 13 (Constructive audit compiler)

Proof. Write $D_T \subseteq \Omega$ for the positive region. For the lower bound, take $L \in \mathcal{E}_\Omega(T)$ and a relatively open patch of positive $\mathcal{H}^{m-1}$-measure $P \subseteq L \cap \partial^e_{\nu,\Omega} D_T$. Any finite hyperplane family representing the same a.e. decision must cover this essential boundary. If it omitted L, its intersections with L would be a finite union of affine sets of dimension at most $m-2$, which cannot cover P. Geometric canonicalization therefore forces every representation to contain every member of $\mathcal{E}_\Omega(T)$.

For the upper bound, finite polyhedrality supplies the missing direction. The relative essential boundary of a finite hyperplane circuit is the closure, in Ω, of the union of its bichromatic full-dimensional facets. Every such facet lies on a presented hyperplane and makes that hyperplane essential. Hence

$$\partial^e_{\nu,\Omega} D_T \subseteq \bigcup_{L \in \mathcal{E}_\Omega(T)} (L \cap \Omega).$$

The relative arrangement theorem then makes $\mathbf{1}_{D_T}$ a.e. constant on every reachable full-dimensional sign cell of $\mathcal{E}_\Omega(T)$. Assign that fitted label to each reachable sign vector and treat unreachable vectors as don't-cares; a DNF over the positive reachable vectors is an equivalent Boolean readout using exactly this atom family. The lower and upper bounds agree, proving $K_{\mathrm{atom},\Omega,\mathcal{W}}([T]) = |\mathcal{E}_\Omega(T)|$ and uniqueness of the minimum geometric family after quotienting by scaling, orientation reversal and repeated identical thresholds. When $K_{\mathrm{atom}} = 0$, the empty disjunction and empty conjunction represent the constant-zero and constant-one rules, respectively.

For an axis-aligned tree on a box, intersect each leaf box with Ω. Two oppositely labeled full-dimensional leaf regions that abut across $\{x_d = t\}$ with positive overlap in every other coordinate provide a lower-bound facet witness. Enumerating all such facets and the complete arrangement-cell truth table provides the two certificate halves. Because every essential hyperplane is presented by the tree, $K_{\mathrm{split}} \geq K_{\mathrm{atom}}$. A positive atomic-count gap is therefore removable by unrestricted Boolean resynthesis; a positive residual after a specified tree-local rewrite says only that its grammar reached a fixed point. □

E.5.5 Proof of Proposition 14 (Complexity of the sample-presented problem)

Proof. Two points lie in a common cell exactly when no chosen hyperplane separates them, so the condition is that every bichromatic pair be separated. Under the weak-inequality convention it suffices to consider thresholds among the observed projections, since replacing t by the smallest observed projection at least t leaves the induced bipartition of the sample unchanged; there are therefore at most Mn combinatorially distinct candidate hyperplanes, computable by sorting. Membership in NP follows: guess k candidates and verify in $O(kn(m+1))$ that no cell is bichromatic.

For hardness, let $U = \{u_1, \dots, u_{n_U}\}$ and $S_1, \dots, S_N \subseteq U$ be a set-cover instance. Take $m = N$, $\mathcal{W} = \{e_1, \dots, e_N\}$, the point $q_0 = \mathbf{0}$ with label 0, and for each $u \in U$ the point $q_u = \sum_{d:\, u \in S_d} e_d$ with label 1. Every coordinate takes only the values 0 and 1 on this sample, so on direction d all separating thresholds are equivalent and choosing one is choosing S_d. The bichromatic pairs are exactly the pairs (q_0, q_u), and (q_0, q_u) is separated by direction d precisely when $u \in S_d$. A family of k named hyperplanes is therefore feasible if and only if the corresponding k sets cover U, and the two optima coincide. Minimum set cover is NP-complete (Karp, 1972), which gives NP-hardness; the converse reduction is the candidate–pair incidence just described.

Because the reduction preserves the objective exactly, (i) follows from the $(1-\varepsilon)\ln$ inapproximability of set cover (Feige, 1998), strengthened to NP-hardness in (Dinur & Steurer, 2014), and because it also preserves the parameter, (ii) follows from $\mathrm{W}[2]$-hardness of set cover parameterized by solution size (Downey & Fellows, 2013). Note that branching on which direction separates a chosen uncovered pair does not give a fixed-parameter algorithm: the threshold remains free, so the branching factor is the number of candidates separating that pair rather than M.

For (iii), the greedy bound is the classical set-cover guarantee (Johnson, 1974; Lovász, 1975; Chvátal, 1979) with ground set the bichromatic pairs. For the factor M, solve the linear relaxation with a variable per candidate and a covering constraint per pair. Each pair receives total mass at least 1, so some direction carries mass at least $1/M$ for it; assign the pair to such a direction. Within one direction the assigned pairs are intervals of the projected order and the constraint matrix has the consecutive-ones property, hence is an interval matrix and totally unimodular (Fulkerson & Gross, 1965; Hoffman & Kruskal, 1956), so the one-dimensional stabbing subproblem is solved exactly by the earliest-right-endpoint rule and its optimum is at most M times the mass that direction carries. Summing over directions gives a feasible family of size at most M times the linear-programming optimum, itself a lower bound on the integral optimum.

For (iv), $\widehat{A}_n(k) = 1$ holds exactly when some $g \in \mathcal{G}_k$ classifies every sample point correctly. A circuit is constant on each cell of its own arrangement and its readout is an unrestricted Boolean function of the reading, so such a g exists exactly when some arrangement of at most k named hyperplanes leaves no cell containing two oppositely-labeled sample points: if no cell is bichromatic, give each reading the common label of the sample points carrying it and any label where it carries none; conversely a bichromatic cell forces one of its two points to be misclassified. The argument runs directly on the sample and invokes no non-atomicity hypothesis, none being available here: the empirical measure violates (3) by construction. The decision problem is therefore the one just shown NP-complete, so computing $\widehat{A}_n(k)$ is NP-hard. □

E.5.6 Proof of Theorem 15 (Accuracy frontier)

Proof. For a fixed arrangement, majority labeling each cell is optimal, proving (i); nested classes give monotonicity. Enumerating rational coordinate thresholds yields increasing finite partitions whose sigma-fields generate $\mathcal{B}(\mathbb{R}^m)$. Lévy's upward theorem gives convergence of cell posterior means to η in L^1, and the map $u \mapsto \max(u, 1-u)$ is 1-Lipschitz.

For attainment, compactify each threshold to $[-\infty, +\infty]$; infinite thresholds encode constant atoms and budgets below k. There are finitely many direction tuples and Boolean truth tables. On each compact parameter block, threshold indicators converge pointwise away from the limit hyperplanes, which are null by (3); bounded convergence makes accuracy continuous. A maximum therefore exists. Finally, (6) and the no-tie assumption imply that an attained Bayes value is possible exactly when the Bayes rule belongs to $\mathcal{C}_k$. □

E.5.7 Proof of Proposition 16 (Reachable sign vectors)

Proof. (i) Let $S_m(k)$ be the supremum of $|\sigma_\theta(\mathbb{R}^m)|$ over all k-atom systems in $\mathbb{R}^m$, each atom the indicator of a half-space P_i, open or closed, with boundary L_i. Every fiber $\sigma_\theta^{-1}(\varepsilon) = \bigcap_i P_i^{\varepsilon_i}$ is an intersection of half-spaces, hence convex, and the fibers partition $\mathbb{R}^m$. Split off the last atom: with $A = P_k$, $B = P_k^{\mathsf{c}}$ and $\sigma' = (\sigma_1, \ldots, \sigma_{k-1})$,

$$\begin{aligned}|\sigma_\theta(\mathbb{R}^m)| &= |\sigma'(A)| + |\sigma'(B)| = |\sigma'(A) \cup \sigma'(B)| + |\sigma'(A) \cap \sigma'(B)| \\ &\le S_m(k-1) + |\sigma'(A) \cap \sigma'(B)|.\end{aligned}$$

If $\varepsilon' \in \sigma'(A) \cap \sigma'(B)$ then the convex set $\bigcap_{i<k} P_i^{\varepsilon'_i}$ contains a point on each side of L_k, so by the intermediate value theorem along the segment joining them it contains a point of L_k. Hence $\sigma'(A) \cap \sigma'(B) \subseteq \sigma'|_{L_k}(L_k)$, and restricting the first $k-1$ atoms to the affine hyperplane $L_k \cong \mathbb{R}^{m-1}$ leaves $k-1$ half-space atoms there, atoms with $w_i \parallel w_k$ becoming constant. Therefore $S_m(k) \le S_m(k-1) + S_{m-1}(k-1)$ with $S_m(0) = 1$ and $S_0(k) = 1$. The same recursion and boundary values hold for $C_m(k)$ by Pascal's rule, so induction on k gives $S_m(k) \le C_m(k)$; all atoms factor through the rank-r projection, so m may be replaced by r. The argument uses neither general position, nor distinctness of the L_i, nor the orientation convention.

(ii) If $w_1, \dots, w_k$ are independent, then for any ε the system $\langle w_i, f\rangle = t_i + (2\varepsilon_i - 1)$ is solvable, so all 2^k readings occur. Conversely suppose $\sum_i \lambda_i w_i = 0$ with $\lambda \neq 0$, and put $c = -\sum_i \lambda_i t_i$; replacing λ by $-\lambda$ if necessary, assume $c \leq 0$. Let $P = \{i : \lambda_i > 0\}$ and $N = \{i : \lambda_i < 0\}$, and note that $\sum_i \lambda_i(\langle w_i, f\rangle - t_i)$ is identically c. If $N \neq \varnothing$ or $c < 0$, the reading $\varepsilon_i = 1$ on P and $\varepsilon_i = 0$ on N makes every summand nonnegative, and strictly positive on N, so the sum would be $> 0 \geq c$ when $N \neq \varnothing$ and $\geq 0 > c$ otherwise. When $N = \varnothing$ and $c = 0$ that reading only forces the sum to be $\geq 0 = c$, which is no contradiction, and indeed it is realisable on the intersection of the boundary hyperplanes whenever that intersection is nonempty, since the weak-inequality convention reads 1 there; so take instead $\varepsilon_i = 0$ for every $i \in P$, a set nonempty because $\lambda \neq 0$, which makes every summand strictly negative and the sum $< 0 = c$. In both cases some reading is unrealisable and $N_{\mathcal{W}}(k) < 2^k$.

(iii) Coordinate directions are pointed, so by Remark 2 the readings are the full-dimensional cells. With k_j atoms on axis j the arrangement is a product of r one-dimensional arrangements, so the cells number $\prod_j (k_j + 1)$. Writing $u_j = k_j + 1 \geq 1$ with $\sum_j u_j = k + r$, the product of positive integers with a fixed sum is maximized at the balanced allocation, since $u_a \geq u_b + 2$ gives $(u_a - 1)(u_b + 1) > u_a u_b$; that allocation is $q + 2$ on ρ parts and $q + 1$ on the rest. Repeating a threshold merges hyperplanes and only decreases the count. A linear change of coordinates carries any r independent directions to $e_1, \dots, e_r$.

(iv) The lower bound is (iii) applied inside $\operatorname{span} \mathcal{W}$; the upper bound is (i). Finally $C_r(k) \sim k^r/r!$ and $G_r(k) \sim k^r/r^r$, so the ratio tends to $r^r/r!$. □

E.5.8 Proof of Proposition 17 (Growth and uniform estimation)

Proof. Fix $X = \{x_1, \dots, x_n\}$. For a fixed direction w the sets $\{i : \langle w, x_i\rangle \geq t\}$ obtained as t varies are the upper sets of the order that $\langle w, \cdot\rangle$ induces on X, of which there are at most $n + 1$, fewer under ties; so the restriction $\theta \mapsto \sigma_\theta|_X$ takes at most $[M(n + 1)]^k$ values. For a fixed θ, a Boolean readout is determined by its restriction to $\sigma_\theta(X)$, so the readouts realize exactly $2^{|\sigma_\theta(X)|}$ labelings of X, and $|\sigma_\theta(X)| \leq \min\{n, |\sigma_\theta(\mathbb{R}^m)|\} \leq \min\{n, N_{\mathcal{W}}(k)\}$ by (8). Multiplying the two counts gives the growth function. Standard growth-function uniform convergence then gives the displayed deviation bound (Anthony & Bartlett, 1999); the suprema over this uncountable class are measurable because $\mathcal{G}_k$ is indexed measurably by finitely many directions, thresholds and truth tables, which is the permissibility required. Accuracy depends only on the ν-a.e. class, so $\sup_{\mathcal{C}_k} \mathrm{Acc} = \sup_{\mathcal{G}_k} \mathrm{Acc} = A_{\mathcal{W}}(k)$, and $|\sup f - \sup \widehat{f}| \leq \sup |f - \widehat{f}|$ transfers the bound to the frontier.

For the two-sided statement, the lower bound is exhibited by a maximizing $\theta^\star$ in (8): choosing one point from each nonempty fiber gives $N_{\mathcal{W}}(k)$ points with pairwise distinct readings, and any labeling of them is realized by a readout defined on those readings, so they are shattered. For the upper bound, if d points are shattered then $2^d \leq \Pi_{\mathcal{G}_k}(d) \leq [M(d + 1)]^k 2^{N_{\mathcal{W}}(k)}$, that is $d \leq N_{\mathcal{W}}(k) + k \log_2(M(d+1))$, and the largest integer satisfying that inequality is the stated bound. The two sides do not meet: shattering may choose a different arrangement for each labeling, so the dimension can exceed the reading count. With $\mathcal{W} = \{e_1, e_2, e_3\} \subset \mathbb{R}^3$ and $k = 1$, the four points $(1, 3, 3)$, $(2, 1, 1)$, $(3, 4, 2)$ and $(4, 2, 4)$ are ordered differently by each coordinate, and the upper sets of those three orders together with their complements realize all sixteen labelings, while $N_{\mathcal{W}}(1) = 2$. □

E.5.9 Proof of Proposition 18 (The plug-in frontier is biased upward)

Proof. Let $d = N^\circ_{\mathcal{W},\Omega}(k)$ and fix a system attaining this maximum, whose full-dimensional cells are $C_1, \dots, C_d$. Mix the declared strictly positive cell densities with weights $1/d$. Arrangement hyperplanes are null and the full cells cover Ω up to those hyperplanes, so the resulting law satisfies $\nu(\Omega) = 1$, is equivalent to $\lambda_m|_\Omega$, and gives every cell probability $1/d$. Let Y be an independent fair coin. Then every rule has population accuracy exactly $1/2$, so $A_{\mathcal{W}}(k) = B^*_F = 1/2$. Write n_c^y for the number of sample points in full cell c carrying label y, and $n_c = n_c^0 + n_c^1$.

The d cell representatives carry pairwise distinct readings, so this N°-attaining system, equipped with the readout that gives each reading the majority label of its cell, lies in $\mathcal{G}_k$; hence $\max_{g \in \mathcal{G}_k} \#\{i : g(x_i) = y_i\} \geq \sum_c \max(n_c^1, n_c^0)$. The relation is an inequality and not an identity: a competing

budget-k system may place a hyperplane inside one of the full-dimensional cells and so adapt its arrangement to the sample. Only this direction is needed. With $\max(a,b) = \frac{a+b}{2} + \frac{|a-b|}{2}$ it gives, for every sample,

$$\widehat{A}_n(k) \;\geq\; \frac{1}{2} + \frac{1}{2n}\sum_{c=1}^{d} |n_c^1 - n_c^0| \;\geq\; \frac{1}{2}.$$

Because Y is a fair coin independent of F, conditionally on the occupancies $n_c^1 - n_c^0$ is distributed as S_{n_c}. For $m \geq 1$ the identity $(m-2j)\binom{m}{j} = m\big[\binom{m-1}{j} - \binom{m-1}{j-1}\big]$ telescopes over $0 \leq j \leq \lfloor (m-1)/2 \rfloor$, so the symmetry of S_m gives

$$\gamma(m) = \mathbb{E}|S_m| = 2^{1-m} \sum_{j<m/2} (m-2j)\binom{m}{j} = \frac{m}{2^{m-1}}\binom{m-1}{\lfloor (m-1)/2 \rfloor},$$

and $\gamma(0) = 0$. Each occupancy is marginally $\mathrm{Bin}(n, 1/d)$ and the d cells are exchangeable, so taking expectations in the display above term by term gives

$$\mathbb{E}\,\widehat{A}_n(k) - \tfrac{1}{2} \;\geq\; \frac{1}{2n}\sum_{c=1}^{d} \mathbb{E}\,\gamma(n_c) = \frac{d}{2n}\,\mathbb{E}_{N\sim\mathrm{Bin}(n,1/d)}\big[\gamma(N)\big] = \beta(d,n),$$

which is (9). No approximation is made at any step, and $\gamma(m) \leq m$ gives $\beta(d,n) \leq (d/2n)\,\mathbb{E}N = 1/2$, with equality only when the occupancies never exceed one.

Sparse end. Put $\lambda = n/d$. Since $m - \gamma(m)$ vanishes for $m \leq 1$ and is at most m otherwise, $\mathbb{E}[N - \gamma(N)] \leq \mathbb{E}[N\mathbf{1}\{N \geq 2\}] = \lambda\big[1 - (1-1/d)^{n-1}\big] \leq \lambda^2$, so $\beta(d,n) \geq \frac{1}{2}(1-\lambda)$ and $\beta(d,n) \to 1/2$ whenever $\lambda \to 0$. If in addition $n = o(\sqrt{d})$, the expected number of colliding pairs is $\binom{n}{2}/d \to 0$, so with probability tending to one each cell holds at most one sample point; the d readings being distinct, a readout then reproduces the observed labels exactly and $\widehat{A}_n(k) = 1$.

Dense end. Suppose $d/n \to 0$, so $\lambda \to \infty$. Cauchy–Schwarz gives $\gamma(m) \leq (\mathbb{E}S_m^2)^{1/2} = \sqrt{m}$, and Stirling applied to the closed form gives $\gamma(m)/\sqrt{m} \to \sqrt{2/\pi}$. The variance of N/λ is $(1-1/d)/\lambda \to 0$, so $N/\lambda \to 1$ in probability, hence $\gamma(N)/\sqrt{\lambda} \to \sqrt{2/\pi}$ in probability; that family is uniformly integrable because $\gamma(N)/\sqrt{\lambda} \leq \sqrt{N/\lambda}$, which is bounded in L^2 by $\mathbb{E}[N/\lambda] = 1$. Vitali's convergence theorem then gives $\mathbb{E}\,\gamma(N) = (1+o(1))\sqrt{2\lambda/\pi}$, so $\beta(d,n) = (1+o(1))\frac{1}{2\lambda}\sqrt{2\lambda/\pi} = (1+o(1))\frac{1}{2}\sqrt{2d/(\pi n)}$. This computation replaces any appeal to Jensen, which for the concave $u \mapsto \sqrt{u}$ bounds $\mathbb{E}\sqrt{N} \leq \sqrt{\mathbb{E}N}$ and so points the wrong way for a lower bound.

Since $\widehat{A}_n(k) \geq A_{\mathcal{W}}(k)$ pointwise,

$$\mathbb{E}\big|A_{\mathcal{W}}(k) - \widehat{A}_n(k)\big| = \mathbb{E}\,\widehat{A}_n(k) - A_{\mathcal{W}}(k) \;\geq\; \beta(d,n),$$

so $\beta(d,n) \leq \varepsilon$ is necessary for a mean absolute deviation of at most ε; in the dense-end regime the requirement reads $n \geq (1+o(1))\,d/(2\pi\varepsilon^2)$. It constrains the expectation only, not any single realized deviation. □

E.5.10 Proof of Corollary 19 (A distribution-free sample-size obstruction)

Proof. For a maximal coordinate arrangement, choose a ball contained in Ω and place each direction's thresholds inside a sufficiently small neighborhood within that ball. Every combinatorial reading counted by $G_r(k)$ then has a nonempty relatively open realisation in Ω, establishing $N^{\circ}_{\mathcal{W},\Omega}(k) = G_r(k)$. Substituting $d = G_r(k)$ into Proposition 18 exhibits a law under which $\mathbb{E}\,\widehat{A}_n(k) - A_{\mathcal{W}}(k) \geq \beta(G_r(k), n)$, and on that law $\widehat{A}_n(k) \geq A_{\mathcal{W}}(k)$ pointwise, so the mean absolute deviation equals the mean deviation and is itself at least $\beta(G_r(k), n)$. If $n < n^*$ then $\beta(G_r(k), n) > \varepsilon$, by the definition of n^* together with monotonicity, and the guarantee fails on that law; a guarantee that was to hold for every law therefore fails outright.

For monotonicity, write $Z_n = \sum_c |S_c^{(n)}|$ and condition on any realized labeled sample $\mathcal{D}_{n+1}$. Delete one observation uniformly. In a cell containing q signs $\varepsilon_i \in \{-1, 1\}$ with sum s,

$$\sum_{i=1}^{q} |s - \varepsilon_i| = (q-1)|s| \quad (s \neq 0),$$

whereas the left side equals q when $s = 0$; deletions outside that cell contribute $(n + 1 - q)|s|$. Summing over all $n+1$ deletions and then over cells yields $\mathbb{E}[Z_n \mid \mathcal{D}_{n+1}] \geq nZ_{n+1}/(n+1)$. Taking expectations and dividing by $2n$ gives $\beta(d, n) \geq \beta(d, n + 1)$.

For the two consequences, $G_r(k) = \prod_j (k_j + 1)$ with $\sum_j k_j = k$ gives $G_r(k) \leq \prod_j 2^{k_j} = 2^k$, since $k_j + 1 \leq 2^{k_j}$ for every non-negative integer, with equality throughout exactly when every $k_j \leq 1$, that is exactly when $k \leq r$. Non-decrease of $G_r(k)$ in r holds because any allocation over r directions extends to one over $r + 1$ by giving the new direction no threshold, which leaves the product unchanged, so the maximum over the larger set is at least the maximum over the smaller. This does not imply strict monotonicity of the integer inverse n^*: exact inversion can have plateaus. □

E.5.11 Proof of Proposition 20 (Identifiability and robust certificate)

Proof. The population statement is the definition of a unique maximizer. For sufficiency, $\widetilde{V}(q) - \widetilde{V}(a) \geq \Delta_{\mathcal{D}}(q) - 2\varepsilon > 0$. For necessity the perturbations range over every such $\widetilde{V}$: if $\Delta_{\mathcal{D}}(q) \leq 2\varepsilon$, setting $\widetilde{V}(q) = V(q) - \varepsilon$, $\widetilde{V}(a^*) = V(a^*) + \varepsilon$ at a strongest competitor a^*, and $\widetilde{V} = V$ elsewhere creates a tie or reversal while respecting the error bound; truncating $\widetilde{V}(a^*)$ to the range of V, as for an AUC in $[0, 1]$, keeps the tie. □

F Protocols, cohorts and source-locked specifications

Every protocol below was fixed in driver source before the corresponding run.

F.1 Experimental design

The experiments answer five bounded questions. First, does a vocabulary extension move the empirical performance curve under a requested fitted split cap that upper-bounds, but need not equal, the atomic complexity in Definition 9? Second, at a fixed operating point, does the fitted circuit change which ECG records a transcript refers? Third, could the decisions of a model fitted outside the declared vocabulary be transcribed inside it at all, and how many names would that take? Fourth, does a validation-set prediction about the strongest dictionary candidate agree with an independent test set? Fifth, how much of a greedy tree's presented-split-minus-atomic gap is removed by one prespecified, support-exact local rewrite, when inference treats an independently generated complete 5×7 cohort trajectory as the sampling unit? The release ledger distinguishes scientific payloads from runtime metadata and binds each headline artifact to its driver, declared dependencies and inputs. A listed `--selfcheck` is accepted only when it recomputes the scientific payload and returns a non-zero status on mismatch. Seeds are not uniform and are reported per experiment: the executed-complexity, capacity and bracket drivers use 20260713; the decision-consequence driver uses 20260713 with a separate bootstrap seed 20260714; the dictionary-selection driver uses $20260713 + i$ for its ith case; and the target-model driver uses 0 for both the PyTorch and the NumPy generator, under the determinism contract its ledger records. The named-circuit and dictionary-selection ledgers carry no seed field, so for those two the seed is fixed in the driver source and covered by the recorded driver hash.

F.2 Executable budgets

A binary CART is fitted with named scalar features. Its *split count* is the number of distinct (w, t) split hyperplanes in the final tree; repeated splits at the same threshold count once. The split count is what the curves are plotted against, and Proposition 13 shows it bounds the atomic vocabulary complexity of the induced hard-decision region from above without equalling it; Section G reports how far apart they are here. Requested fitted split caps are $b \in \{1, 2, 4, 8, 16, 32, 64\}$, imposed on the tree as `max_leaf_nodes` $= b + 1$, with a runtime assertion that the split count never exceeds b. A leaf cap is strictly stronger than $K \leq b$: an arrangement of b named hyperplanes over a rank-r vocabulary admits up to $N_{\mathcal{W}}(b)$ distinct readings (Proposition 16), 83,521 at $r = 4$, $b = 64$, and in general far more than $b + 1$, and a tree cannot reuse one hyperplane on two branches without spending an extra leaf. The searched family is therefore a proper subset of the budget class, and the reported curve is conservative twice over: greedy search inside a leaf-restricted subfamily. Tree leaves, nonzero

coefficients, and latent dimensions are reported only as diagnostics and never called K_{atom}. The count of distinct split hyperplanes is an upper bound on atomic vocabulary complexity, not a measure of emitted gate cost or a query policy; by Proposition 13 the exact atom value is obtained from the same tree by retaining only the hyperplanes that carry a bichromatic facet, and we report both. CART returns a feasible member of the budget class. Its test accuracy estimates that member's population accuracy; finite-sample noise means the observed curve is neither a deterministic lower bound on, nor an estimate certified to equal, $A_{\mathcal{W}}(b)$.

The controlled distribution has four independent $\mathrm{Unif}[-1,1]$ coordinates and posterior

$$\eta_\omega(x) = \sigma(3\,[(1-\omega)\{\mathrm{sgn}(x_1)+\mathrm{sgn}(x_2)\}+4\omega x_3 x_4])\,.$$

We use $\omega \in \{0, .25, .5, .75, 1\}$. The base feature map is $F_0(x) = x$ with coordinate directions. The extended feature map is $F_1(x) = (x_1, x_2, x_3, x_4, x_3x_4)$, again with coordinate directions; thus the composite is a named coordinate rather than a nonlinear atom inserted into $\mathcal{W}$. The law of F_1 lies on a four-dimensional graph in $\mathbb{R}^5$, so the full-density assumption of Theorem 11 does not apply to this lifted experiment. The hyperplane-null and frontier arguments do apply. Because every base atom remains available after lifting, $\mathcal{H}_k(F_0, \mathcal{W}_0) \subseteq \mathcal{H}_k(F_1, \mathcal{W}_1)$ on the original instance space. The posterior is given in closed form, so the Bayes rule and its transcription complexity in each vocabulary can be computed by hand; the ledger additionally records the Bayes accuracy of every arm, and Section G uses it as the reference level for the fitted curves. Training and test sets are independently generated. Hard-decision accuracy is primary; AUC is a diagnostic.

The $\omega = 0$ arm sits outside the hypotheses of Theorem 15. There $\eta_0(x) = \sigma(3\{\mathrm{sgn}(x_1)+\mathrm{sgn}(x_2)\})$ equals exactly $1/2$ whenever $\mathrm{sgn}(x_1)+\mathrm{sgn}(x_2) = 0$, an event of probability $1/2$, so the no-tie condition $\nu(\eta = 1/2) = 0$ fails on this arm with mass one half; for $\omega > 0$ the tie set is ν-null. Without that condition part (iv) of the theorem also fails here: under the weak-inequality convention the single atom $\mathbf{1}\{x_1 \geq 0\}$ already attains the Bayes value ($B_F^* = 0.7488$ analytically, 0.7457 as estimated in the ledger), while $K_{\mathrm{id},\mathcal{W}}([h_F^*]) = 2$. We report this arm as a budget-saturation control, not as an instance of the frontier theorem.

F.3 Prespecified local compression and trajectory inference

For this control, $b \in \{1, 2, 4, 8, 16, 32, 64\}$ denotes the requested CART split cap, K_{split} the number of distinct presented hyperplanes, and K_{atom} the essential-hyperplane count. The initial gap is $g = K_{\mathrm{split}} - K_{\mathrm{atom}}$. A deterministic local operator $\mathcal{A}$ first collapses same-label sibling leaves. It then constructs the left-child and right-child replacement candidates independently and accepts only replacements exactly equivalent to the fitted geometric tree on the declared box. If both are admissible, the frozen tie-break minimizes, in order, the resulting number of distinct presented hyperplanes, the number of reachable internal nodes, and then selects the left child. Reachable nodes are scanned in preorder; after each accepted rewrite the tree is compacted and scanning restarts until a second application changes nothing. The residual is $a = K_{\mathrm{split}}(\mathcal{A}(T)) - K_{\mathrm{atom}}(T)$. Thus $a > 0$ means only that this local grammar reached a fixed point; unrestricted Boolean resynthesis can still use the exact essential family from Proposition 13.

Each accepted rewrite must preserve every training prediction, the fitted tree's `float32` routing on the stored per-cell test table, and the exact geometric decision on every full-dimensional cell of the common box refinement. A 20,000-draw uniform support probe per circuit is an additional falsification check, not the equivalence proof. The permanent ILPD $b = 4$ counterexample requires a right-child splice while the left candidate fails, preventing regression to a left-only Pass B; repeated application is asserted to be a fixed point. Every cell records input, driver, dependency and prediction hashes, and `--selfcheck` recomputes rather than trusting stored summaries.

The synthetic sampling unit is one independently generated cohort of 30,000 records with one frozen data seed and a complete $5 \times 7\ \omega \times b$ trajectory. This trajectory experiment's 30,000-record cohorts are distinct from the earlier 20,000-draw single-seed frontier illustration. Within a trajectory, ω, budget and circuit rows are paired repeated measurements. The registered primary estimand conditions on non-constant greedy cells with $g > 0$ and is the ratio of sums

$$\rho_{\mathcal{A}} = \frac{\sum_s A_s}{\sum_s B_s}, \qquad A_s = \sum_i a_{si}, \quad B_s = \sum_i g_{si},$$

not a mean of cell or seed ratios. Whole trajectories are resampled in the cluster bootstrap. The planned prefixes were 30, 50, 75 and 100 trajectories; four-look Bonferroni adjustment gives a two-sided 98.75% interval, and sampling stops when its half-width is at most 0.05, at least 20 trajectories have positive denominators, and invalid bootstrap replicates do not exceed 0.5%. The complete all-cell 5×7 matrix is always reported. Because clinical controls retain constant cells, a separately labeled post-hoc sensitivity also adds constant positive-gap cells to the synthetic denominator while preserving whole-trajectory resampling.

For the clinical illustration, we use PTB-XL version 1.0.3 and its recommended folds and patient identifiers (Wagner et al., 2022). The locally pinned metadata contains 21,799 ten-second twelve-lead ECG records. Folds 1–9 train the circuit and fold 10 is the held-out evaluation fold—no circuit fitting or threshold selection uses it—with 2,198 records, 1,904 patients, and 214 positive LVH labels. One released record (ECG ID 5865, fold 4) had no waveform in the precomputed feature cache and was excluded, leaving an analytic cohort of 21,798; no fold-10 record was excluded. Metadata checks assert that no patient crosses folds and cached labels match metadata. Both PTB-XL analyses read the locked P1 feature-cache lineage `pr36p4_c1200`, and no untagged cache enters a reported PTB-XL result. The Sleep-EDF row of Table 11 is computed from a precomputed feature file whose name carries no lineage tag; it is pinned by SHA-256 in the ledger instead, as Appendix H records. The base vocabulary contains nine named ECG scalars, including five voltage coordinates plus QRS duration, axis, ST-J60, and T-wave amplitude, each of the last four taken as the median across the twelve leads; the extension adds separately named Sokolow–Lyon and Cornell constructions. Sokolow–Lyon is implemented as $S_{V1} + \max(R_{V5}, R_{V6})$, not as $S_{V1} + R_{V5}$, and Cornell remains a distinct construction (Sokolow & Lyon, 1949; Casale et al., 1985). Test uncertainty uses 2,000 paired patient-cluster bootstrap replicates and percentile intervals. These are descriptive resampling intervals; no superpopulation coverage is asserted. We do not report a bootstrap sign proportion as a p value, and cluster-wise score exchangeability is not assumed.

The local-compression control uses two related PTB-XL trajectories: folds 1–8 fit the circuit and fold 9 evaluates it, then folds 1–9 fit and fold 10 evaluates. Patients are disjoint across the fitting and evaluation folds in each trajectory; the two nested trajectories are not treated as independent replications. Sleep-EDF uses a frozen eight-subject fitting/two-subject evaluation split, plus eight leave-one-fitting-subject-out influence refits. Those refits are correlated sensitivity analyses, not eight independent cohorts. ILPD uses the full table because no patient identifier or external cohort is available, has no held-out endpoint, and remains diagnostic. Unlike the registered synthetic estimand, all three clinical controls retain constant cells.

This quantized vocabulary satisfies (3) only approximately. On the locked cache the largest single-value mass over all named directions is 0.0063 (S_{V1}, training); median QRS duration takes 5,140 distinct values among 19,600 training records, and its largest fold-10 point mass is 12 of 2,198 records at 90 ms. Eight fold-10 records (96 in training) have all nine named scalars exactly zero; all eight are LVH-negative. That extraction degeneracy is retained, unlike the single record excluded above for a missing waveform. Nothing reported here is derived from (3): split counts, accuracy, and the bootstrap intervals are measured on the fitted circuits, and the weak-inequality convention of Section 2 keeps those circuits well defined on ties. Theorem 11 is accordingly not asserted for this cohort, as it is not asserted for the lifted synthetic experiment.

F.4 A THIRD VOCABULARY

The read-off measurement is repeated on a vocabulary whose resolution profile is the opposite of the ECG one. The cohort is the Sleep-EDF feature cache the dictionary-selection analysis already uses: 27,597 thirty-second epochs from ten subjects, eight named spectral and artifact scalars, and the N1 sleep stage as the label, 1,005 positives at 3.64% prevalence. The declared extension is the three composites the frozen selection driver builds, θ/α, $\theta - \alpha$ and δ/α; the new driver calls that function and compares all eleven columns bitwise, so a different constant in the ratios would be caught rather than absorbed. Fitting uses the frozen driver's own subject-disjoint partition, 22,199 epochs against 5,398 held out with no subject in common, and the same budget ladder and `max_leaf_nodes` mechanism as the ECG arm. The essentiality criterion is not reimplemented: the new driver imports it from the executed-complexity driver, which is what makes the three vocabularies comparable.

Two properties of this cohort are settled before any circuit is fitted, and both matter later. Six of the eight base columns take a distinct value on every epoch, but `swa_frac` takes only 31 distinct values with 21.1% of its mass at zero. Against the ECG vocabulary's largest training point mass, 0.0063 on S_{V1}, that is a factor of 33; on distinct-value counts of each cohort's least-resolved direction, 31 against 5,140, a factor of 166. Equation (3) is therefore violated here far more sharply than on PTB-XL, and the ECG wording is not reused. Neither vocabulary contains a single exact-tie group. A feasibility census of the kind Section 4.3 runs is consequently model-independent on this cohort (with no tie group there is nothing for obstruction B1 to act on, whatever model is asked), so no classifier was trained for one, and the absence is a consequence of that measurement rather than a gap.

F.5 Referral consequence at a fixed operating point

The budget curves are accuracy curves. A third experiment fixes a label-defined threshold and asks what a budget costs in referrals. It reuses the PTB-XL setting without change: the locked cache `pr36p4_c1200`, folds 1–9 for fitting, fold 10 untouched for evaluation, and the extended vocabulary of eleven named columns. Two circuits are fitted under requested split caps 64 and 1; their exact atomic budgets are 59 and 1. Nothing is refitted on fold 10.

The policy is hypothetical and the label is a label. PTB-XL's LVH annotation is an SCP-ECG statement drawn from the report accompanying each record (Wagner et al., 2020); it is not an echocardiographic diagnosis, and no imaging was available to this study. What follows therefore counts LVH-*labeled* records under a hypothetical referral policy, and nothing in it is a statement about hypertrophy in the patients.

A record is referred when the training risk of the leaf it falls in is at least p_t, under the weak-inequality convention of Section 2. Net benefit is

$$\mathrm{NB}_k(p) = \frac{\mathrm{TP}_k(p)}{n} - \frac{p}{1-p} \cdot \frac{\mathrm{FP}_k(p)}{n},$$

and the pre-specified primary statistic is $1000\,[\mathrm{NB}_{64}(p_t) - \mathrm{NB}_1(p_t)]$; the subscripts denote requested split caps. It is evaluated at $p_t = 0.10$, in net LVH-labeled records identified per 1000 screened after assigning a penalty of one ninth of an identified record to each referral of an LVH-label-negative record. *Pre-specified* here means fixed in the driver source, whose SHA-256 the ledger records, before the number was computed; there is no external registry entry and none is claimed. That threshold is the smallest round value at which both trivial strategies are non-positive under the empirical label/prevalence arithmetic. This comparison does not establish clinical utility or identify a causal source of discrimination. The minimum interesting difference is 5.0. Five verdict branches (material, equivalent, unresolved, structurally null, and moot) were written out with their exact wording before the statistic was computed, and a volume-matched guard is mandatory on the material branch.

Uncertainty uses 2,000 paired patient-cluster bootstrap replicates at seed 20260714, with percentile intervals. They quantify resampling uncertainty. No p value is computed, no hypothesis is tested, and cluster-wise exchangeability is not assumed.

Three further quantities are registered. A label-blind gate computes, before the labels are read, the largest difference arithmetically compatible with the discordance between the two referral sets; a comparison it closes is not run. The volume-matched guard fixes a target referral volume of 214, the number of fold-10 positives, ranks leaves by training risk, and reports the capture count on both sides of the leaf that straddles the target, together with the yield per referred record at each arm's own nearest attainable volume. An exhaustive one-atom arm scores every midpoint between consecutive distinct training values, on every named column, in both orientations, by training net benefit, and evaluates the single maximizer once on fold 10; it exists so that the one-atom comparison cannot be dismissed as an artifact of greedy search.

Confusion counts are recorded at every budget in both vocabularies from the classifier's argmax decision, which fires at leaf risk strictly above one half. The referral analysis uses the weak convention instead. The two agree off records whose leaf risk is exactly one half, and those are counted rather than absorbed. Before any of this is computed the driver recomputes the frozen named-circuit ledger and halts unless every shared value reproduces.

F.6 A FROZEN BLACK-BOX TARGET

The budget curves and the referral analysis both fit circuits inside the declared vocabulary. A fourth experiment turns the question around: it takes a model fitted outside that vocabulary and asks whether its decisions could be transcribed inside it at all.

The target is a ResNet1D over the twelve raw leads, trained on folds 1–9 of the same locked cache and scored once on fold 10. The reference accuracy figures come from a separate, unpublished study of the same cohort; this paper re-uses that study's frozen artifacts rather than recomputing them. The network behind those figures was not recoverable, because its driver wrote neither a checkpoint nor a score vector. What is used here is a network of the same architecture and the same recipe, retrained under a determinism contract (fixed seeds, deterministic algorithms, no autotuning) on byte-identical inputs. It is not that network, and nothing below is a claim about its weights. Two tolerances were fixed in the driver source before the run: macro AUC within 0.010 of the earlier 0.9559, and LVH AUC within 0.020 of the earlier 0.9384. A run outside either is recorded as a re-instantiation rather than a reproduction, and the ledger carries whichever verdict fired.

The census itself is combinatorial and fits nothing. Group the fold-10 records by their exact named feature vector. A transcript over the declared vocabulary is a function of that vector, so it is constant on each group. If the network's decision is not constant on some group, then no transcript over that vocabulary agrees with it on this cohort at any budget: the decision does not factor through the names, which is obstruction B1 of Theorem 12 seen on a finite sample. If the decision is constant on every group, some finite named arrangement realizes it, and the census says nothing about how large that arrangement has to be. The census runs at six operating points: the fixed binary-decision threshold $\tau = 1/2$, the training-prevalence quantile, and the four thresholds firing on the top 22, 110, 220 and 440 records. At each it records the conflicting groups and the smallest number of records whose removal would restore constancy.

The census answers whether a transcript exists. A second computation asks how long it has to be. Call a pair of records bichromatic at a given threshold when the network's decision differs between them. A named cut is a threshold placed between two consecutive distinct values of a named column, and it separates a bichromatic pair when the two records fall on opposite sides of it. A family of named cuts realizes the decision exactly when it separates every bichromatic pair, so the shortest transcript over the declared vocabulary is a minimum set cover whose ground set is the bichromatic pairs and whose sets are the candidate cuts. That is Proposition 14 applied to the presented sample, and what it returns is the sample-presented minimum under the empirical measure, not the population complexity of Definition 9.

Because that problem is NP-complete, the experiment brackets it rather than solving it. The lower end is a feasible dual solution of the linear relaxation, found by row generation and then checked against every candidate cut, so the bound survives even if the solver stopped early; the reported value is its ceiling. The upper end is an exhibited family verified to separate every bichromatic pair. Three are built (greedy covering, two-stage rounding of the relaxation, and a decision tree fitted on ranks so that no float cast can merge two records), and each is reduced to an irredundant subfamily, which is minimal and not minimum. The smallest verified family is the upper bound. A pre-committed permutation control repeats the entire computation on ten decision vectors carrying the network's firing rate permuted across records; the rule for reading it and the sentence for each outcome were fixed in the driver source before any number existed. An anchor instance whose answer is known, the decision $\{\text{Sokolow–Lyon} \geq 3.5\}$, is required to come back bracketed at exactly one.

F.7 A POST-V1 ONE-BIT COLLISION-REFINEMENT DIAGNOSTIC

This diagnostic was designed after the v1 census had exposed the eight-record fiber. It is therefore adaptive, not preregistered from the outset. Its narrower question is still falsifiable: can one non-injective program bit, chosen without the evaluation score values, open a measurable gap between the score intervals inside that fiber? Record identity, patient identity, labels and model scores were forbidden as candidate quantities.

Fourteen frozen provenance-linked program quantities were crossed with sixteen frozen bit-depth quantizers. For stored binary32 candidate value z, the implemented operators are

$$Q_1(z) = \mathbf{1}\{z \geq \tfrac{1}{2}\}, \qquad Q_b(z) = \mathrm{round}_{\mathrm{even}}\big((2^b - 1)z\big), \quad b = 2, \ldots, 16,$$

where ties in the latter operation round to the even integer. This gives 224 candidate–quantizer witnesses before this successor decoded the design or evaluation score values. Exact screening removed 162 trivial witnesses and 34 witnesses that isolated a singleton. The remaining 28 admissible witnesses deduplicated to two partitions of the same eight ECG records: P35, with block sizes $5/3$, and P44, with block sizes $4/4$. Neither partition is injective and neither has a singleton block. The selected real partition, P35/c10-Q1 with restricted-growth signature `00001101`, places ECG records $\{4922, 7445, 7889, 8263, 19305\}$ and $\{10160, 18916, 20966\}$ in its two blocks. The eight records belong to eight different patients. The blocks have no asserted clinical or physiological meaning.

For score artifact s, endpoint e, partition $\mathcal{P}$ and block B, write

$$I_e(B; s) = \left[\min_{i \in B} s_{ie}, \max_{i \in B} s_{ie}\right], \qquad \Phi_s(\mathcal{P}) = \frac{1}{11} \sum_{e=1}^{11} \lambda\left(\bigcup_{B \in \mathcal{P}} I_e(B; s)\right).$$

The base partition $\mathcal{P}_0$ has one eight-record block, and the registered contrast is $\Delta\Phi_s(\mathcal{P}) = \Phi_s(\mathcal{P}_0) - \Phi_s(\mathcal{P})$. Every interval of a refined block is contained in the base interval. Hence $\Delta\Phi_s(\mathcal{P}) \geq 0$ for every score array by construction. For two blocks and one endpoint, the contribution is the gap between their score intervals when they are disjoint, and zero when they overlap. A positive value measures this finite interval separation. It does not measure accuracy, calibration, discrimination, clinical utility or patient outcome, and the lack of negative values is not evidence of safety or absence of harm.

Five historical score artifacts were assigned the design role. P35 or P44 was chosen for each candidate library by the smaller exact mean repaired Φ over those five artifacts, with equal artifact and endpoint weights. Exact equality was broken by lower structural partition entropy and then the lexicographically smaller restricted-growth signature. All 56 partitions of size $3/5$ were crossed with all 35 partitions of size $4/4$, producing an exhaustive matched domain of 1,960 libraries. The real library appears once, at frozen index 77, and the same selector is applied to every library.

Seed0 had the single evaluation role. Before the selection receipt was sealed, production read its bytes only to verify the frozen SHA-256 binding; its score values were not decoded. A separate one-shot authorization then permitted one value decode, fixed the primary raw $\Delta\Phi$, and forbade retry or substitution. Better/equal/worse counts and ranks are exhaustive orderings of the fixed 1,960-library domain, not a randomization null or a p value. The six-artifact leave-one-artifact-out calculation ran only after the primary file was fixed and cannot change it. No confidence interval, standard error, significance test, power calculation or population inference is attached to either result.

F.8 Independent dictionary selection

Each case uses disjoint fit, validation, and test partitions. The validation set locks both the predicted sign of Δ and the strongest competitor; the test-set argmax is then observed once. Only the locked sign is scored. The audit asks whether the target is the test argmax, so its verdict is independent of the locked competitor, and the locked competitor did differ from the observed test argmax in two of the five cases: both ILPD rows locked alkaline phosphatase, while total bilirubin was observed. Four of the five locked predictions were negative, that is, they predicted that the target is not the argmax; against dictionaries of 8–11 candidates such a prediction is low-information, and the agreement count should be read on those terms. Endpoints are the PTB-XL LVH label for Sokolow–Lyon, the PTB-XL LNGQT label for Bazett QTc (its 11 test positives are 11 LNGQT records among 2,198), the Sleep-EDF N1 sleep stage for the N1 candidate, and the ILPD liver-disease label for both ILPD rows. PTB-XL uses official folds 1–8/9/10 and patient clusters. Sleep-EDF uses subject groups and the official PhysioNet resource (Kemp et al., 2018). ILPD keeps exact duplicate rows in the same group and is cited to the UCI record (Ramana & Venkateswarlu, 2022). A locally vendored Framingham-labeled table was excluded from the manuscript because its official accession and license could not be verified. Paired cluster bootstrap intervals use 2,000 replicates. They quantify resampling uncertainty but are not null-hypothesis tests. The candidate list also included Bazett's QT correction and the De Ritis ratio with their historical definitions (Bazett, 1920; De Ritis et al., 1957). Their inclusion records a prespecified algebraic candidate; it does not validate either candidate for the local endpoint.

G ADDITIONAL EXPERIMENTS AND COMPLETE RESULTS

The sections below retain diagnostics outside the nine-page narrative and give the complete table for the one-bit result reported there. Each subsection states its own scope.

G.1 ONE-BIT COLLISION REFINEMENT

The five-artifact design objective selected P35 strictly. The two admissible candidates had objectives

$$J(\mathrm{P35}) = \frac{171728636330494379}{495395959010754560} = 0.3466492473,$$
$$J(\mathrm{P44}) = \frac{171771861890824619}{495395959010754560} = 0.3467365019.$$

On the single evaluation artifact, the base and P35 values were 0.37259132021290897 and 0.37259124924743514. Their exact difference was

$$\Delta\Phi_{\mathrm{seed0}} = \frac{858301}{12094627905536} = 7.096547382058237 \times 10^{-8}.$$

This is 1.90465×10^{-7} of the base Φ (about 0.0000190%), so the primary contrast is positive but near zero. Among all $1{,}960$ matched libraries, 624 had a larger value, 31 tied it and $1{,}305$ had a smaller value, giving rank interval 625–655 from best. Within the $1{,}274$ libraries whose selector also chose a $3/5$ partition, the corresponding counts were $368/31/875$ and the rank interval was 369–399. These are exact finite rankings, not inferential probabilities.

Table 9: Retrospective leave-one-artifact-out description for the real program library. Selection uses the other five artifacts; $\Delta\Phi$ is then evaluated on the held artifact. The six rows share the same records and development history and are not independent validations.

Held artifact	Selected block sizes	Held $\Delta\Phi$
seed13	$5/3$	0
seed42	$5/3$	0
seed101	$5/3$	0
seed202	$5/3$	$5.2034147930 \times 10^{-4}$
seed314	$4/4$	0
seed0	$5/3$	$7.0965473821 \times 10^{-8}$
Equal-weight mean	—	$8.6735407463 \times 10^{-5}$

Table 9 shows the six leave-one-artifact-out contrasts. The complete secondary ledger contains $6 \times 1{,}960 = 11{,}760$ library-fold rows. Its real-library mean ranks 557 of $1{,}960$, but four held-artifact contrasts are exactly zero and the seed202 row supplies 99.9864% of the positive sum. The secondary result is therefore heterogeneous and cannot rescue or strengthen the near-null primary. The experiment establishes only that the sealed program partition and matched comparison were executed and reconstructed exactly on the fixed artifacts. It does not establish meaningful, stable, clinical or general anti-memorization repair.

G.2 PATIENT-DISJOINT PTB-XL ILLUSTRATION

Across all seven requested fitted split caps the extended voltage vocabulary raised ordinary test accuracy by 0.0096, 0.0096, 0.0059, 0.0136, 0.0055, 0.0091, and 0.0005 at $b = 1, 2, 4, 8, 16, 32, 64$. At $b = 8$, accuracy was 0.9195 versus 0.9331 (lift 0.0136; patient-cluster bootstrap percentile interval $[0.0063, 0.0211]$). That cap carries the largest of the seven lifts, so its interval is a selected one and we read it descriptively, as promised in Section F; no multiplicity adjustment is made across caps. The $b = 1$ interval $[0.0000, 0.0192]$ touches zero, and at $b = 64$ the interval crossed zero ($[-0.0073, 0.0085]$). Balanced accuracy, reported because positives formed only 9.7% of the test records, was 0.634 versus 0.700 at $b = 8$ and 0.752 versus 0.763 at $b = 64$. The experiment therefore shows a descriptive held-fold difference under a selected fitted split cap, not a general clinical-utility claim.

One property of that headline is derivable from the numbers just given, so it is stated here rather than left to be found. Fold 10 holds 2,198 records and 214 positives, so the always-negative rule scores accuracy 0.9026. Most of both accuracies above is available without reading an ECG. The accuracy and balanced-accuracy pair already reported fixes the detection counts, and those run the other way: at $b = 8$ the base circuit detects 60 of the 214 positives and the extended circuit 88, the fewest of the seven caps in each vocabulary, against 113 and 118 at $b = 64$. The cap carrying the largest accuracy lift is the cap at which each circuit says least. At this prevalence accuracy rewards silence, and the lift above is measured on a scale where 0.9026 is free. Table 10 gives the counts at every cap. They use the classifier's argmax decision, which fires at leaf risk strictly above one half; Section F.5 refers at a declared threshold instead, and the two conventions can differ only on records whose leaf risk is exactly one half.

At the separately declared referral threshold $p_t = 0.10$, the pre-specified primary contrast was 17.996 net LVH-labeled records per $1{,}000$ screened, with paired patient-cluster percentile interval $[10.666, 26.052]$ over 2,000 replicates. The mandatory volume guard contradicted a rank-based reading of that contrast: the $b = 64$ circuit referred 417 records, versus 183 for $b = 1$. At each arm's nearest attainable volume below the target of 214, label yield was $134/213 = 0.6291$ versus $114/183 = 0.6230$ (difference 0.0062); the fixed-volume capture-difference bracket was $[-80, 21]$ because the two-leaf $b = 1$ circuit has no attainable referral volume between 183 and 2,198. The frozen verdict is therefore `MATERIAL_BUT_VOLUME_DRIVEN`, not rank superiority. Against the single named threshold selected by exhaustive training-set net benefit and evaluated once on fold 10, the $b = 64$ contrast was 9.049 per $1{,}000$, interval $[3.033, 15.423]$. These are consequences for the PTB-XL report label under the stated arithmetic, not evidence of echocardiographic disease, clinical utility or care benefit.

G.3 Selection stability is dictionary- and endpoint-specific

Validation predictions agreed with the held-group test argmax in all five source-verifiable cases. This count is a protocol audit, not five independent replications or a universal recovery rate. Three splits carry the five rows: the two PTB-XL rows share one 17417/2183/2198 split, and the two ILPD rows share a split and a byte-identical candidate dictionary, so both resolve to the same test argmax, total bilirubin. Table 11 exposes the sample and cluster counts. The test global AUC gap was positive for Sokolow–Lyon (0.0688, cluster-bootstrap percentile interval $[0.0406, 0.0789]$) and negative for the other four reported candidates. This global gap is not the pre-specified contrast. Table 12 gives the locked one; its interval contains zero for Bazett QTc, albumin/globulin, and De Ritis, while none of the reported global intervals does. The global statistic re-selects the strongest competitor inside every bootstrap replicate, so for a target that is not the test argmax an interval excluding zero is close to structurally automatic, and a percentile bootstrap of a maximum is not consistent near ties. The Bazett test split contained only 11 positives, and the Sleep-EDF test partition contained only two subject clusters, which admit three distinct cluster resamples, so its interval is degenerate rather than a stable population uncertainty estimate and is not reported. These limitations preclude prevalence or clinical-validity claims. The analysis establishes only that validation-defined predictions were evaluated on held groups under the stated dictionaries. For ILPD, grouping is defined only by exact duplicate rows because no patient identifier exists; patient independence is not identifiable.

G.4 Presented split count, atomic complexity, and restricted local residual

Proposition 13 separates the requested cap b, the distinct hyperplanes K_{split} presented by a fitted tree, the exact atomic complexity K_{atom} of its hard-decision region, and the emitted Boolean-DAG gate cost of a Boolean readout. The control below estimates a more limited quantity: how much of $g = K_{\mathrm{split}} - K_{\mathrm{atom}}$ remains after the frozen local operator $\mathcal{A}$. A positive residual is evidence only about that rewrite grammar; it is not evidence that the representation is globally irreducible.

Across 30 independent complete synthetic trajectories, the registered primary numerator was $\sum_s A_s = 52$ and denominator $\sum_s B_s = 5794$, hence $\widehat{\rho}_{\mathcal{A}} = 0.00897$. The whole-trajectory bootstrap 98.75% interval was $[0.00654, 0.01183]$ (half-width 0.00265), all 30 trajectory denominators were positive, and no bootstrap replicate was invalid. The precision rule therefore stopped sampling at the prespecified 30-trajectory look; the 50-, 75- and 100-trajectory looks were not run. On the parallel count scale, $47/770 = 0.0610$ positive-gap cell observations retained a positive residual,

Table 10: Fold-10 detections and referral volumes at each requested fitted split cap, in both PTB-XL vocabularies. Counts use the classifier's argmax decision, which fires at leaf risk strictly above one half, and are fixed by the accuracy and balanced-accuracy pair already reported. Fold 10 holds 2,198 records and 214 positives, so the always-negative rule scores accuracy 0.9026. Cap 8 detects the fewest positives in both vocabularies. The referral analysis of Section F.5 uses the weak convention at a declared threshold, which enlarges the extended $b = 64$ referral set to 417 and leaves the extended $b = 1$ set unchanged.

	Base		Extended	
b requested	Detected	Referred	Detected	Referred
1	71	118	114	183
2	71	118	114	183
4	100	168	114	183
8	60	83	88	109
16	84	109	110	149
32	92	124	111	142
64	113	161	118	170

Table 11: Held-group dictionary-selection audit. Δ_{test} compares the target with the strongest test-set competitor; it is not the pre-specified contrast, which is locked on validation and reported in Table 12. Reported intervals are paired cluster percentile intervals over 2,000 replicates. The maximizing competitor is re-selected inside each replicate, so for a target that is not the test argmax an interval excluding zero is close to automatic and a percentile bootstrap of a maximum is not consistent near ties; these intervals are not tests. Five rows, three splits: the two PTB-XL rows share one split, and the two ILPD rows share one exact-duplicate-group split and one candidate dictionary; ILPD patient independence is not identifiable. †Two subject clusters admit three distinct cluster resamples; the interval is degenerate and is not reported, though the JSON ledger records its nominal value.

Case	Fit	Val.	Test	Pos./clusters	Δ_{test} (percentile interval)
Sokolow	17417	2183	2198	214/1904	$0.0688\ [0.0406, 0.0789]$
Bazett QTc	17417	2183	2198	11/1904	$-0.0918\ [-0.3067, -0.0018]$
N1 candidate	16607	5592	5398	225/2	-0.1738 —†
Albumin/glob.	355	114	114	81/114	$-0.1268\ [-0.2611, -0.0193]$
De Ritis	355	114	114	81/114	$-0.2527\ [-0.4086, -0.1316]$

with 98.75% whole-trajectory-bootstrap interval $[0.0465, 0.0760]$. These continuous quantities and Tables 13–15, not a cutpoint label, are the main result. Applying the prespecified 0.10 cutpoint would be a descriptive secondary classification only.

The conditioning rule excludes nine constant positive-gap cells with total initial gap 11. Because the clinical diagnostics retain constants, we additionally ran an explicitly post-hoc all-cell sensitivity: $52/5805 = 0.00896$, with trajectory-bootstrap interval $[0.00653, 0.01181]$; the corresponding positive-cell fraction was $47/779 = 0.0603$, with descriptive post-hoc whole-trajectory-bootstrap interval $[0.0457, 0.0753]$. This denominator-only change does not materially change the qualitative low-residual reading, but it prevents unlike synthetic and clinical cell scopes from being silently conflated. A separate single-seed reconstruction remains an implementation fixture and is not an inferential result.

Every one of the 56 paired single-seed cells preserved the exact box decision, the stored float32 cell predictions, and 20,000 support-probe predictions. Pass B was exercised by the permanent ILPD $b = 4$ counterexample, where only the independently constructed right-child candidate is valid. The zero empirical probe discrepancy is a falsification result, not a proof; exact full-cell comparison supplies the proof under the declared box/tree contract.

The clinical controls are diagnostic and are not pooled with the theorem-covered synthetic arm. Two related PTB-XL trajectories used patient-disjoint train/test folds and retained residuals of $0/41$ and $0/51$ initial gap units. Sleep-EDF used a frozen eight-fit/two-held-out subject split and retained

Table 12: Pre-specified contrast. Δ_{loc} compares the target with the single competitor locked on validation; this is the quantity the protocol registered, whereas Table 11 reports the global gap against the strongest test-set competitor. The two coincide for Bazett QTc and the N1 candidate, the global gap is the larger in magnitude for both ILPD rows, and it is the smaller for Sokolow–Lyon. Three of the five locked intervals contain zero. †Not reported, for the reason given in Table 11.

Case	Δ_{loc} (percentile interval)
Sokolow	$0.0815\ [0.0574, 0.1070]$
Bazett QTc	$-0.0918\ [-0.2057, 0.0056]$
N1 candidate	-0.1738 —†
Albumin/glob.	$-0.0152\ [-0.1273, 0.1011]$
De Ritis	$-0.1410\ [-0.2910, 0.0039]$

Table 13: Complete synthetic mean-gap matrix over 30 independent cohort seeds, budgets 1–8. Each entry is $\bar{g} \to \bar{a}$ over all 30 fitted cells at that $\omega \times b$ location. The registered ratio-of-sums inference conditions on non-constant greedy cells with $g > 0$; this table deliberately also displays zeros and constants.

$\omega \backslash b$	1	2	4	8
0	$.00 \to .00$	$.40 \to .00$	$1.03 \to .00$	$1.70 \to .00$
.25	$.00 \to .00$	$.67 \to .00$	$1.03 \to .00$	$1.13 \to .00$
.50	$.00 \to .00$	$.77 \to .00$	$2.70 \to .00$	$1.90 \to .00$
.75	$.00 \to .00$	$.70 \to .00$	$2.57 \to .00$	$1.70 \to .00$
1	$.23 \to .00$	$.37 \to .00$	$.43 \to .00$	$1.57 \to .00$

$0/44$; eight leave-one-subject-out influence refits also all returned zero residual. ILPD, which has no patient identifier and no external cohort here, retained $0/8$ and remains diagnostic only. Across all 161 clinical cells, training and exact-refinement prediction-disagreement counts were zero; held-out disagreement was zero in the 154 PTB-XL and Sleep-EDF cells for which that check was applicable, while ILPD had no held-out cohort. Patient-disjoint evaluation rules out cross-fold reuse of the same patient in these cached PTB-XL analyses, but two related fold trajectories do not support a population prevalence claim; the ten-subject Sleep-EDF control is likewise sensitivity evidence rather than an external replication.

Finally, the audit compiler makes the atomic claim constructive. Given an externally bound box, named axis-threshold vocabulary, and exact binary-rational stored tree, it enumerates the presented refinement, extracts essential bichromatic facets, emits a Boolean readout, and returns a certificate whose lower and upper bounds agree at K_{atom}. A two-split witness compiles to $K_{\mathrm{atom}} = 1$ and one emitted basis gate; XOR needs two atoms and the emitted minimum-DNF DAG uses five gates. At synthetic $\omega = 0, b = 16$, the exact atomic count is 15, whereas the canonical DNF has 3929 gates—an upper bound on general Boolean-circuit cost, not a minimum. These records provide no certificate from which the emitted-gate coordinate can be inferred from atomic K; the joint Pareto frontier is not solved. The verifier checks exact cells, semantics, support and hash binding and rejects a falsified-K certificate; separate tests reject selected semantic tampering after rehashing. It shares the compiler's exact interval implementation, however, so this is a reference verifier rather than an independently implemented proof checker; scalability beyond the verified complete 5×5 low-budget grid is also left open.

Proposition 18 supplies a counterfactual distribution-free obstruction, not a lower bound on the bias of any PTB-XL, Sleep-EDF or synthetic law analyzed here. Evaluated at the same (d, n) as the ECG vocabulary—$N^{\circ}_{\mathcal{W},\Omega}(8) = 256$ and $n = 19{,}600$—the constructed fair-coin law has the exact expected-MAD obstruction $\beta = 0.0455$; the dense-end form returns 0.0456. At $N^{\circ}_{\mathcal{W},\Omega}(64) = 83{,}521$ and $n = 20{,}000$, the dense-end formula is out of scope because $d/n = 4.2 > \pi/2$, while the exact obstruction is $\beta = 0.446645$. These values say that a distribution-free expected-MAD guarantee below those levels cannot hold at the same (d, n); they do not estimate the realized frontier error.

Table 14: Complete synthetic mean-gap matrix, budgets 16–64; notation and cohort seeds are as in Table 13.

$\omega\backslash b$	16	32	64
0	$2.37 \to .00$	$3.80 \to .00$	$6.33 \to .00$
.25	$4.33 \to .00$	$8.83 \to .07$	$10.57 \to .17$
.50	$4.60 \to .00$	$14.83 \to .10$	$24.33 \to .30$
.75	$2.73 \to .00$	$14.47 \to .20$	$25.63 \to .50$
1	$7.20 \to .00$	$19.70 \to .03$	$24.87 \to .37$

Table 15: Companion complete positive-cell matrix. Each entry is $n(g > 0)/n(a > 0)$ out of 30 cohort seeds. It is the count companion to Tables 13–14, not a replacement for the ratio-of-sums mass estimand.

$\omega\backslash b$	1	2	4	8	16	32	64
0	0/0	12/0	24/0	24/0	26/0	30/0	30/0
.25	0/0	20/0	24/0	18/0	28/0	30/2	30/5
.50	0/0	23/0	30/0	30/0	30/0	30/3	30/9
.75	0/0	21/0	30/0	30/0	30/0	30/6	30/13
1	7/0	9/0	12/0	24/0	27/0	30/1	30/8

Under this particular obstruction, the nine-scalar $n = 19{,}600$ calculation excludes guarantees at $\varepsilon = 0.05$ beyond $k = 8$ and at $\varepsilon = 0.01$ beyond $k = 2$ on the tabulated ladder. Thus $k = 8$ is the largest tabulated budget not excluded by this construction at the looser target; escaping the same obstruction at $k = 16$ and 32 requires $554{,}716$ and $50{,}728{,}389$ records, respectively. This is a necessary counterfactual screen, not evidence that every cohort of those sizes can or cannot settle its own frontier. The PTB-XL lift of 0.0136 at requested fitted split cap $b = 8$ remains only a comparison of two fitted circuits on one held-out fold, as Section F claims. It is not promoted into a population-frontier statement. Together with Proposition 14, the calculation keeps computational intractability, distribution-free obstruction and realized held-out comparison as three distinct claims.

H Reproducibility ledger

The internal development archive contains additional KBS-stage drivers and ledgers that are not redistributed in the anonymous conference bundle. The submitted bundle contains the compact scientific artifacts used for the conference claims, a payload-complete manifest and a standalone verifier. The verifier checks hashes and recomputes the headline aggregates, including the complete E0 cell censuses, from the staged scientific exports. The following subsections enumerate the submitted evidence and its scope.

H.1 ICLR evidence extension: complete frozen results

The conference version adds five non-overwriting evidence envelopes. The tables below are generated from their scientific payloads at appendix-build time; the builder first recomputes each scientific SHA-256 and fails on a mismatch. A bracket of ∞ means that the decision does not factor through the declared vocabulary on the literal held-out records.

Matched-form placebo.

Requested splits	Base	Placebo	Named product	Placebo lift	Product lift
1	0.59555	0.59555	0.78570	+0.00000	+0.19015
8	0.63155	0.63155	0.82880	+0.00000	+0.19725
16	0.65330	0.65325	0.82880	-0.00005	+0.17550
64	0.74950	0.74915	0.82570	-0.00035	+0.07620

Complete deletion sensitivity.

Deleted name	0.5 bracket	Effect	Prevalence bracket	Effect
R_aVL	[9,19]	floor up	[10,23]	floor up
S_V3	[9,19]	floor up	[11,23]	floor up
S_V1	[9,20]	floor up	[10,22]	floor up
R_V5	[8,18]	unchanged	[10,21]	floor up
R_V6	[8,17]	cover changed	[10,21]	floor up
med_QRS	[8,18]	unchanged	[10,21]	floor up
med_axis	[8,18]	unchanged	[10,21]	floor up
med_ST_J60	[9,19]	floor up	[10,21]	floor up
med_T_amp	[9,19]	floor up	[10,22]	floor up

In the table, 'floor up' means a larger certified lower bound, 'cover changed' means that only the exhibited upper bound changed, and 'unchanged' means that neither endpoint changed. The full-vocabulary brackets are [8,18] at the 0.5 threshold and [9,21] at the prevalence threshold. These are deletion effects on the certificate, not claims that a name occurs in every minimum transcript.

All 66 base-vocabulary (A_0) operating-point brackets.

Decision	0.5	Prev.	Top 22	Top 110	Top 220	Top 440
1AVB	[8,17]	∞	[8,15]	∞	∞	∞
CLBBB	∞	∞	∞	∞	∞	∞
IRBBB	[10,24]	[10,24]	[8,15]	[10,24]	[11,26]	[12,27]
LVH	[8,18]	[9,21]	[5,8]	[8,16]	[9,21]	∞
LAFB	∞	∞	[6,10]	∞	∞	∞
AFIB	∞	∞	[8,16]	∞	∞	∞
LNGQT	[5,9]	[6,10]	[7,15]	[10,22]	∞	∞
IMI	[10,23]	[10,24]	[6,14]	[9,21]	[10,23]	[11,27]
ASMI	∞	∞	[7,15]	[9,20]	∞	∞
PVC	[10,22]	[10,22]	[7,16]	[10,22]	∞	∞
WPW	[4,4]	[6,10]	[8,15]	[10,21]	[11,25]	∞

The A_0 grid contains 30 infeasible and 36 solved cells; among the solved cells the certified floors range from 4 to 12 and exhibited covers from 4 to 27.

Exact-solver axis and resolution control.

Quantiles	Node budget	Realized nodes	Distinct cuts	Accuracy
16	15	14	11	0.80625
32	15	15	12	0.80955
64 (uncertified)	15	15	12	0.81020

At $\omega = 1$, node budgets 2 and 4 both use two distinct named cuts while accuracy rises from 0.6685 to 0.8364. The optimization axis is therefore node budget, not the traceability budget. The 64-quantile row is a sensitivity only and carries no optimality certificate.

H.2 ICLR EVIDENCE EXTENSION: ARCHIVED WORKSPACE LEDGER

The five envelopes are bound below to their scientific payloads and the workspace drivers that created them. These paths and hashes record provenance; they are not a promise that every historical workspace command is an entry point of the anonymous bundle.

- `companion/results/p5_v12b_placebo_extension.json` result SHA-256 `756002a6c8907518a18cb145d35de466c452c0d72ed518d85c99971b2ebaad82`; scientific SHA-256 `8924e56a23be22c9aafd000c45e0d4b0d767ef93679bdd2e`

3a08e701f85fdaf0; driver companion/code/run_p5_v12b_placebo_extension.py; driver SHA-256 85a993694810e9be44480deee359da4f52fb7050ea3992980ab73d4bc4559d4b.

- companion/results/p5_v13b_vocabulary_certificate_sensitivity.json result SHA-256 95908b10a115bc2dcc87d55cda4ee63c8e25aab038dd45bb0aa1db744a782658; scientific SHA-256 a3ddfd6c4940091ddd9a83dedbb3f38f02cd02ab61dee0915d83b48315ed8eef; driver companion/code/run_p5_v13b_vocabulary_certificate_sensitivity.py; driver SHA-256 86586d5101d6f8d144fb8b9cdcbb728a6818f9c70c1e192e913b888d8eaf009f.
- companion/results/p5_v14b_node_budget_baseline.json result SHA-256 48571fbbb24ceb3e3d20f03b533c869f274f598b296249fb92b4cc8161817b83; scientific SHA-256 20d0babb71c675010cd30f038adf846a65c41fa822d61883c7c6d222fe68cf4e; driver companion/code/run_p5_v14b_node_budget_baseline.py; driver SHA-256 c19f7ef384d75191cb31b86b5bdb617e2b07237f4979a60780f47bfa2986974b.
- companion/results/p5_v15b_multilabel_census.json result SHA-256 37bcb0ab5e540902e002d8b7521f0e73634c25ad173e52eadd0937b35358fb3f; scientific SHA-256 8718aacca6fa707f4268728b852a6a9234920252f5ab81e8a574275b1f33b5ec; driver companion/code/run_p5_v15b_multilabel_census.py; driver SHA-256 53d1838f3f396d87301eaf5dc189bcf7d914b8d44465a3f97ab4a6e5b8cf3931.
- companion/results/p5_v16b_bracket_exact_grid.json result SHA-256 f9f13a199be11d235dcbd30f014f6b8c32b3e565a2c9af81c528fe58eadc9d6f; scientific SHA-256 ba1a41bdda75872de36bf8964de22fd4dce667bd02f01bf1ce6d7fa32d919774; driver companion/code/run_p5_v16b_bracket_exact_grid.py; driver SHA-256 b4c5bbf06321ff0505641fb9134db9ee9f14b7f26ecd9018b4cdce17faa5ffcd.

The union of declared dependencies, inputs and migration source results is:

- companion/code/_result_hash.py SHA-256 40271eb47f683cc68f5c7d0a03641d137037513e01edc662f010194da6818825.
- companion/code/p5_evidence_contract.py SHA-256 35967e0095b3494b2b8164a289ff9f19b98ba5d45f94ba8e4ce559c04c6faf44.
- companion/code/run_p5_v12_placebo_extension.py SHA-256 3f8c44caf25c6b6f9a84e7460d28961cc6ff65df4942760b8af721988c1aa308.
- companion/code/run_p5_v14_optimal_tree_baseline.py SHA-256 73e6e5d1d119a29731c03e51587ad3d6d6a373c632f3955b23634ecdf3177025.
- companion/code/run_p5_v15_multilabel_census.py SHA-256 4f5a1aec884b54732ce76606ee84850a2bf813ae28bce0a78661c42585e9d077.
- companion/code/run_p5_v2_named_circuit.py SHA-256 601268c88d747670c1addf5ccebbd915e18abf7938134c56177edb91ba821a27.

- companion/code/run_p5_v4_blackbox_bracket.py SHA-256 5ac8be51905580dc11c5819a7770dee949cfdb5c6b87d588b5572d65ad38ef03.
- companion/results/_p1c_pr36p4_c1200_te.npz SHA-256 ed1c363327e581debd4e96934100f1b66f99472710400bf8bf36a272798f55ec.
- companion/results/_p1c_pr36p4_c1200_tr.npz SHA-256 a45261cf4db5758ca7efbe4084240d28e3d158cce89df8b42e297a1fc803d815.
- companion/results/_p5_v4_resnet1d_fold10_scores_seed0.npy SHA-256 a873d58808d5668c5bb45ecaa88190bd0e361ddd50b32a309e85a3f615dfd447.
- companion/results/p5_v12_placebo_extension.json SHA-256 6fa49ac8bb5e69b5c1bba9769051238173ce1820fb253b788aa4bb08f3c442d3.
- companion/results/p5_v14_optimal_tree_baseline.json SHA-256 9436c551b869eceb5ba64242a39009fbe983c81f9a7cd235d8a8c33e27d98eac.
- companion/results/p5_v15_multilabel_census.json SHA-256 15efeeaf13b92295bb4affb8967445ec7d26ee1a0ca78153d32daf7af7d30eb6.
- companion/code/run_p5_v4_decision_consequence.py SHA-256 b2fd627f8e772cb682b7fb19d3dc038b1e7e7228ec447846362ed4ba00ff6ab2.
- companion/results/p5_v4_decision_consequence.json file SHA-256 7890f5504dadc66e3f3dc974fe1de982ec93ce11dffd759d03c087f3bedca4cc; canonical scientific-content SHA-256 e644746debf6ee9397fdf81ef1140589cc2c95b29ca1dd9c987c4d41eb86fcce.
- companion/results/p5_v4_target_model.json SHA-256 3594bb3ccf84b4019dda468375d57137501ce54f6242560988e719333d590c8e.
- data/ptbxl/ptbxl_database.csv SHA-256 7600de9c1b27d181d850b3c6038a35d7c3ddb6bb33b702e3a20252a6859d216b.

The anonymous bundle retains the two exact referral files above under legacy/referral/ solely to preserve their original byte identities. Narrative verdict strings written before the final construct audit are deprecated and are not manuscript or package claims. The verifier instead evaluates the claim-bounded results/referral_scientific_v1.json, which preserves the frozen numerical fields and source bindings while omitting the deprecated narrative field. No production value was changed and the experiment was not rerun. The supported package-level audit entry point is python -B verify_v14r5.py. It checks the staged envelopes and recomputes the claims used in the paper. Full raw-data replays remain available through the archived workspace drivers and the cited public datasets.

H.3 One-shot collision-refinement ledger

The anonymous bundle places the frozen one-bit protocol and its independent verifier under code/onebit/, and immutable receipts and results under results/onebit/. Their file SHA-256 values are:

- protocol: 88E687DEB64F13B099BF370023278299F0CF9E6306A761147062BBB9CC3E650;
- runner: E51505919CFD812E6D49F3A85E9820150015B586F0D2BEBBF6A8416CEC86AD32;
- independent verifier: 7068C446A629FF75AFF84F4A1E928D1A783D18B7212372F4AF44BD4FA2C01078;
- selection receipt: 7E240C9DD0F8FD597D4629BEEDD3BB895C71BE9A6B5C9955FBC33FD4D8765124;

- access receipt: `E0BCF9FC125C5DE4422D48E2B7FF89D06079FC6AE472A1768FBCC6C79A43749A`;
- decode receipt: `B296761CB5D0C77B250C2D62E3E843EDEC58E4E523658A624B1449F036B934CA`;
- primary: `46F8F82F832B135237321100E5B7AF8BD8FAC593B2E0E54C8FD8BB3D3A32FA02`;
- secondary: `59F19529AFF5D5554E7BE4F1ED20DC20E054D0BB57B4D0CB5ECE793EC6F41080`;
- aggregate: `2C892B3552432CA4D0B070B025C996E2C71DA353645168DE82726B9AF982E625`;
- independent postrun report: `5AE5070DD9D6AB029EF690990BAF30F4861A5A603A9B25C0B6C95C833A3A33BA`.

The final report's canonical self-content SHA-256 is `6027EA57E1997BA38948F3AF2730A2881C4F618A57EF28263054367590DA08B4`. It independently reconstructs the primary array and all 11,760 secondary rows from the six fixed score artifacts, records one evaluation artifact read and fixes `post_access_retry_allowed=false`. The aggregate result file retains its point-in-time status "pending independent postrun verification"; the separately bound report is the later PASS record and is not rewritten into that historical receipt.

H.4 E0 FIXED-PROFILE SENSITIVITY AND FIXED-ARTIFACT PERSISTENCE/LOCALIZATION CENSUS

E0 was frozen before endpoint execution. E0-A applies zero-anchored numerical stress profiles and two hypothetical exclusions to the same score artifact; these profiles are engineering perturbations, not claims about instrument resolution. E0-B exhaustively uses all six available fixed score artifacts that met the frozen recipe and provenance contract. The artifacts are a retrospective fixed set, not a random sample of training seeds.

E0-A: resolution and population localization.

Vocabulary	Numerical profile	Raw N_{sc}	Record-excluded	Patient-closed
A0_base_9	Q0_native_float32	30/66	0/66	0/66
A0_base_9	Q1_zero_anchor_fine	30/66	0/66	0/66
A0_base_9	Q2_zero_anchor_mid	30/66	0/66	0/66
A0_base_9	Q3_zero_anchor_coarse	30/66	0/66	0/66
A2_extended_11	Q0_native_float32	30/66	0/66	0/66
A2_extended_11	Q1_zero_anchor_fine	30/66	0/66	0/66
A2_extended_11	Q2_zero_anchor_mid	30/66	0/66	0/66
A2_extended_11	Q3_zero_anchor_coarse	30/66	0/66	0/66

Across all fourteen blocks on the full record set, the count is $30/66$: both vocabularies, all four numerical profiles, and the explicit degeneracy indicator present or absent. Across all sixteen blocks on the record-excluded and patient-closed sets, the count is $0/66$. Adding two named criteria, rounding to the coarsest physical grid, or naming the degeneracy therefore removes no collision. The indicator is constant on the tied group; a name for the obstruction is not a coordinate that resolves it.

The raw cohort contains 2,198 records. Excluding the eight declared-vocabulary zero-vector records leaves 2,190; closing that exclusion over patient identity leaves 2,184. Every positive cell has exactly one conflicting representation class, comprising those same eight rows. Thus the obstruction is exact but concentrated rather than diffuse.

E0-B: all six frozen score artifacts.

Score artifact	Base vocabulary	Extended vocabulary
seed0_frozen	30/66	30/66
seed13_retrospective	29/66	29/66
seed42_retrospective	27/66	27/66
seed101_retrospective	21/66	21/66
seed202_retrospective	29/66	29/66
seed314_retrospective	25/66	25/66

The six counts are 30, 29, 27, 21, 29 and 25 (median 28; range 21–30). Eighteen operating-point cells are positive in every artifact and 36 in at least one. No confidence interval, hypothesis test or seed-population inference is attached. The five retrospective score files join patient and record identifiers exactly; the historical seed0 artifact remains tied to the frozen cache order, and checkpoint-to-score forward identity was not recomputed. AUCs are report-only diagnostics and were not used for selection or weighting.

Frozen scientific-result ledger.

- `companion/results/p5_iclr_e0a_discretization_degeneracy_v11/aggregate.json`
  `SHA-256 22544b86c4dde62209d7a288b9a15ef14a4b9d2b3d66ec99a1611d2b69b2e39c.`
- `companion/results/p5_iclr_e0a_discretization_degeneracy_v11/raw_cells.json`
  `SHA-256 69f6275c855fc56f97cc220694a20303a9e351d8b623589c67c06cc3d491b7bc.`
- `companion/results/p5_iclr_e0a_discretization_degeneracy_v11/verification.json`
  `SHA-256 1f9eb79d20cdabfba150a6fc0fa11445cf765d182e51954db0ddd52dcbd86faa.`
- `companion/results/p5_iclr_e0b_six_score_artefacts_v11/aggregate.json`
  `SHA-256 2b5d4b1b4abffa4749c81329727880958168f465b5c4e7359eca59b3c53139c9.`
- `companion/results/p5_iclr_e0b_six_score_artefacts_v11/raw_cells.json`
  `SHA-256 e2295351f4b0f7973e585bc41de50a9ba4b84dffa7326e1cfc3ddc7b998f93dc.`
- `companion/results/p5_iclr_e0b_six_score_artefacts_v11/verification.json`
  `SHA-256 59fc9e36b2ed72e116310b7e36bdd768a8fdf767e7e1f59ed370761ab2c40737.`

The independent pre-integration audit completed 54,321 assertions over all 1,980 E0-A and 660 E0-B cells, their order, dependencies, 18 mandatory checkpoint/score/run-result artifacts, denominators, and tie-aware AUC recomputations. The anonymous bundle contains the scientific-only E0 exports; `verify_v14r5.py` independently checks their hashes, cell counts, block aggregates and localization identity.

I FROZEN ONE-USE PTB-XL TOP-22 EXACT SEARCH

This Gate-II audit used the frozen fold-10 LVH score artifact and its raw top-22 decision. The raw threshold was 0.9705759286880493; the six-decimal display value was not used to reconstruct labels. The sample comprised $2{,}198$ records from $1{,}904$ patients, with 22 positive and $2{,}176$ negative decisions, hence $47{,}872$ bichromatic pairs. The base vocabulary A_0 contained nine named ECG scalars; A_2 added Sokolow–Lyon and Cornell. Their finite dictionaries contained $19{,}098$ and $23{,}478$ distinct sample cuts. Before the search, $m_n(A_0) \in [5, 8]$ and $m_n(A_2) \in [4, 8]$.

The one-use protocol fixed three branches before execution: a valid four-cut witness using an added direction was success; exhaustive closure without a witness raised the A_2 lower bound to five; and a timeout was `NOT_ASSESSED`. No retry or change of endpoint, threshold or vocabulary was allowed.

For a fixed four-direction multiset $(c_1, \dots, c_4)$, let R_{ic} be the rank of record i along direction c. A cut rank r_ℓ separates a bichromatic pair (i, j) exactly when

$$r_\ell \in \big[\min(R_{ic_\ell}, R_{jc_\ell}) + 1, \max(R_{ic_\ell}, R_{jc_\ell})\big].$$

The search solved the resulting finite clauses by unit propagation and exhaustive branching over nonempty interval intersections. Any transcript with fewer than four cuts can be padded by repeating a cut, so closure of the four-slot space also excludes smaller transcripts.

The eleven directions yield $\binom{14}{4} = 1{,}001$ four-direction multisets. The $\binom{12}{4} = 495$ base-only cases were already excluded by $m_n(A_0) \geq 5$. The exact search closed all 506 remaining multisets containing Sokolow–Lyon or Cornell, visiting $17{,}956$ nodes, and found no witness. A separate verifier reconstructed the raw decision, checked the frozen input and scientific-content hashes and matched the $506/506$ completion record. It did not rerun the combinatorial search. Because $A_0 \subset A_2$ and an eight-cut A_0 cover is available, the strongest fixed-sample conclusion is

$$5 \leq m_n(A_2) \leq m_n(A_0) \leq 8.$$

This tightens the A_2 lower bound but does not determine either optimum: a strict A_2 improvement may still occur within $[5, 8]$. The result is a sample Gate-II bound refinement, not a population or clinical repair claim.

The anonymous bundle stores the frozen protocol, one-use claim, searcher, result and separate verification together with their SHA-256 inventory. The scientific-content hash of the result is `59F2DF4932548B141846CC3BEE023068 4101AB95716893A3F01EA22F71005DDB`.

J Post-hoc PTB-XL engineering missingness stress test

This separate post-hoc exploratory diagnostic asks whether explicit measurement status resolves the fixed eight-record mixed-decision fiber in Section 4.3; it is not a retry or recovery. The eight patients, decision vector, nine coordinate families, measurement rules and dictionary were held fixed.

The typed representation did not assert numerical ECG measurement availability or clinical validity. All 72 coordinates returned `UNAVAILABLE_FIDUCIAL` with diagnostic `FEWER_THAN_THREE_VALID_QRS_FIDUCIALS`; none had a numerical value. The records therefore remained one typed cell with seven negative and one positive decision, giving $U(W_1) = \min\{7, 1\} = 1 = U(W_0)$ for $U(W) = \sum_{C \in \Pi_W} \min(n_C^+, n_C^-)$. All 36 status atoms were constant and no value-cut atom existed, so none separated the seven bichromatic pairs. Gate I failed and Gate II was not entered.

Table 16 records the complete fixed-sample outcome.

Table 16: Fixed-sample outcome of the post-hoc engineering missingness stress test.

Evidence item	Observed value
Cohort and decisions	8 records from 8 patients; 7 negative and 1 positive decision
Numerical coordinates	0 of 72 available; all 72 explicitly `UNAVAILABLE_FIDUCIAL`
Unavoidable conflicts	$U(W_0) = 1$ and $U(W_1) = 1$
Gate sequence	Gate I failed: 0 of 36 status atoms separated a bichromatic pair; Gate II not entered
Evidence tier	Post-hoc fixed-engine, fixed-dictionary diagnostic on the eight records; no clinical-variable, clinical-validity, B1–B3 or alternative-representation claim
Provenance	Public summary binds the aggregate quantities and verification hashes; machine-local paths are not redistributed

K VOCABULARY-RESOLUTION SENSITIVITY ON FROZEN PTB-XL ARTIFACTS

K.1 QUESTION, CONSTRUCTION AND EXACT REFERENCE

This analysis changes only the declared representation. The fold-10 records, nine named ECG coordinates, eleven frozen score vectors and six operating points per score vector remain fixed. For each finite q, the $q-1$ interior cut-points of every coordinate are computed on training folds 1–9, then applied unchanged to fold 10. The native arm retains the stored floating-point values. Gate I fails in a cell precisely when the frozen decision splits a fiber of the resulting representation. The 66 cells are an enumerated artifact, not sampled replicates.

The reference calculation conditions separately on each cell's observed number m of positives. Suppose the representation has non-singleton fibers of sizes $g_1, \dots, g_J$ and F remaining records. Under a positive set drawn uniformly from all m-subsets of the N records, the probability that no fiber is split is

$$p_0 = \frac{\sum_{S \subseteq \{1,\dots,J\}} \binom{F}{m - \sum_{j \in S} g_j}}{\binom{N}{m}} = \frac{[z^m](1+z)^F \prod_{j=1}^{J}(1+z^{g_j})}{\binom{N}{m}}, \tag{10}$$

where an out-of-range binomial coefficient is zero. Thus the reference failure probability is $1 - p_0$. A subset-sum dynamic program evaluates the numerator exactly; exhaustive enumeration on five small instances gives identical values. The expected number of failed cells is the sum of the 66 cell probabilities. The reported standard deviation and standardized difference additionally treat the cells as independent. This last coupling is a reference convention, not a sampling model for the overlapping decisions.

K.2 COMPLETE SWEEP

Table 17 reports the ten frozen resolution settings.

Table 17: Vocabulary-resolution sweep on the frozen fold. E_0 is the exact expected failure count under Equation 10; $z = (O - E_0)/\{\sum_i p_i(1-p_i)\}^{1/2}$ uses the cellwise-independent reference.

q	Fibers	Largest	Tied records	$O/66$	E_0	z
2	274	41	2102	66	65.9981	0.0435
3	347	8	840	66	65.5593	0.8137
4	84	8	182	64	61.9658	1.3155
6	9	8	24	47	44.5832	0.8430
8	2	8	10	30	30.8178	−0.2434
12	2	8	10	30	30.8178	−0.2434
16	2	8	10	30	30.8178	−0.2434
32	1	8	8	30	27.2960	0.7987
64	1	8	8	30	27.2960	0.7987
Native	1	8	8	30	27.2960	0.7987

On this frozen artifact, the number of cells failing Gate I falls from 66 under the two-bin construction to 30 from $q = 8$ onward, and the native arm reproduces the manuscript's $30/66$ count and its single eight-record fiber. The largest standardized excess under the stated reference is 1.3155 at $q = 4$. We therefore make no aggregate model-specific claim. In particular, tracking this reference neither proves equivalence to a null mechanism nor shows that the curve is a property of the vocabulary alone; both the observed and reference curves depend on the frozen decision counts.

A transparent post-hoc sensitivity using one record permutation jointly across all 66 decision columns and ten partitions (19,999-draw pilot; separate 99,999-draw one-sided max-T test) did not detect family-wise excess splitting ($T_{\max} = 1.069$, 95% randomization critical value 1.947, $p = 0.417$); this is failure to detect under synthetic record exchangeability, not equivalence or falsification of model-specific alignment, and record permutation does not preserve patient clustering.

If W' genuinely refines W, then every W-fiber is partitioned into W'-fibers and

$$\sum_{C' \subseteq C} \min(n^+_{C'}, n^-_{C'}) \leq \min(n^+_C, n^-_C), \tag{11}$$

so $U(W') \leq U(W)$. The quantile partitions used here are not nested across all tested values of q: their cut-points move when q changes. The observed sequence must therefore not be presented as a general refinement theorem.

K.3 Protocol deviations and claim boundary

A before-run internal plan fixed the grid and frozen artifacts, but two parts of its null specification do not match the retained calculation. First, it proposed multiplying Bernoulli unsplit probabilities within fibers, which is incorrect after conditioning on a fixed positive count; the retained code uses Equation 10. Second, the plan mentioned a 95% interval and Holm family, whereas the retained standardization requires an additional cell-independence coupling that the plan did not identify. An initial completed run also exposed a verdict branch weaker than the written criterion and was discarded before retention. For these reasons the null-calibrated component is reported as a descriptive reference analysis, not as confirmatory inference or external preregistration. Wilson intervals are likewise omitted because the 66 cells are an enumerated census.

The release evidence-binding file records the exact identities of the before-run plan, retained driver and result artifact. It also marks the raw artifact's automated null interpretation as superseded by the claim boundary above. Byte identity does not elevate the evidence tier.

L Post-hoc rate augmentation on the fixed PTB-XL fiber

L.1 Fixed object and evidence tier

This analysis is restricted to the eight records in the native all-zero fiber at the frozen LVH top-440 decision. In their fixed order, the decision vector is $(0, 0, 1, 0, 0, 0, 0, 0)$. No model was trained, loaded, scored or evaluated in this analysis. The atom and threshold were selected after inspecting these decisions and pilot rate measurements. The evidence tier is therefore an explicitly post-hoc fixed-eight exploration, regardless of the subsequent execution and verification controls.

In this section, $U(W)$ abbreviates $U_{\mathrm{mc}}(A_W)$ for the binary atom family A_W induced by vocabulary W. The median lead-II RR intervals in milliseconds are

$$(690, 852, 376, 611, 446, 598, 842, 552). \tag{12}$$

The typed atom $a(x) = \mathbf{1}\{\mathrm{median}(RR_{\mathrm{II}}) < 400\,\mathrm{ms}\}$ has truth vector $(0, 0, 1, 0, 0, 0, 0, 0)$ and hence reproduces the frozen decision on these records. The baseline partition has one mixed fiber, so Equation 1 gives $U(W_0) = 1$; after augmentation, $U(W_0 \cup \{a\}) = 0$.

There are seven bichromatic record pairs. The atom separates all seven, proving $m_8 \leq 1$. Because both decision classes occur, a zero-atom transcript is constant and cannot be faithful, so $m_8 \geq 1$. Thus $m_8 = 1$ exactly. By contrast, the coarse control $\mathbf{1}\{\mathrm{median}(RR_{\mathrm{II}}) < 600\,\mathrm{ms}\}$ has truth vector $(0, 0, 1, 0, 1, 1, 0, 1)$, leaves three bichromatic pairs uncovered and retains $U = 1$.

L.2 Same-record robustness and scientific interpretation

Splitting every record's RR sequence into temporal halves preserves the primary atom for all eight records. Deleting each RR interval in turn and recomputing the median also preserves it for all eight. Three PTB-XL+ outputs—12SL, Uni-G and ECGdeli—produce the same rate-side truth vector on the same records. These are robustness checks against local temporal and measurement-engine variation; none is an independent cohort or a held-out confirmation.

The positive record, ECG 7889, has a PSVT annotation and a median RR of $376\,\mathrm{ms}$ (approximately 160 beats per minute). Thus the atom has a rhythm interpretation but is not an LVH criterion. Exact agreement does not show that the frozen network used rate: it supplies only a one-atom factorization for these eight records, without establishing clinical validity or a unique missing measurement.

The release stages the amended result, frozen dependencies, measurement driver and two verification programs under `evidence/ptb8_rate_repair_v2/`. The package verifier recomputes the finite-sample certificate from the stored measurements; rerunning waveform extraction requires the public PTB-XL and PTB-XL+ files cited above.

M CUB CONCEPT-INTERFACE AUDIT

M.1 FINITE-LABEL GATE-I AND AGREEMENT CERTIFICATE

This section proves Proposition 3. Fix a presented sample and a finite atom family A, and write $a_A(x_i)$ for its joint atom readout. Any deterministic readout of A is constant on each cell $C \in \Pi_A$. If it assigns label a_C to that cell, its disagreement count there is $|C| - n_C^{a_C}$, which is minimized by a modal label. Summing the independent cellwise minima gives

$$\min_g \sum_i \mathbf{1}\{g(a_A(x_i)) \neq R_i\} = \sum_{C \in \Pi_A} \Big(|C| - \max_a n_C^a\Big) = U_{\mathrm{mc}}(A).$$

Equivalently, the maximum agreement over all deterministic functions of the joint atom readout is

$$\mathrm{Acc}_n^\star(A) = \frac{1}{n} \sum_{C \in \Pi_A} \max_a n_C^a = 1 - \frac{U_{\mathrm{mc}}(A)}{n}.$$

Choosing a modal label in every occupied cell attains the maximum. A tie can make the maximizing readout non-unique, but does not change its value. The minimum is zero exactly when every occupied cell is decision-homogeneous, which is precisely sample factorization through the interface.

Now let $A \subseteq B$. Each A-cell C is partitioned into B-cells $C_1, \dots, C_r$, and

$$\sum_{j=1}^r \max_a n_{C_j}^a \geq \max_a \sum_{j=1}^r n_{C_j}^a.$$

Subtracting from $|C|$ and summing over the old cells proves $U_{\mathrm{mc}}(B) \leq U_{\mathrm{mc}}(A)$ and $\mathrm{Acc}_n^\star(A) \leq \mathrm{Acc}_n^\star(B)$. Thus a positive value for a complete finite atom family rules out every subfamily as an exact readout, and the complete-family agreement is an upper bound for every subfamily. This is a finite-sample interface statement; the binary population boundary theory for K is unchanged.

M.2 PREDICTORS, ENDPOINT AND DECLARED INTERFACES

We audited three public joint-sigmoid concept-bottleneck checkpoints released with the CUB experiments of Koh et al. (2020). CUB-200-2011 contains 200 bird species and 312 annotated attributes (Wah et al., 2011); the released models use a fixed subset of 112 named attributes. The validation and test splits contain 1,198 and 5,794 images, respectively. For checkpoint s, the audited endpoint is the frozen soft-head decision

$$R_s(x) = \arg\max_a h_{s,a}(p_s(x)), \qquad p_s(x) \in [0,1]^{112},$$

not the ground-truth bird label. The checkpoints, splits and deterministic execution were fixed before the audit; the scientific object is therefore each frozen checkpoint–execution pair.

The binary interface is

$$\mathcal{I}_2 = \{\mathbf{1}[p_j \geq 4/8] : j = 1, \dots, 112\}.$$

The four-level interface uses numerators $\{2, 4, 6\}$ over eight, and the eight-level interface uses all numerators $\{1, \dots, 7\}$ for every concept. Hence

$$\mathcal{I}_2 \subset \mathcal{I}_4 \subset \mathcal{I}_8, \qquad |\mathcal{I}_8| = 112 \times 7 = 784.$$

The candidate augmentation library is the 672 atoms in $\mathcal{I}_8 \setminus \mathcal{I}_2$. Validation selection minimized candidate count and then applied a fixed lexicographic tie-break. It selected zero-based concept index 5 at $5/8$ for Seed 1 and index 2 at $7/8$ for Seed 2; the Seed-3 validation conflicts were infeasible in the frozen library. Both selected atoms reduced validation conflict to zero. Test data were not used for selection.

M.3 HELD-OUT RESULTS

Table 18 reports the complete test census. "Selected" means $\mathcal{I}_2 \cup \{a^\star\}$; for Seed 3 no atom was selected, so its row equals the binary base.

Table 18: Complete CUB test census. Occupied and mixed fibers are counted after forming the joint interface code; mixed mass is the number of images in mixed fibers.

Checkpoint	Interface	Occupied fibers	Mixed fibers	Mixed mass	U_{mc}	Conflicting pairs
Seed 1	$\mathcal{I}_2$	1916	20	165	34	404
	$\mathcal{I}_4$	3202	4	18	5	18
	$\mathcal{I}_8$	4346	2	6	2	4
	Selected	1925	20	164	34	403
Seed 2	$\mathcal{I}_2$	1885	14	129	49	820
	$\mathcal{I}_4$	3148	5	46	16	170
	$\mathcal{I}_8$	4341	3	22	8	38
	Selected	1928	17	126	41	507
Seed 3	$\mathcal{I}_2$	1885	15	114	35	419
	$\mathcal{I}_4$	3200	2	21	8	79
	$\mathcal{I}_8$	4381	1	3	1	2
	Selected	1885	15	114	35	419

For Seeds 1–3, the binary interfaces have exact best held-out agreements of 99.41%, 99.15% and 99.40%. The full $\mathcal{I}_8$ family raises these values to 99.97%, 99.86% and 99.98%, respectively, but its U_{mc} values of 2, 8 and 1 remain positive. By the refinement result above, no subset of the 784 frozen atoms can exactly reproduce the corresponding held-out soft-head decision vector, and no subset can exceed the checkpoint-specific $\mathcal{I}_8$ agreement ceiling. The validation-selected atoms also fail to transport exactly: Seed 1 remains at 34, Seed 2 falls from 49 to 41, and Seed 3 has no candidate. Gate II is therefore not entered.

A diagnostic frozen before execution reconstructed all six $\mathcal{I}_8$ seed–fibers: 31 record occurrences spanning 29 official image identifiers and all 44 opposite-decision pair occurrences. No pair shared a source-file SHA-256, decoded RGB array, 299×299 model crop or normalized model input, and none met the fixed operational near-duplicate screen (raw- or crop-image pHash Hamming distance at most 6, or crop SSIM at least 0.98). The observed pHash minima were 20 and 22, and the maximum SSIM was 0.549. A separate verifier reproduced the complete census in the frozen runtime. These checks exclude exact duplicates and positives under the stated screen; CUB provides neither bird-individual nor capture-session identifiers.

As a construction control, we recomputed the endpoint after feeding the 0.5-thresholded concept vector to the same head. Its decision has $U_{\mathrm{mc}} = 0$ for all three checkpoints, as any deterministic function of that vector must. A separate verifier reconstructed the endpoints and partitions from the frozen extracted tensors and returned PASS for each checkpoint. The result identifies information discarded by the finite declared semantic interface. It does not rule out the continuous concept state, finer thresholds, derived concepts, another checkpoint or population factorization.

M.4 EXECUTION-FROZEN JOINT1 ORDINAL REPAIR

Scientific object and sequence. The second CUB study audited three released Joint1 checkpoints. For checkpoint s, raw concept logits $z_s(x) \in \mathbb{R}^{112}$ enter the released 200-class head,

$$R_s(x) = \arg\max_a g_{s,a}(z_s(x)).$$

The sigmoid values $p_s(x) = \mathrm{sigmoid}(z_s(x))$ define the displayed vocabulary only; they are not substituted for $z_s(x)$ in the class head. The official $5{,}794$-image test split was ranked by a fixed SHA-256 function of the official image identifier and relative path, then divided into $2{,}897$ discovery and $2{,}897$ confirmation records. The split was constructed before any Joint1 checkpoint was downloaded or evaluated.

This was a sequential, V1-informed study. The preceding one-use exact q16 search was frozen before execution on a different released independent-bottleneck recipe. Across its three fixed checkpoints, $U_2 \to U_{16}$ changed from $66 \to 1$, $65 \to 8$ and $59 \to 1$; all three stopped without a finite transcript. For the primary checkpoint, $3{,}212/5{,}794 = 55.44\%$ of records remained in non-singleton q16 fibers. This stage is a sequential negative comparator from a different recipe, not a matched control. Before opening any Joint1 endpoint output, the second stage fixed Joint1 Seed 1 as the sole primary endpoint, Seeds 2 and 3 as non-rescuing replications, the q32 family, the outcome-blind half split and every terminal disposition. Every terminal outcome stopped the CUB repair search. The second stage is therefore an execution-frozen evaluation of a previously unaudited endpoint within the same public test set.

For $q \in \{2, 4, 8, 16, 32\}$, the interface is

$$\mathcal{I}_q = \{\mathbf{1}[p_{s,j} \geq r/q] : j = 1, \ldots, 112,\ r = 1, \ldots, q-1\}.$$

Threshold equality evaluates as true. Thus $\mathcal{I}_2$ has 112 predicates and $\mathcal{I}_{32}$ has $112(32-1) = 3{,}472$, with $\mathcal{I}_2 \subset \mathcal{I}_4 \subset \mathcal{I}_8 \subset \mathcal{I}_{16} \subset \mathcal{I}_{32}$. No task logit, class margin, image identifier, hash, embedding or data-dependent threshold enters an interface.

Fixed-policy result. Table 19 gives every frozen grid point. The primary success rule required $U_{\mathrm{mc}}(\mathcal{I}_2) > 0$, $U_{\mathrm{mc}}(\mathcal{I}_{32}) = 0$, at least 5% of all confirmation records in non-singleton q32 fibers, exact class-decision replay and an independently reconstructed zero-error transcript.

Table 19: Complete sequential Joint1 ordinal-interface census. All rows contain 2,897 records. The q32 mass is the number and percentage of records in non-singleton q32 fibers.

Phase	Seed	U_2	U_4	U_8	U_{16}	U_{32}	q32 mass
Discovery	1	31	7	1	0	0	1,342 (46.32%)
Confirmation	1	34	7	3	2	0	1,377 (47.53%)
Confirmation	2	25	5	1	1	0	1,319 (45.53%)
Confirmation	3	26	1	1	0	0	1,313 (45.32%)

Although q16 appeared sufficient on primary discovery, it left two unavoidable disagreements on confirmation. The held-out half therefore prevented an optimistic sufficiency claim based on discovery alone; q32 was fixed before any Joint1 output was opened.

Seed 1 passed the frozen primary rule. Its binary-interface agreement ceiling is $2{,}863/2{,}897 = 98.83\%$, whereas the q32 ceiling is 100%, a 1.17-percentage-point increase. Because every occupied q32 fiber is decision-homogeneous, a finite confirmation-sample transcript exists. With $K_{n,\mathrm{mc}}$ denoting the number of threshold atoms in an exhibited sample-level multiclass transcript, compiling the modal decision of each occupied cell and replaying it without error gives

$$K_{n,\mathrm{mc}} \leq 3{,}472.$$

This is a valid upper bound for the presented confirmation sample, not the minimum budget and not a population value of K. The transcript is compiled on the confirmation partition as an existence certificate; it is not a discovery transcript transported to confirmation. Seeds 2 and 3 reproduce the fixed-policy transition but do not replace the primary endpoint.

Compact local secondary endpoint. Discovery also searched exactly for at most four q32 additions inside any binary fiber of size at least 20 and with $U_{\mathrm{mc}} \geq 2$. It selected two atoms in one 26-record fiber:

$$\texttt{has_bill_shape::hooked_seabird} \geq 1/32,$$
$$\texttt{has_wing_shape::rounded-wings} \geq 1/32.$$

The four discovery signatures were pure, with supports 15, 3, 6 and 2. On confirmation, the same base fiber contained 30 records but had $U_{\mathrm{mc}} = 1$, below the frozen opportunity gate of two. One signature had support one, and the frozen discovery transcript made two errors. The terminal status is therefore `NO_CONFIRMATION_LOCAL_REPAIR_OPPORTUNITY`. The positive primary result is a fixed full-family factorization; it is not a compact transported repair.

Independent reconstruction and numerical scope. A separate verifier rebuilt the sigmoid scores, nested partitions, all U_{mc} values, the local exact search, terminal dispositions and transcript replays for discovery and all three confirmation checkpoints. A first verifier call stopped because CPU and recorded CUDA logits differed by at most 1.91×10^{-5}, with no argmax change. The retained amendment requires both tolerant CPU replay with identical decisions and exact deterministic CUDA replay; it changes no endpoint, vocabulary or success rule. The final reconstruction passed. The anonymous package retains the protocol, amendments, split projection, result fragments, transcripts and verification programs under `evidence/joint1_q32_repair_v1/`, `results/joint1_q32_repair_v1/` and `code/joint1_q32_repair_v1/`.

M.5 A FITTED READOUT AND THE SAMPLE AGREEMENT CEILING

The interface certificate concerns the best possible readout on a presented sample. To distinguish this quantity from the performance of a learned readout, we conducted a secondary analysis of the previously examined Joint1 Seed-1 data in Section M.4. The original discovery and confirmation halves each contain 2,897 records. Here these names identify the existing split; the analysis adds no fresh confirmation data.

Fixed comparison. For each $q \in \{2, 16, 32\}$, one multi-output ridge regression was fitted to the frozen model's 200 class logits on the discovery records, with penalty $\alpha = 1$, an unpenalized intercept and no tuning. Its inputs were exactly the $112(q-1)$ binary threshold atoms of $\mathcal{I}_q$; prediction used the lowest-index argmax of the fitted logits. Fits used float64 arithmetic and a Cholesky solver. All three sets of predictions were saved before evaluation against the confirmation decisions. Ground-truth bird labels were not targets. The execution-frozen primary comparison was q16; q2 and q32 supplied fixed context. The descriptive criterion, agreement of at least 95% with positive U_{mc}, was not met on the primary confirmation split.

Table 20: Fixed-penalty ridge agreement and the exact sample ceiling on Joint1 Seed 1. Each row has $n = 2{,}897$ records. E counts disagreements with the frozen model; $U = U_{\mathrm{mc}}(\mathcal{I}_q)$ is their minimum over all deterministic interface readouts. Ridge agreement is $100(1 - E/n)$ and the ceiling is $100(1 - U/n)$.

Split	q	E	U	$E - U$	Ridge agreement (%)	Sample ceiling (%)
Discovery	2	227	31	196	92.16	98.93
	16	50	0	50	98.27	100.00
	32	26	0	26	99.10	100.00
Confirmation	2	198	34	164	93.17	98.83
	16	208	2	206	92.82	99.93
	32	252	0	252	91.30	100.00

Table 20 separates two questions. At q32, the confirmation interface admits exact reproduction, while the discovery-fitted readout agrees on $2{,}645/2{,}897 = 91.30\%$ of records. The interface repair certifies existence on the confirmation sample; it does not guarantee that a discovery-fitted readout attains the ceiling. The gap $E - U$ can reflect the readout class, fitting objective, regularization and estimation across samples, rather than isolating optimization error.

Joint codes already present in discovery account for 2,352, 1,625 and 1,387 confirmation records at q2, q16 and q32. The corresponding error counts are 39, 0 and 0; unseen codes account for the remaining 159, 208 and 252 errors. Thus every q16 and q32 error occurs at a code unseen in discovery. The higher sample ceiling coexists here with lower ridge agreement; these counts locate the errors but do not identify the cause of the decline.

The anonymous package retains the frozen configuration, driver, saved outputs and verification receipts under `code/surrogate_ceiling_v1/`, `evidence/surrogate_ceiling_v1/` and `results/surrogate_ceiling_v1/`. All three comparisons are retained. They form one fixed-checkpoint case study, not independent replications.

N ECG HARD CONCEPT-BOTTLENECK AUDIT

This appendix records the audit reported in Section 4.1. The ledger was frozen before any concept activation was computed, and the driver writes a content hash over its own payload so that a replay is byte-comparable.

N.1 FROZEN INPUTS

The audited object is a trained twelve-lead concept-bottleneck classifier, stored as a checkpoint and loaded once. Its concept layer holds 22 binary units: 19 learned clinical rule detectors, each carrying a name and a cut-point, concatenated with three frozen morphology anchors ϕ_{WPW}, ϕ_{RBBB} and ϕ_{LBBB} that carry no clinical name. Concept scores are hard-thresholded at 0.5 through a straight-through estimator, and the resulting binary vector enters a sign-constrained masked linear head. Nothing is trained, fine-tuned or selected in this appendix.

The evaluation set is the $2{,}198$ patient-disjoint fold-10 records of the same PTB-XL cache tag used throughout the paper, with no exclusions. Eleven diseases are audited at the six-threshold ladder $\{\tau_{1/2}, \tau_{\mathrm{prev}}, \tau_{22}, \tau_{110}, \tau_{220}, \tau_{440}\}$ derived per disease from the model's own logits, giving $11 \times 6 = 66$ cells per vocabulary.

N.2 THE TWO VOCABULARIES AND THE ENDPOINT

The audit runs twice over the same frozen decisions. The *full* vocabulary is all 22 binary concepts; the *named* vocabulary is the 19 that carry clinical names. In both runs the endpoint is Proposition 2: the decision fails to factor through a vocabulary exactly when it splits a tied group, that is, a maximal set of records sharing an identical binary concept vector.

The full vocabulary is the construction anchor rather than a result. The decision is a deterministic function of the 22-bit vector, so it cannot split a full-vocabulary tied group, and a failure there would indicate a defect in the harness rather than a property of the model.

N.3 PRESPECIFIED CHECKS

Four checks were fixed before execution, and each was mutation-tested against a clean state, a broken state and a degenerate state to confirm it can actually fail.

- C1 The full 22-concept vocabulary passes at all 66 cells. A failure invalidates the run.
- C2 The effective head mask read from the loaded checkpoint matches the declared concept graph. The declared edges are $\phi_{\mathrm{WPW}} \to \mathrm{WPW}$, $\phi_{\mathrm{RBBB}} \to \mathrm{IRBBB}$ and $\phi_{\mathrm{LBBB}} \to \mathrm{CLBBB}$, so for the other eight diseases the logit cannot depend on any anchor bit.
- C3 Latent anchor activations are bitwise reproducible from the stored anchor file.
- C4 The comparison is not vacuous: the named vocabulary must induce at least one tied group, and at least one cell must differ between the two vocabularies.

Check C2 makes the finding falsifiable. Because the mask is read from the checkpoint rather than from the architecture description, a named-vocabulary failure at any of the eight diseases with no anchor edge would contradict the mask and was prespecified as a setup defect rather than a result.

N.4 RESULTS

Table 21 gives the full per-disease breakdown. The full vocabulary passes $66/66$ as required by C1. The named vocabulary passes $53/66$. All 13 failures fall on the three endpoints with an anchor edge, and the eight endpoints without one pass at all 48 of their cells, so the registered verdict is `FAILURE_CONFINED_TO_LATENT_SUPPORTED`.

The mechanism is visible in the tied-group census. Removing three of 22 concepts merges the record set from 167 tied groups into 116, and raises the largest group from 255 to 547 records. The splits that Proposition 2 detects occur inside those merged groups.

Table 21: Gate-I failures by endpoint over the named 19-concept vocabulary, out of six operating points each. Every failure falls on an endpoint whose declared head mask gives an unnamed anchor an edge. The eight endpoints with no anchor edge are 1AVB, LVH, LAFB, AFIB, LNGQT, IMI, ASMI and PVC.

Endpoint	Anchor edge	Failures (named)	Failures (full)
WPW	ϕ_{WPW}	$6/6$	$0/6$
CLBBB	ϕ_{LBBB}	$4/6$	$0/6$
IRBBB	ϕ_{RBBB}	$3/6$	$0/6$
Eight remaining endpoints	none	$0/48$	$0/48$
Total		$13/66$	$0/66$

WPW is the extreme case, failing at every threshold. That is consistent with why the anchor exists: the delta wave that defines pre-excitation was among the morphological patterns that resisted expression as a threshold on the named measurement set during development, which is what motivated adding an unnamed unit in the first place. The audit prices that decision. For WPW, CLBBB and IRBBB the named concept subset is not an exact transcript of the decision, so an explanation quoting only the named concepts is incomplete at those operating points.

N.5 Scope

The audit examines one architecture on one cohort. It establishes that the named 19-concept subset does not determine every presented decision and localizes each failure to an endpoint whose head uses an omitted anchor. It does not establish that the anchors are necessary, measure the accuracy effect of removing them or support a clinical interpretation. The 66 cells are an enumerated census of a deterministic function of frozen artifacts, so no sampling interval is attached.

O A language-model display audit

O.1 Decision, representation and displayed information

This analysis applies the finite-label certificate to a deliberately restricted display of a frozen language model. It uses GPT-2 and the released GPT2-Small-SAEs-Reformatted suite (Radford et al., 2019; Bloom, 2024). The SAE input is the residual stream before blocks 3, 5, 8 and 11; the block indices are zero-based and each dictionary has 24,576 features. Layer 8 and $k = 3$ define the primary descriptive comparison. Model and SAE weights remain fixed.

We used the `wikitext-103-raw-v1` configuration. We joined its nonempty test rows with newlines, tokenized the text once and retained the first 2,000 complete consecutive blocks of 128 tokens (Merity et al., 2016). Positions 8–127, inclusive, contribute 120 observations per block. Even-indexed blocks form the discovery half; odd-indexed blocks form the 120,000-position analysis half, called confirmation in the stored files. These interleaved blocks reuse a previously examined corpus and are not an independent-document confirmation sample. Each target label is the frozen model's next-token argmax at that position, rather than the observed next token.

Let $a_L(x)$ be the nonnegative SAE activation vector at one position and j_1, j_2, j_3 its three largest positive coordinates. Ties are resolved by increasing feature index. We retain positions with at least three strictly positive activations; all 120,000 analysis positions qualify at each of the four layers. The two displays are

$$S_L(x) = \{j_1, j_2, j_3\}, \qquad Q_L(x) = (j_1, j_2, j_3),$$

where the tuple follows decreasing activation. Different positions may select different IDs. The comparison preserves the three selected IDs at each position and adds their relative order. It does not preserve the information budget: conditional on the set, at most $3! = 6$ orders are possible. Neither display includes activation magnitudes, other positions or later-layer states.

For each finite code, Equation 1 counts unavoidable disagreement by taking the majority decision in each cell. The cellwise argument in Proposition 3 applies directly to these categorical codes.

Forgetting tuple order maps Q_L deterministically to S_L, so its partition is a refinement on the same support. Hence $U_{\mathrm{mc}}(Q_L) \le U_{\mathrm{mc}}(S_L)$. In contrast, unordered top-k sets at different k need not form a refinement chain. The displayed feature count k is not the traceability budget K.

The frozen SAE computes $\mathrm{ReLU}((h - b_{\mathrm{dec}})W_{\mathrm{enc}} + b_{\mathrm{enc}})$, where $h = h_{\mathrm{HF}} - \overline{h}_{\mathrm{HF}}\mathbf{1}$ subtracts the position's mean across 768 residual coordinates. Direct block-input comparisons and numerical centered-writer checks, together with representative independent forward replay, verified the coordinates used in this extraction. This check does not recover historical SAE training flags. Exact tie ordering, positive-activation eligibility and code-to-decision alignment were checked before interpreting the stored arrays. Model, tokenizer, corpus and SAE revisions and their source hashes accompany the artifact.

O.2 Partial recovery across the fixed layers

Table 22 reports all four fixed layers. Ranking reduces unavoidable disagreement at each layer, but every ranked display retains mixed cells. Layer 8 gives the main-text result: U_{mc} falls from 17,610 to 12,791, raising the exact sample optimum from 85.3250% to 89.3408%. This is partial recovery of the restricted display. The layers are repeated analyses of the same frozen model and corpus, not independent model replications.

Table 22: Fixed-k display refinement on the same 120,000 positions per layer, $k = 3$. M counts occupied cells; the last column counts records in non-singleton ranked cells.

Layer	U_{set}	U_{rank}	M_{set}	M_{rank}	Ranked non-singletons
3	31,946	26,564	67,809	76,592	56,936
5	24,105	19,442	77,603	85,311	46,316
8	17,610	12,791	84,653	93,532	37,107
11	9,810	5,329	95,490	105,056	22,747

O.3 Separating singleton gains from the occupancy reference

An empirical optimum can increase when a partition becomes finer even if its new distinctions have no special relationship to the labels. We therefore report two different diagnostics on the same saved arrays. Both were specified before computing their outputs, after the original display outcomes were known. They are secondary descriptive analyses.

First, we isolate every singleton of the final ranked partition and only then refine the remaining records. For an unordered parent cell C, let c_C be its largest label count, s_C the number of records that become ranked singletons, and r_C the largest label count after removing those records, with $r_C = 0$ for an empty remainder. Let q_C sum the majority counts in its non-singleton ranked children. The gain has the exact accounting decomposition

$$G = \sum_C (s_C + q_C - c_C) = \underbrace{\sum_C (s_C + r_C - c_C)}_{G_{\text{singleton first}}} + \underbrace{\sum_C (q_C - r_C)}_{G_{\text{remaining}}}.$$

Both terms are nonnegative: isolating records and then refining the remainder cannot reduce its optimal correct count. This is an explicitly ordered accounting decomposition, not a causal attribution. At layer 8, $4{,}819 = 4{,}107 + 712$. On the identical remaining 37,107 records, the optimal correct counts are 23,604 and 24,316, giving the second row of Table 4.

Second, an occupancy reference keeps every unordered parent, its decision-label multiset, and its actual ranked child cells fixed. It uniformly permutes label assignments within each parent, preserving all child-cell sizes. We use 499 assignments per layer, NumPy PCG64 seed $20260909 + L$, sample standard deviation with one degree-of-freedom correction, and linear interpolation for empirical quantiles. Table 23 reports the complete comparison. The intervals describe this artificial finite assignment reference, not population uncertainty.

The layer-8 reference mean has Monte Carlo standard error 0.961 disagreements. The observed gain exceeds that mean by 170.33, or 0.1419 percentage points of the full sample. All four observed gains

Table 23: Observed refinement gains and the occupancy-preserving reference. Every entry counts fewer unavoidable disagreements. The reference range is its empirical central 95% range over 499 assignments; it is not a confidence interval.

Layer	G	$G_{\text{singleton first}}$	$G_{\text{remaining}}$	Reference mean	Reference range
3	5,382	4,032	1,350	5,168.11	[5,113.35, 5,220.00]
5	4,663	3,867	796	4,433.33	[4,388.00, 4,480.00]
8	4,819	4,107	712	4,648.67	[4,607.00, 4,689.00]
11	4,481	4,067	414	4,315.11	[4,270.00, 4,360.55]

exceed their reference's empirical 97.5th percentile. The excess over the reference and the gain on the common non-singleton subset answer different questions and are not additive components.

O.4 Artifact scope and interpretation

The supplementary artifact includes compact aligned codes and frozen model decisions for the four fixed layers, exact count routines, all per-assignment gains and a seeded saved-data replay command. These inputs reproduce the reported finite statistics without loading a language model. Original extraction code and release identifiers describe the separate forward-reconstruction step. Two executors used the same independently written verifier for complete raw-array recomputation; each matched all 1,996 assignment gains and all four layers' counts. This is two executions of one independent implementation, not two independent datasets.

The display audit complements SAE reconstruction, sparse probing and intervention evaluations (Karvonen et al., 2025; Gao et al., 2025). It asks what an explicitly chosen display allows every deterministic readout to reproduce. The custom index-only display is not a model of all SAE explanation practice, and its information loss need not originate in the SAE itself. The experiment establishes a sample-level display gap and partial recovery by restoring order. It leaves learned-readout transfer, feature semantics and causal mechanism identification to separate tests.

The initial random-dictionary interpolation and direction-only interpretation are not used here. The released analysis uses actual discrete partitions, strict positive-coordinate eligibility, deterministic tie handling and the coordinate-verified cache. Its reused-data status is retained throughout; internal execution freezes are not external preregistrations.