WHEN DO WORLD MODELS RECOVER LATENT FACTORS?

Anonymous authors
Paper under double-blind review

ABSTRACT

World models can mix latent variables while task predictors compensate, preserving predictive accuracy. The existing low-degree guarantee requires direct coordinate readouts to remain first-order. We ask when minimizing the average Fourier degree of predictors sharing a representation recovers coordinates without this separate constraint. For cumulative task families defined by maximum interaction order, we completely characterize coordinate identifiability among reversible encodings of d binary variables under generic continuous and uniform finite-output task laws. Population minimizers recover coordinates, up to permutations and independent sign flips, if and only if the selected task spaces' joint projection kernels distinguish pairs differing in one variable from all other distinct pairs. Cumulative cubic tasks suffice for $d \geq 5$; every failing selection admits an explicit reversible mixed encoding. When each variable takes $q \geq 3$ values, every proper, nonconstant cumulative family identifies coordinates. We also prove approximate recovery guarantees and give a reconstruction algorithm running in time polynomial in the state count $n = 2^d$. Readouts supervised on all states show that fixed representations from frozen Transformers support coordinate recovery through task kernels. After graph correction, cumulative cubic readouts achieve 100% coordinate accuracy for $d = 6$–11, while the pipeline rejects all tested non-identifying controls.

1 INTRODUCTION

World models learn from observations an internal state that can be reused across prediction, planning, or control tasks (Ha & Schmidhuber, 2018; Friston et al., 2021). We consider worlds composed of binary facts, such as whether a door is open or a light is on, with each fact corresponding to a *true coordinate*. Yet a model can mix these facts when encoding the world and let task predictors compensate, leaving task answers unchanged. For example, it can record the door's state and whether the door and light states agree, then combine these two entries to infer the light's state. Task accuracy alone therefore cannot guarantee coordinate recovery, but different encodings can make prediction tasks differ in complexity. Recording the light's state directly requires reading only one coordinate, whereas the mixed code above requires an interaction between two coordinates. This motivates the low-order preference, which favors representations that solve shared tasks through lower-order coordinate interactions. We ask whether complexity constraints on higher cumulative task families can still rule out mixed codes and recover true coordinates when tasks that directly read latent variables are no longer required to remain first-order.

Zhang et al. (2025) formalized this preference as a population objective minimizing the average Fourier degree of shared-task predictors, organizing tasks by their maximum allowed interaction order. For example, the up-to-three family allows first-, second-, and third-order interactions. They proved that an optimal representation recovers the true coordinates when all interaction orders are sampled and, in particular, the first-order family receives positive weight. But that family contains tasks that directly read individual latent variables, so requiring these tasks to remain first-order already restricts coordinate mixing. The result therefore leaves open whether complexity constraints on higher-order families alone can determine the true coordinates without separately constraining the first-order family. Removing this separate constraint still leaves lower-order components in the higher cumulative families, but no longer requires those components to separately retain their origi-

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.

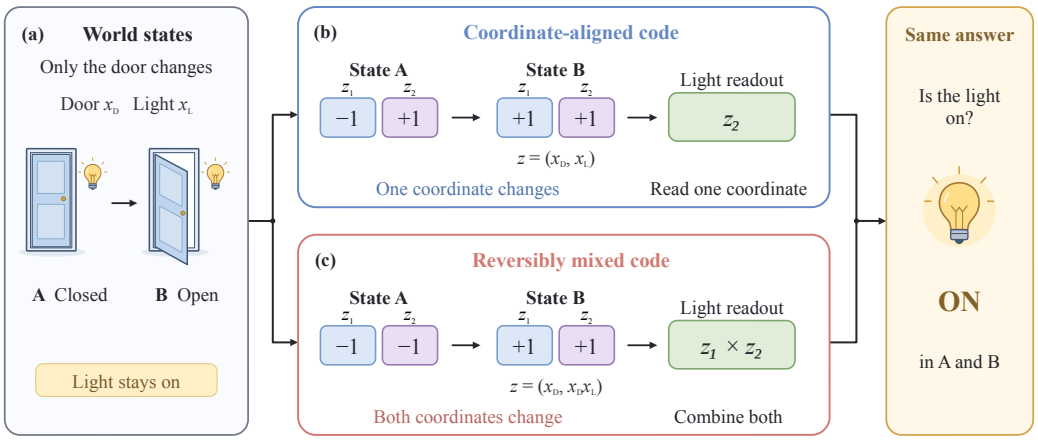


Figure 1: Task predictors can absorb an unknown latent-state relabeling, hiding different coordinates behind identical outputs. We classify when higher cumulative families identify coordinates without a separate first-order constraint.

nal orders. Prior work gives only one failure example in three dimensions, leaving higher dimensions and joint constraints from multiple higher-order families without a complete characterization.

Our main finding is that higher-order task-family constraints can recover true coordinates even without separately requiring tasks that directly read individual latent variables to remain first-order. These constraints can complement one another, with families that individually admit mixed representations jointly determining the true coordinates. Recovery conditions also vary with the number of latent variables, and in some higher-dimensional settings a single suitably chosen higher-order family suffices. We give the complete success and failure boundary for every dimension and every collection of higher-order task families. Under the task laws considered here, minimizing the average Fourier degree of predictors is equivalent to preserving the selected task spaces. We therefore ask whether the state similarities jointly determined by these spaces distinguish pairs differing in one variable from all other distinct pairs. Identifying this adjacency determines the coordinate structure. For every failing case, we construct an explicit reversible mixed code showing that different coordinate systems can satisfy the same task constraints.

Contributions.

❶ **Complete recovery classification.** We classify all dimensions and higher-order cumulative task-family combinations, construct explicit reversible mixed codes for every failure case, and characterize task-complexity minimizers under the stated task laws.

❷ **Task signatures identify coordinates.** Projection kernels reduce recovery to distinguishing Hamming neighbors through task-induced similarities, forcing every task-space-preserving relabeling to recover coordinates.

❸ **Robust finite-data recovery.** Fourier leakage bounds graph and coordinate errors and gives an exact-recovery threshold. A recovery algorithm handles finite, noisy, or partial responses in time polynomial in the state count.

❹ **Recovery beyond binary spaces.** We extend the classification to product spaces with at least three values per coordinate, where every proper, nonconstant cumulative task family identifies coordinates.

2 Related Work and Technical Background

Observable structure and latent identifiability. Distinct latent parameterizations can produce the same observation distribution through nonlinear observation maps, so observations alone gen-

erally cannot identify the true coordinates (Hyvärinen & Pajunen, 1999; Locatello et al., 2019). Prior work uses additional information to restrict this ambiguity. Nonlinear ICA uses auxiliary variables and assumptions on latent variation (Hyvärinen et al., 2019; Khemakhem et al., 2020); contrastive and multi-view methods use paired observations (Saunshi et al., 2019; Tosh et al., 2021; Zimmermann et al., 2021); causal and invariance methods use interventions and distribution shifts (Schölkopf et al., 2021; Ahuja et al., 2023; Arjovsky et al., 2019). Shared tasks also constrain latent representations (Bullins et al., 2019; Maurer et al., 2016; Tripuraneni et al., 2021), but predictors can compensate for changes in an invertible code, allowing different representations to yield the same task answers. A preference for simple solutions (Dingle et al., 2018; Huh et al., 2023) may further constrain encodings that increase task complexity.

Task complexity and coordinate recovery. Fourier–Walsh degree measures the highest order of coordinate interactions in a Boolean function (O'Donnell, 2014). Neural networks' spectral biases are related to this low-complexity preference (Rahaman et al., 2019; Basri et al., 2020; Tancik et al., 2020); related work also studies sparse-parity learning, degree-based curricula, and transformers' preference for low-sensitivity solutions (Barak et al., 2022; Abbe et al., 2023; Bhattamishra et al., 2023). Zhang et al. (2026) measure effective degree through polynomial fits along data interpolation paths and use it as a differentiable regularizer. Our coordinate-identification results concern exact Fourier degree on the Boolean cube. Zhang et al. (2025) apply the low-order preference to Boolean world models, selecting representations by minimizing the average Fourier degree of shared-task predictors and establishing a level-wise lower bound. When all degrees receive positive weight, optimality requires preserving the degree-one space spanned by constants and coordinate functions, thereby guaranteeing coordinate recovery.

Hamming symmetry and coordinate preservation. Connecting states that differ in exactly one binary fact gives the Boolean Hamming graph. Invertible relabelings preserving this graph are exactly coordinate permutations and independent value flips, so adjacency determines the coordinate structure (De Winter & Korb, 2016). Projection kernels of cumulative degree spaces connect task spaces to these state relations. They are radial kernels of the Hamming association scheme, depending only on Hamming distance, with values given by Krawtchouk polynomials (Delsarte, 1973; Bannai & Ito, 1984; Brouwer et al., 1989). Weak-isometry results directly require selected Hamming distances to be preserved (De Winter & Korb, 2016), whereas our task constraints require only preservation of the joint signatures assigned to state pairs by the projection kernels. Different distances may still have the same signature. We must determine whether these signatures uniquely identify distance-one pairs. Related structures also appear in automorphism groups of Reed–Muller codes (Berger & Charpin, 1993), but our invariant objects are nested real function spaces rather than a single code.

Coordinate recovery without a separate first-order constraint. The degree-one recovery argument does not extend directly to higher orders, because preserving a higher cumulative space still allows coordinate functions to become interactions among several variables. The three-dimensional counterexample of Zhang et al. (2025) shows that some higher-order constraints are insufficient, but leaves the recovery conditions for other dimensions and task-family combinations unresolved. We give a complete success and failure classification and prove, under the task laws considered here, that minimizing the average Fourier degree is equivalent to preserving the selected task spaces (Appendix C). The classification therefore also determines when this objective guarantees coordinate recovery.

3 Problem Setup

Each state in $\Omega_d = \{-1, 1\}^d$ records d binary facts. An unknown observation bijection $\psi : \Omega_d \to \mathcal{X}$ and a candidate representation bijection $\Phi : \mathcal{X} \to \Omega_d$ induce $T = \Phi \circ \psi$. In this finite representation class, task f is predicted by $f \circ T^{-1}$.

Use the normalized inner product $\langle f, g \rangle = 2^{-d} \sum_{x \in \Omega_d} f(x)g(x)$. This uniform reference measure is independent of state visitation frequencies. The Fourier-Walsh characters $\chi_S(x) = \prod_{i \in S} x_i$ form an orthonormal basis for real functions on Ω_d (O'Donnell, 2014). The Fourier degree of a nonzero function is $\deg(f) = \max\{|S| : \langle f, \chi_S \rangle \neq 0\}$, with $\deg(0) = 0$.

For $0 \le k \le d$, define the pure level and cumulative degree flag

$$\mathcal{H}_k = \mathrm{span}_{\mathbb{R}}\{\chi_S : |S| = k\}, \qquad \mathcal{F}^d_{\le k} = \bigoplus_{j=0}^{k} \mathcal{H}_j.$$

Thus $\mathcal{F}^d_{\le 0}$ contains only constants and $\mathcal{F}^d_{\le d}$ is the full function space. An active degree k always means $\mathcal{F}^d_{\le k}$; a pure degree-k parity is a character χ_S with $|S| = k$.

For weights $p_k \ge 0$ summing to one and task laws μ_k on $\mathcal{F}^d_{\le k}$, we minimize $\sum_{k=0}^{d} p_k \mathbb{E}_{f\sim\mu_k} \deg(f \circ T^{-1})$. Under the generic continuous and uniform finite-output laws of Section 4.3, minimizers are exactly the encodings preserving every cumulative space with positive weight. We classify when these preservation conditions force coordinate recovery.

Any cube bijection $T : \Omega_d \to \Omega_d$ induces the unitary pullback $U_T f = f \circ T$. It preserves $\mathcal{F}^d_{\le k}$ when $U_T \mathcal{F}^d_{\le k} = \mathcal{F}^d_{\le k}$.

The signed coordinate group $\mathcal{G}_d$ consists of maps $(T(x))_i = \varepsilon_i x_{\pi(i)}$, where $\varepsilon_i \in \{-1, 1\}$ and $\pi \in S_d$. These maps preserve every pure Fourier level and cumulative flag.

The orthogonal projection P_k onto $\mathcal{F}^d_{\le k}$ has kernel $K_k(x, y) = \sum_{|S|\le k} \chi_S(x)\chi_S(y)$. If T preserves $\mathcal{F}^d_{\le k}$, its projector and kernel satisfy

$$K_k(Tx, Ty) = K_k(x, y) \tag{1}$$

for all $x, y \in \Omega_d$.

The kernel depends only on the Hamming distance $r = \mathrm{dist}_H(x, y) = |\{j : x_j \ne y_j\}|$. Write $\kappa_k(r)$ for this value. Explicit formulas and their derivations are given in Appendix A.

4 Identifiability of Latent Factors

4.1 Kernel Signatures for Coordinate Recovery

Under the task laws in Section 3, optimal representations preserve the selected task spaces. Each space determines a projection kernel on state pairs. Preserving the space is equivalent to preserving its orthogonal projector and kernel (Theorem 4, Appendix B).

On the Boolean cube, each edge connects states differing in one variable. It suffices for the joint similarities to distinguish these pairs from all other distinct pairs.

Definition 1 (Edge-separating degree sets)**.** *A set $A \subseteq \{1, \dots, d-1\}$ is* edge-separating *in dimension d if, for every $r \in \{2, \dots, d\}$, there exists $k \in A$ such that $\kappa_k(r) \ne \kappa_k(1)$.*

Different task families may distinguish different distances. Thus their joint constraints can determine every adjacency even when individual families leave ambiguities.

Proof idea. For an edge x, y, kernel invariance gives $\kappa_k(\mathrm{dist}_H(Tx, Ty)) = \kappa_k(1)$ for every $k \in A$. If an encoding mapped this pair to nonadjacent states, edge separation would force at least one kernel value to change, contradicting invariance. The encoding therefore preserves the entire cube graph. Its automorphisms can only permute coordinates and independently flip their signs, which rules out mixed codes (Theorem 5, Appendix B).

4.2 Classification of Identifying Task Families

The cumulative kernels follow from Krawtchouk polynomials (Lemma 2, Appendix A). For each family, we list nonadjacent distances sharing its edge kernel value. A selection separates edges exactly when these collision sets have empty intersection.

Lemma 1 (All Boolean edge collisions)**.** *Let $d \ge 3$, $1 \le k \le d-1$, and $2 \le r \le d$. Then $\kappa_k(r) = \kappa_k(1)$ holds exactly in the following cases.*

$$\begin{array}{c|l} 1 \le k \le d-2 & k \text{ is even and } r = d, \\ k = d-1 & r \text{ is odd}. \end{array}$$

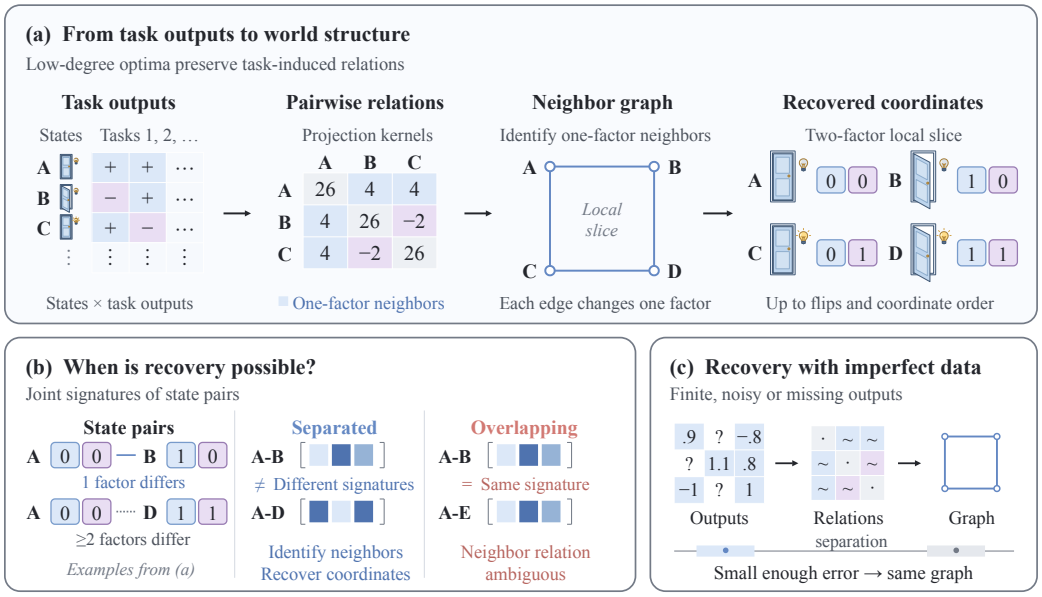


Figure 2: Overview. (a) Under the stated task laws, optimal encodings preserve selected task spaces and their pairwise relations. Complete responses spanning each space yield projection kernels; separating signatures identify coordinates. Entries illustrate the cumulative cubic family at $d = 5$; graphs show a local slice. (b) Signatures illustrate different task selections. Here E is a nonneighbor outside the slice; explicit mixed codes establish every failing selection. (c) Under the stated sampling and noise conditions, sufficiently small kernel estimation error relative to the separation margin permits graph recovery. Signature colors and error intervals are schematic.

These collisions explain when task families can complement one another. Lower even orders confuse only maximally separated states with neighbors, whereas the highest proper order leaves every larger odd distance ambiguous. In even dimension, the maximum distance is even, so the two sets of ambiguities are disjoint and joint constraints identify adjacency. In odd dimension, the maximum distance remains a shared ambiguity.

Theorem 1 (Complete Boolean degree-support classification)**.** *Fix $d \geq 3$ and $A \subseteq \{1, \ldots, d-1\}$. The following are equivalent.*

1. *Every cube bijection preserving $\mathcal{F}^d_{\leq k}$ for all $k \in A$ is a signed coordinate permutation.*
2. *The set A is edge-separating.*
3. *A contains an odd $k \leq d-2$, or d is even, $d-1 \in A$, and A contains an even $k \leq d-2$.*

In particular, a singleton flag $\mathcal{F}^d_{\leq k}$ identifies coordinates if and only if k is odd and $k \leq d-2$.

Proof sketch. Lemma 1 gives the degree condition by intersecting the collision sets. In particular, lower odd orders have no edge collisions. Theorem 5 yields coordinate recovery for every separating selection; Proposition 1 supplies counterexamples for all others. □

Kernel collisions alone do not establish failure. We also need mixed encodings preserving every selected space. Proposition 1 in Appendix B gives two constructions covering all failures. The even shear $A(x) = (x_1, x_1x_2, \ldots, x_1x_d)$ preserves every even flag and, in odd dimension, the highest proper flag. A second mixed map fixes total parity and preserves the highest proper flag in every dimension.

4.3 Population Minimizers and Coordinate Recovery

We now show why minimizing the average Fourier degree in Section 3 requires preserving every selected task space.

For the encoding $T = \Phi \circ \psi$, let $h_k(T)$ be the highest degree needed to represent all tasks in the cumulative family $\mathcal{F}^d_{\leq k}$ after re-encoding. We average these heights using the task-family weights p_k, giving

$$h_k(T) = \min\{m : U_{T^{-1}}\mathcal{F}^d_{\leq k} \subseteq \mathcal{F}^d_{\leq m}\}, \tag{2}$$

$$\mathcal{L}_p(T) = \sum_{k=0}^{d} p_k h_k(T). \tag{3}$$

Under a law absolutely continuous on each $\mathcal{F}^d_{\leq k}$, a sampled task almost surely attains this height, so the average equals the expected predictor degree. Uniform finite-output task laws have the same minimizing encodings (Appendix C).

Re-encoding preserves the dimension of each space, so $h_k(T) \geq k$. For $k < d$, equality holds exactly when the space is preserved (Lemma 4). Hence $\mathcal{L}_p(T) \geq \mathcal{L}^\star_p = \sum_{k=0}^{d} p_k k$. No family can fall below its own bound, and coordinate maps attain every bound. An optimal average therefore requires each family with positive weight to attain its bound, preserving every selected space.

Corollary 1 (Population minimizers)**.** *Let $A_p = \{1 \leq k < d : p_k > 0\}$. Under either task law, the global minimizers are exactly $\Phi = S \circ \psi^{-1}$ with $S \in \mathcal{G}_d$ if and only if A_p is edge-separating. Without degree one, $p_3 > 0$ suffices for $d \geq 5$ and $p_2, p_3 > 0$ for $d = 4$; for the full mixture in $d \geq 3$, $p_1 = 0$ and $p_k > 0$ $(2 \leq k \leq d)$, recovery holds iff $d \geq 4$.*

For an edge-separating support, integer degrees give a positive gap excluding mixed encodings. Under generic continuous laws, one complete truth table per selected family also identifies the global minimizers almost surely (Corollaries 2 and 3, Appendix C). This conclusion concerns optimization over all cube bijections; Section 5.2 constructs coordinates from tasks that span the selected spaces.

5 Approximate and Finite-Data Recovery

Exact degree jumps when any higher-level coefficient appears. We use Fourier leakage to quantify departures from space preservation, then study recovery from finite, noisy, or partial responses.

5.1 Fourier Leakage and Recovery Guarantees

Averaging the Fourier mass that leaves a task space after re-encoding over an orthonormal basis gives normalized leakage, which vanishes exactly when the space is preserved.

Write $n = 2^d$, $D_k = \dim \mathcal{F}^d_{\leq k}$, and let $\|\cdot\|_{\mathrm{HS}}$ denote the Hilbert–Schmidt norm on $L^2(\Omega_d)$. For a cube bijection T, define its normalized level k leakage

$$\ell_k(T) = \frac{1}{D_k}\left\|(I - P_k)U_{T^{-1}}P_k\right\|^2_{\mathrm{HS}}. \tag{4}$$

For an edge-separating set A and weights $\lambda_k > 0$, put $\mathcal{R}_{A,\lambda}(T) = \sum_{k\in A}\lambda_k\ell_k(T)$. Let $\rho_0(T, \mathcal{G}_d)$ be the minimum fraction of states on which T disagrees with a signed coordinate permutation. The positive margin $\mu_{A,\lambda}$ measures the minimum weighted squared separation between the joint kernel values of adjacent and nonadjacent pairs (Appendix E). Theorem 7 gives

$$\rho_0(T, \mathcal{G}_d) \leq \min\left\{1, \frac{4C_{\mathrm{FKN}}(d-1)n}{\mu_{A,\lambda}}\mathcal{R}_{A,\lambda}(T)\right\}.$$

Here C_{FKN} is universal. At fixed dimension and leakage, a larger margin yields a tighter coordinate error bound. The proof bounds broken edges through squared kernel changes, then applies balanced FKN rounding to obtain coordinate disagreement (Lemma 6, Appendix E).

Every mixed encoding has leakage at least $\beta_{A,\lambda}/n$ (Theorem 8), so leakage below this threshold forces a signed coordinate permutation. For the cubic family, the minimum nonzero leakage is

$\Theta((nd)^{-1})$ for all $d \geq 5$, with universal constants (Theorem 14). If $d \geq 11$ and $d > 2592\sqrt{C_{\mathrm{FKN}}}$, the minimizing non-coordinate maps are precisely signed coordinate maps composed with a transposition of two adjacent states. Appendix E.9 gives the exact minimum and full proof.

When leakage is estimated from data, a uniform objective estimation bound gives coordinate error bounds for approximate minimizers and an exact-recovery condition (Corollary 5, Appendix E.2). The appendix also extends the analysis to nonuniform discrete and continuous task laws whose second moments are well-conditioned on each active flag.

5.2 Coordinate Recovery from Task Responses

Under the observation map ψ of Section 3, task f supplies the response vector $g = f \circ \psi^{-1} \in \mathbb{R}^{\mathcal{X}}$ on all observed states. Once these vectors span a task space, their orthogonal projector gives the kernel used in Section 4 to identify adjacency.

Theorem 2 (Span-complete graph recovery)**.** *Let $A \subseteq \{1, \ldots, d-1\}$ be edge-separating. For each $k \in A$, suppose the observed tasks $g_{k,1}, \ldots, g_{k,D_k}$ have latent pullbacks forming a basis of $\mathcal{F}^d_{\leq k}$. Let Π_k project onto their observed span. Then distinct $x, y \in \mathcal{X}$ are latent Hamming neighbors exactly when*

$$\big(n(\Pi_k)_{xy}\big)_{k \in A} = \big(\kappa_k(1)\big)_{k \in A}. \tag{5}$$

The resulting graph determines the latent coordinates up to $\mathcal{G}_d$.

In the recovered cube graph, choose a root and assign its neighbors to coordinate directions. Graph distances then label all states. The choices of root and direction account for the independent sign flips and coordinate permutations.

Computing the projectors, testing all pairs, and labeling the graph cost $O(n \sum_{k \in A} D_k^2 + n^2 \sum_{k \in A} D_k + d^2 n)$ operations. For $A = \{3\}$ in identifying dimensions, $D_3 = 1 + d + \binom{d}{2} + \binom{d}{3} = \Theta(d^3)$ linearly independent tasks span the space; fewer than D_k vectors cannot span a D_k-dimensional flag. The algorithm processes all states, so its runtime is polynomial in the state count $n = 2^d$, which grows exponentially with the latent dimension d.

Noisy and partial responses. With noisy or missing responses, we use the known task sampling law to estimate the kernel directly. For each $k \in A$, sample m_k cumulative Walsh characters uniformly. The expected product of a task's responses at two states, scaled by the task-space dimension, equals their kernel value.

We average these products over tasks observed at both states. All state identifiers are known. Masks are independent across tasks and of task values, with pair co-inclusion probabilities at least $q_k > 0$. Noise is bounded, conditionally mean-zero and pair-uncorrelated across states, with the independence conditions of Theorem 11. Under its task budget, every pair is co-observed and every kernel estimate is within a quarter of the edge separation margin with probability at least $1 - \delta$. Thresholding at half the margin recovers the graph; coordinates follow from the same labeling procedure.

If each state is observed independently with probability p_k per task, $q_k = p_k^2$. For $d \geq 8$, constant observation rate, fixed confidence, and a fixed noise bound, cubic recovery uses $O(d^3)$ tasks. Without noise, a one-sided kernel test needs $O(d^2)$ tasks by a sharper variance bound (Theorem 12). These estimates use the specific task sampling law and do not require the response vectors to span the whole task space. Appendix E gives the full conditions, proofs, and lower bounds for state coverage and pairwise testing.

6 Recovery Beyond Binary Factors

For binary variables, some cumulative task families assign the same kernel value to adjacent and nonadjacent pairs, allowing mixed encodings to persist. We now ask whether these obstructions remain when each variable has more values.

On $\Omega_{q,d} = [q]^d$, split each coordinate space into constants and its mean-zero complement. Let $\mathcal{H}_S$ be mean-zero on coordinates in S and constant elsewhere, and define the ANOVA flag $\mathcal{F}^{q,d}_{\leq k} =$

$\bigoplus_{|S|\le k} \mathcal{H}_S$. Its projection kernel $\kappa_{q,k}(r)$ depends only on the Hamming distance $r = \mathrm{dist}_H(x, y)$. The dimension and kernel formulas are given in Appendix F.

Theorem 3 (Complete q-ary support classification)**.** *Let $q \ge 3$, $d \ge 2$, and $A \subseteq \{1, \dots, d-1\}$. Every bijection preserving $\mathcal{F}^{q,d}_{\le k}$ for all $k \in A$ is a coordinate permutation followed by independent within-coordinate symbol permutations if and only if A is nonempty.*

Thus any cumulative task space containing nonconstant functions but excluding some functions already determines the coordinate structure. The Boolean requirements on degree parity and combinations of families disappear here.

Proof sketch. For each selected space, the kernel sums products of k distinct entries from a list of $d-1$ numbers, each $q-1$ or -1. Adjacent pairs have only $q-1$ entries. At every other distance, some entries become -1; because $q-1 > 1$, each product's absolute value cannot increase and at least one strictly decreases. Hence $\kappa_{q,k}(r) < \kappa_{q,k}(1)$ for $r \ge 2$. The kernel principle and Hamming graph automorphisms give the claimed coordinate and symbol permutations. The empty support is unconstrained. Full details appear in Appendix F. □

7 Experiments

We check the classification by exact arithmetic, reconstruct coordinates from complete task responses, and test distance classification under perturbations. Frozen Transformer readouts separately assess task accessibility in fixed activations. This last block uses supervision on all states and does not test end-to-end coordinate discovery. Intervals are 1.96 standard errors over independent task draws with fixed encoders where applicable; protocols and seed groups appear in Appendix G.

7.1 Theory Checks and Recovery from Task Responses

Classification and margins. Cubic margins are $2, 8, 16$ at $d = 5, 6, 7$ and $(d-2)(d-3)$ for $d \ge 8$. In dimension four, quadratic and cubic gap vectors $(4, 4, 0)$ and $(2, 0, 2)$ give joint margin 2. Figure 3(a) to (e) checks these quantities, the singleton boundary, and the two obstructions to recovery from finite data.

Complete response reconstruction. At $d = 7$, projection onto the estimated task spaces followed by edge testing and BFS labeling recovers every coordinate and state in all 180 runs across $\{3, 5\}$, $\{2, 3, 4\}$, and $\{1\}$. Each support uses 8192 tasks with Gaussian response noise $\sigma = 0.25$, over two groups of 30 seeds. Figure 3(i) compares this procedure with calibrated MDS/FastICA on the same response matrices. The projector uses no distance calibration labels. Appendix G.2 gives both algorithms and their rejection criteria.

Calibrated distance classification. Disjoint calibration and test pairs evaluate a separate distance decoder. With 8192 tasks, nearest-centroid witness accuracy for $\{3, 5\}$ is 81.88% at $d = 7$ and 100% at $d = 11, 13, 14$. Witness accuracy evaluates a d-class predictor on distances 2 and $d-1$; it does not measure coordinate recovery. Identifying supports can still have collisions between these two distances. Under perturbations, feature standardization fitted on calibration data improves logistic regression's mean macro accuracy by 7.35 percentage points across 40 conditions on independent seeds. Nine witness conditions decline. Figure 4 in Appendix G.2 retains every perturbation and the unscaled baselines. State masking evaluates visible pairs only.

q-ary graph recovery. An analytic sweep covers 48 flags for $q = 2, \dots, 5$ and $d = 4, 5, 6$. At $d = 5$, the span projector recovers the exact graph for all four ternary flags and the two identifying binary flags in all 240 runs across both seed groups. Neither non-identifying binary flag recovers the complete graph. Figure 5 in Appendix G.3 compares projection with response averaging under the same normalization. The appendix also reports failures under limited budgets and noise.

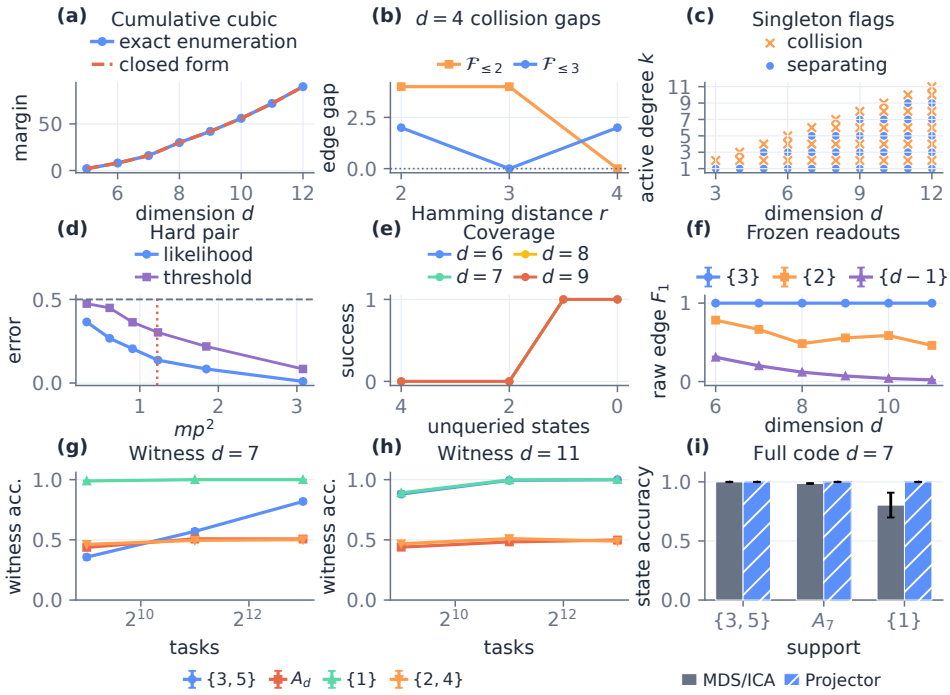


Figure 3: Theory and reconstruction. (a,b) Cubic margins and complementary $d = 4$ collisions. (c) Singleton boundary. (d,e) Co-observation and coverage obstructions. (f) Frozen readouts, with edge F_1 before graph correction. (g,h) Nearest-centroid witness accuracy, with $A_7 = \{2, 3, 4\}$ and $A_{11} = \{2, 4, 5\}$. (i) Whole-state accuracy from complete responses at $d = 7$, comparing MDS/FastICA with projection and BFS. Error bars show 1.96 standard errors over 10 task seeds in (f) and 30 seeds in (g) to (i).

7.2 Frozen Representation Readouts

One Transformer per dimension is trained on exact cubic parities and frozen. With 3000 cumulative cubic readouts, dimensions 6 through 11 attain 100% coordinate and state accuracy after graph correction, with raw edge $F_1 \geq 0.9997$ over the original ten task seeds (Figure 3(f)). For controls $\{2\}$ and $\{d-1\}$, both fitted and oracle responses are rejected in all 120 cases. Rejection is a pipeline outcome, not itself proof of a kernel collision. The highest family also exceeds readout capacity at larger dimensions. Appendix G.4 separates readout error, kernel error, and graph correction, and reports independent cubic task seeds.

8 Conclusion

Under the task laws considered here, we prove that the globally optimal representations minimizing the average Fourier degree of shared-task predictors are exactly the state relabelings preserving the selected task spaces. Projection kernels then reduce coordinate recovery to uniquely identifying state pairs that differ in one coordinate. Without a separate first-order constraint, the complete classification shows that recovery fails for $d = 3$, $d = 4$ requires both the up-to-two and up-to-three families, and the up-to-three family alone suffices for $d \geq 5$; every failing selection admits an explicit invertible mixed code. Our approximate-recovery theory bounds coordinate error and supports recovery from finite, noisy, or partial responses under the stated conditions, with runtime polynomial in the state count $n = 2^d$. Complete-response experiments recover coordinates, calibrated experiments test distance classification under perturbations, and supervised readouts from frozen Transformers recover coordinates on the evaluated states after graph correction. With at least three values per coordinate, every proper, nonconstant cumulative family identifies coordinates, making these collisions specific to binary spaces.

AI USE STATEMENT

We used generative AI tools to assist with mathematical derivations and proof writing, numerical verification, manuscript drafting and editing, and figure preparation. The authors take responsibility for the accuracy, claims, and final content of this work.

ETHICS STATEMENT

This work is theoretical, and its experiments use synthetic task responses and representations from Transformers trained on a synthetic Boolean environment. No human subjects were involved and no personal data were collected.

REPRODUCIBILITY STATEMENT

The assumptions and complete proofs are provided in the paper and appendices. Appendix G specifies task generation, pair splits, kernel estimators, graph reconstruction, evaluation metrics, and seed groups. The frozen readout protocol in Appendix G.4 specifies how saved activation tables are used for readout fitting and recovery. Reported statistics use the seed counts stated with each experiment.

REFERENCES

Emmanuel Abbe, Samy Bengio, Aryo Lotfi, and Kevin Rizk. Generalization on the unseen, logic reasoning and degree curriculum. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 31–60. PMLR, 2023. URL https://proceedings.mlr.press/v202/abbe23a.html.

Kartik Ahuja, Divyat Mahajan, Yixin Wang, and Yoshua Bengio. Interventional causal representation learning. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 372–407. PMLR, 2023. URL https://proceedings.mlr.press/v202/ahuja23a.html.

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019. URL https://arxiv.org/abs/1907.02893.

Eiichi Bannai and Tatsuro Ito. *Algebraic Combinatorics I: Association Schemes*. Benjamin/Cummings, Menlo Park, CA, 1984.

Boaz Barak, Benjamin Edelman, Surbhi Goel, Sham Kakade, Eran Malach, and Cyril Zhang. Hidden progress in deep learning: SGD learns parities near the computational limit. In *Advances in Neural Information Processing Systems*, volume 35, pp. 21750–21764, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/884baf65392170763b27c914087bde01-Abstract-Conference.html.

Ronen Basri, Meirav Galun, Amnon Geifman, David Jacobs, Yoni Kasten, and Shira Kritchman. Frequency bias in neural networks for input of non-uniform density. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 685–694, 2020. URL https://proceedings.mlr.press/v119/basri20a.html.

Thierry Berger and Pascale Charpin. The automorphism group of generalized Reed–Muller codes. *Discrete Mathematics*, 117(1–3):1–17, 1993. doi: 10.1016/0012-365X(93)90321-J.

Satwik Bhattamishra, Arkil Patel, Varun Kanade, and Phil Blunsom. Simplicity bias in transformers and their ability to learn sparse Boolean functions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5767–5791. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.acl-long.317. URL https://aclanthology.org/2023.acl-long.317/.

J. Bretagnolle and C. Huber. Estimation des densités: risque minimax. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 47(2):119–137, 1979. doi: 10.1007/BF00535278.

Andries E. Brouwer, Arjeh M. Cohen, and Arnold Neumaier. *Distance-Regular Graphs*, volume 18 of *Ergebnisse der Mathematik und ihrer Grenzgebiete. 3. Folge*. Springer, Berlin, 1989.

Brian Bullins, Elad Hazan, Adam Kalai, and Roi Livni. Generalize across tasks: Efficient algorithms for linear representation learning. In *Proceedings of the 30th International Conference on Algorithmic Learning Theory*, volume 98 of *Proceedings of Machine Learning Research*, pp. 235–246, 2019.

Stefaan De Winter and Mason Korb. Weak isometries of the Boolean cube. *Discrete Mathematics*, 339(2):877–885, 2016. doi: 10.1016/j.disc.2015.10.006.

Philippe Delsarte. *An Algebraic Approach to the Association Schemes of Coding Theory*. PhD thesis, Université catholique de Louvain, 1973. Philips Research Reports Supplements, No. 10.

Kamaludin Dingle, Chico Q. Camargo, and Ard A. Louis. Input–output maps are strongly biased towards simple outputs. *Nature Communications*, 9(1):761, 2018. doi: 10.1038/s41467-018-03101-6.

Ehud Friedgut, Gil Kalai, and Assaf Naor. Boolean functions whose Fourier transform is concentrated on the first two levels. *Advances in Applied Mathematics*, 29(3):427–437, 2002. doi: 10.1016/S0196-8858(02)00024-6.

Karl Friston, Rosalyn J. Moran, Yukie Nagai, Tadahiro Taniguchi, Hiroaki Gomi, and Josh Tenenbaum. World model learning and inference. *Neural Networks*, 144:573–590, 2021. doi: 10.1016/j.neunet.2021.09.011.

David Ha and Jürgen Schmidhuber. Recurrent world models facilitate policy evolution. In *Advances in Neural Information Processing Systems*, volume 31, pp. 2451–2463. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/hash/2de5d16682c3c35007e4e92982f1a2ba-Abstract.html.

Minyoung Huh, Hossein Mobahi, Richard Zhang, Brian Cheung, Pulkit Agrawal, and Phillip Isola. The low-rank simplicity bias in deep networks. *Transactions on Machine Learning Research*, 2023. URL https://openreview.net/forum?id=bCiNWDmlY2.

Aapo Hyvärinen. Fast and robust fixed-point algorithms for independent component analysis. *IEEE Transactions on Neural Networks*, 10(3):626–634, 1999. doi: 10.1109/72.761722.

Aapo Hyvärinen and Petteri Pajunen. Nonlinear independent component analysis: Existence and uniqueness results. *Neural Networks*, 12(3):429–439, 1999. doi: 10.1016/S0893-6080(98)00140-3.

Aapo Hyvärinen, Hiroaki Sasaki, and Richard E. Turner. Nonlinear ICA using auxiliary variables and generalized contrastive learning. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pp. 859–868, 2019.

Ilyes Khemakhem, Diederik P. Kingma, Ricardo Pio Monti, and Aapo Hyvärinen. Variational autoencoders and nonlinear ICA: A unifying framework. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 2207–2217. PMLR, 2020. URL https://proceedings.mlr.press/v108/khemakhem20a.html.

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 4114–4124, 2019.

Andreas Maurer, Massimiliano Pontil, and Bernardino Romera-Paredes. The benefit of multitask representation learning. *Journal of Machine Learning Research*, 17(81):1–32, 2016. URL https://www.jmlr.org/papers/v17/15-242.html.

Seyed Morteza Mirafzal and Meysam Ziaee. A note on the automorphism group of the Hamming graph. *Transactions on Combinatorics*, 10(2):129–136, 2021. doi: 10.22108/toc.2021.127225.1817.

Ryan O'Donnell. *Analysis of Boolean Functions*. Cambridge University Press, Cambridge, 2014.

Nasim Rahaman, Aristide Baratin, Devansh Arpit, Felix Draxler, Min Lin, Fred A. Hamprecht, Yoshua Bengio, and Aaron Courville. On the spectral bias of neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5301–5310, 2019. URL https://proceedings.mlr.press/v97/rahaman19a.html.

Nikunj Saunshi, Orestis Plevrakis, Sanjeev Arora, Mikhail Khodak, and Hrishikesh Khandeparkar. A theoretical analysis of contrastive unsupervised representation learning. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5628–5637, 2019. URL https://proceedings.mlr.press/v97/saunshi19a.html.

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021. doi: 10.1109/JPROC.2021.3058954.

Matthew Tancik, Pratul P. Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan T. Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. In *Advances in Neural Information Processing Systems*, volume 33, pp. 7537–7547, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/55053683268957697aa39fba6f231c68-Abstract.html.

Christopher Tosh, Akshay Krishnamurthy, and Daniel Hsu. Contrastive learning, multi-view redundancy, and linear models. In *Proceedings of the 32nd International Conference on Algorithmic Learning Theory*, volume 132 of *Proceedings of Machine Learning Research*, pp. 1179–1206, 2021. URL https://proceedings.mlr.press/v132/tosh21a.html.

Nilesh Tripuraneni, Chi Jin, and Michael I. Jordan. Provable meta-learning of linear representations. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 10434–10443, 2021. URL https://proceedings.mlr.press/v139/tripuraneni21a.html.

Tianren Zhang, Guanyu Chen, and Feng Chen. When do neural networks learn world models? In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 74639–74670. PMLR, 2025. URL https://proceedings.mlr.press/v267/zhang25j.html.

Tianren Zhang, Xiangxin Li, Minghao Xiao, Guanyu Chen, and Feng Chen. Quantifying and optimizing simplicity via polynomial representations. In *International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2605.29823.

Roland S. Zimmermann, Yash Sharma, Steffen Schneider, Matthias Bethge, and Wieland Brendel. Contrastive learning inverts the data generating process. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 12979–12990. PMLR, 2021. URL https://proceedings.mlr.press/v139/zimmermann21a.html.

A KERNEL CALCULATIONS

This appendix gives the kernel identities used in the classification proof. Let $u_j = x_j y_j$ and $r = \mathrm{dist}_H(x, y)$. Then $u_j = -1$ on exactly r coordinates, $u_j = 1$ on the remaining coordinates, and

$$s = \sum_{j=1}^{d} u_j = d - 2r.$$

A.1 THE DEGREE-TWO KERNEL

For $k = 2$,

$$K_2(x, y) = 1 + \sum_{j=1}^{d} u_j + \sum_{1 \le i < j \le d} u_i u_j.$$

Since

$$\begin{aligned}\left(\sum_{j=1}^{d} u_j\right)^2 &= \sum_{j=1}^{d} u_j^2 + 2 \sum_{1 \le i < j \le d} u_i u_j \\ &= d + 2 \sum_{1 \le i < j \le d} u_i u_j,\end{aligned}$$

we obtain

$$\kappa_2(r) = 1 + s + \frac{s^2 - d}{2} = \frac{(s+1)^2 + 1 - d}{2}. \tag{6}$$

If $\mathrm{dist}_H(x, y) = 1$, then $s = d - 2$. If $r' = \mathrm{dist}_H(Tx, Ty)$ and $s' = d - 2r'$, kernel-invariance gives

$$\frac{(s'+1)^2 + 1 - d}{2} = \frac{(d-1)^2 + 1 - d}{2}.$$

Hence

$$(d - 2r' + 1)^2 = (d - 1)^2.$$

The two integer solutions in $0 \le r' \le d$ are $r' = 1$ and $r' = d$.

A.2 THE DEGREE-THREE KERNEL

For $k = 3$, Newton's identities give

$$\sum_{1 \le i < j < \ell \le d} u_i u_j u_\ell = \frac{s^3 - 3sd + 2s}{6} = \frac{s^3 - (3d-2)s}{6}.$$

Hence

$$\begin{aligned}\kappa_3(r) &= 1 + s + \frac{s^2 - d}{2} + \frac{s^3 - (3d-2)s}{6} \\ &= \frac{(s+1)(s^2 + 2s + 6 - 3d)}{6}.\end{aligned} \tag{7}$$

The separation proof uses the factorization

$$\begin{aligned}\kappa_3(1) - \kappa_3(r) = \frac{d - s - 2}{6} \\ \cdot \big(s^2 + (d+1)s + d^2 - 4d + 6\big), \\ s = d - 2r.\end{aligned}$$

For $d \ge 5$ and $r \ge 2$, the first factor is $2(r-1) > 0$. The second factor is at least its real quadratic minimum

$$d^2 - 4d + 6 - \frac{(d+1)^2}{4} = \frac{3d^2 - 18d + 23}{4} > 0,$$

so the cubic flag alone separates edges from all other distances.

A.3 THE $d-1$ KERNEL

Assume $r > 0$. The sum of all Fourier levels is

$$\sum_{S\subseteq[d]} \chi_S(x)\chi_S(y) = \prod_{j=1}^{d}(1+u_j) = 0.$$

The degree d term is

$$\mathsf{K}_d^{(d)}(r) = \prod_{j=1}^{d} u_j = (-1)^r.$$

Hence

$$\kappa_{d-1}(r) = -\mathsf{K}_d^{(d)}(r) = -(-1)^r.$$

For an edge, $\kappa_{d-1}(1) = 1$. If d is even, then $\kappa_{d-1}(d) = -1$. These two values separate edges from antipodal pairs.

A.4 THE $d-2$ KERNEL

Again assume $r > 0$. Let

$$\mathsf{K}_m^{(d)}(r) = \sum_{|S|=m} \prod_{j\in S} u_j$$

be the pure level m kernel. Since the sum over all levels is zero,

$$\kappa_{d-2}(r) = -\big(\mathsf{K}_{d-1}^{(d)}(r) + \mathsf{K}_d^{(d)}(r)\big).$$

We already have $\mathsf{K}_d^{(d)}(r) = (-1)^r$. For $\mathsf{K}_{d-1}^{(d)}$,

$$\begin{aligned}\mathsf{K}_{d-1}^{(d)}(r) &= \sum_{i=1}^{d}\prod_{j\neq i} u_j \\ &= \left(\prod_{j=1}^{d} u_j\right)\sum_{i=1}^{d} u_i = (-1)^r(d-2r),\end{aligned}$$

using $u_i^{-1} = u_i$. Hence

$$\kappa_{d-2}(r) = -(-1)^r(d-2r+1).$$

When d is odd, $\kappa_{d-2}(1) = d-1$ and $\kappa_{d-2}(d) = 1-d$, so this kernel separates edges from antipodal pairs for every odd $d \geq 5$.

A.5 PROOFS OF THE CUMULATIVE IDENTITY AND THE CLASSIFICATION

Write the pure degree j polynomial as $\mathsf{K}_j^{(m)}(s) = [t^j](1+t)^{m-s}(1-t)^s$.

Lemma 2 (Cumulative-kernel identity)**.** *For $1 \leq k \leq d-1$ and $1 \leq r \leq d$,*

$$\kappa_k^{(d)}(r) = \mathsf{K}_k^{(d-1)}(r-1), \qquad \kappa_k^{(d)}(1) = \binom{d-1}{k}.$$

Proof of Lemma 2. The generating polynomial for the pure Fourier-level kernels at distance r is

$$\sum_{j=0}^{d} \mathsf{K}_j^{(d)}(r)t^j = (1+t)^{d-r}(1-t)^r.$$

Taking cumulative coefficients is equivalent to dividing by $1-t$, giving

$$\sum_{j=0}^{k} \mathsf{K}_j^{(d)}(r) = [t^k]\frac{(1+t)^{d-r}(1-t)^r}{1-t} = \mathsf{K}_k^{(d-1)}(r-1).$$

At $r = 1$ the last generating polynomial is $(1+t)^{d-1}$. □

Proof of Lemma 1. Put $N = d - 1$ and $s = r - 1$. By Lemma 2, $\kappa_k(r)$ is the sum of the products of all k subsets of N signs, exactly s of which are -1. Its edge value is $\binom{N}{k}$, the maximum possible sum of the $\binom{N}{k}$ summands in $\{-1, 1\}$. Equality therefore holds if and only if every k subset has product $+1$.

Assume first that $k \leq N - 1$. For any two indices i, j, choose a $(k - 1)$ subset disjoint from them (with the empty set allowed when $k = 1$). Comparing the two resulting k subset products shows the signs at i and j are equal. Hence all N signs are equal. Since $s \geq 1$, they must all be -1, and then every k-fold product is positive exactly when k is even. This is exactly $s = N$, or $r = d$. If $k = N$, there is one summand, equal to $(-1)^s$; it is positive exactly when s is even, equivalently when r is odd. □

Proof of Theorem 1. Collecting the collisions of Lemma 1 over the active degrees shows that an odd $k \leq d-2$ leaves no collision, that a support of even degrees below the top retains exactly $r = d$, that the top flag alone retains every odd distance and in particular $r = 3$, and that the top flag removes the surviving distance $r = d$ precisely when d is even. Appendix A.6 carries out the four-case enumeration in full. Consequently, conditions (2) and (3) are equivalent, and Theorem 5 proves that either condition implies (1).

For the converse, suppose condition (3) fails. If d is odd, the active support consists only of lower even degrees and possibly $d - 1$; the universal even shear in Proposition 1 preserves all of them and is non-coordinate. If d is even and $d - 1 \notin A$, the same shear preserves every active lower even flag. If d is even and $d - 1 \in A$ but there is no active lower degree, the top-parity shear from that proposition preserves the sole active flag. The empty-support case is immediate as well. Thus every non-edge-separating support has a non-coordinate flag preserver. This proves (1)⇒(3). □

A.6 Enumeration of edge-separating supports

This subsection completes the equivalence of conditions (2) and (3) in Theorem 1. For a single degree k let

$$C_k = \{r \in \{2, \ldots, d\} : \kappa_k(r) = \kappa_k(1)\}$$

be its set of edge collisions. By Definition 1, a support A is edge-separating exactly when

$$\bigcap_{k \in A} C_k = \emptyset,$$

where the intersection over $A = \emptyset$ is all of $\{2, \ldots, d\}$. Lemma 1 gives the three possible cases

$$C_k = \begin{cases} \emptyset, & k \leq d - 2 \text{ odd}, \\ \{d\}, & k \leq d - 2 \text{ even}, \\ \{r \in \{2, \ldots, d\} : r \text{ odd}\}, & k = d - 1. \end{cases}$$

Case 1. A **contains an odd** $k \leq d - 2$**.** Then $C_k = \emptyset$ for that degree, so the intersection is empty and A is edge-separating. This is the first alternative in condition (3).

In each of the three remaining cases A contains no odd $k \leq d - 2$, so every element of A is an even $k \leq d - 2$ or equals $d - 1$, and the first alternative in condition (3) fails. The cases split on how $d - 1$ and the even lower degrees occur.

Case 2. $d - 1 \notin A$**.** Every element of A is then an even $k \leq d - 2$, so the intersection is $\{d\}$ when A is nonempty and $\{2, \ldots, d\}$ when $A = \emptyset$. Both are nonempty, so A is not edge-separating. The second alternative in condition (3) fails because it requires $d - 1 \in A$.

Case 3. $d - 1 \in A$ **and** A **contains no even** $k \leq d - 2$**, that is,** $A = \{d - 1\}$**.** The intersection is C_{d-1}, which contains $r = 3$ because $d \geq 3$. Hence A is not edge-separating, and the second alternative in condition (3) fails because it requires an active even $k \leq d - 2$.

Case 4. $d-1 \in A$ **and** A **contains an even** $k \le d-2$**.** The intersection is $\{d\} \cap C_{d-1}$. For odd d the distance d is odd, so the intersection is $\{d\}$ and A is not edge-separating; the second alternative in condition (3) also fails because it requires even d. For even d the distance d is even, so $d \notin C_{d-1}$, the intersection is empty, and A is edge-separating; this is exactly the second alternative in condition (3).

Case 1 is delimited by the presence of an odd $k \le d-2$, and Cases 2, 3 and 4 partition the supports without such a degree according to whether $d-1 \in A$ and whether an even $k \le d-2$ is active. The four cases are therefore mutually exclusive and exhaust every $A \subseteq \{1, \dots, d-1\}$. In each case A is edge-separating precisely when one of the alternatives in condition (3) holds, which proves (2)⇔(3).

As a check, $d = 3$ admits no even $k \le d-2 = 1$, so Cases 1 and 3 are the only ones with nonempty A. The support $\{1\}$ separates and $\{2\}$ does not. For $d = 4$ the support $\{2, 3\}$ falls in Case 4 with d even and separates, while $\{2\}$ and $\{3\}$ fall in Cases 2 and 3 and do not. For $d \ge 5$ the singleton $\{3\}$ falls in Case 1.

A.7 Exact verification of the enumeration

The case analysis above was checked mechanically in exact integer arithmetic, with the kernel values obtained from the cumulative-Krawtchouk identity of Lemma 2 as $\kappa_k(r) = [t^k](1+t)^{d-r}(1-t)^{r-1}$; no floating-point arithmetic enters the audit.

The collision sets of Lemma 1 were recomputed directly and compared with the three predicted forms for every pair (d, k) with $3 \le d \le 80$ and $1 \le k \le d-1$. All 3159 pairs agree. The passage from those collision sets to the two alternatives in condition (3) was then checked exhaustively, by intersecting $\bigcap_{k \in A} C_k$ and comparing with Theorem 1 for every one of the 2^{d-1} supports at each $3 \le d \le 20$. All 1,048,572 supports agree, and each falls in exactly one of the four cases.

Case	delimiting condition	supports	separating
	odd $k \le d-2$		
1	active	1,045,506	1,045,506
	no such k,		
2	$d-1 \notin A$	1,533	0
3	$A = \{d-1\}$	18	0
	$d-1 \in A$,		
4	even $k \le d-2$ active	1,515	1,013

The separating supports in Case 4 are exactly the 1,013 with d even, which is the branch isolated by the second alternative in condition (3). Here the top flag resolves the edge collision left by an even interior flag, and the exact audit checks this joint-signature mechanism directly.

B Cube Graph Automorphisms

Theorem 4 (Finite-space kernel stabilizer principle)**.** *Let $\mathcal{X}$ be finite with the uniform measure, and let $\{V_a : a \in I\}$ be function subspaces with projection kernels K_a. A permutation T preserves every V_a if and only if*

$$(K_a(Tx, Ty))_{a \in I} = (K_a(x, y))_{a \in I} \qquad (x, y \in \mathcal{X}).$$

Proof. Because U_T is unitary, preserving V_a is equivalent to preserving its projector, whose conjugated kernel is $|\mathcal{X}|^{-1} K_a(Tx, Ty)$. □

Lemma 3. *Let $A : \Omega_d \to \Omega_d$ be a bijection such that*

$$\mathrm{dist}_H(x, y) = 1 \quad \Longrightarrow \quad \mathrm{dist}_H(Ax, Ay) = 1.$$

Then A is a signed coordinate permutation.

Proof. Since A is a bijection on the finite vertex set and sends every edge to an edge, it is an automorphism of the cube graph. Let $\mathbf{1} = (1, \dots, 1)$ and set

$$B(x) = A(\mathbf{1}) \odot A(x),$$

where $\odot$ is coordinatewise multiplication. This output sign change is a cube graph automorphism, and $B(\mathbf{1}) = \mathbf{1}$.

Identify vertices with subsets $S \subseteq [d]$, where S records the coordinates flipped from $\mathbf{1}$. The neighbors of $\mathbf{1}$ are the singletons. The map B fixes $\mathbf{1}$ and therefore permutes the singletons; call this permutation π.

We prove that $B(S) = \pi(S)$ for all $S \subseteq [d]$. The claim is immediate for $|S| = 0$ and $|S| = 1$. Assume it holds for sets of size $m-1$, and take $|S| = m$. Graph automorphisms preserve graph distance from $\mathbf{1}$, so $B(S)$ is a set of size m. For every $i \in S$, the set S is adjacent to $S \setminus \{i\}$, hence $B(S)$ is adjacent to

$$B(S \setminus \{i\}) = \pi(S \setminus \{i\}).$$

Adjacency between sets of sizes m and $m-1$ requires containment. Thus $B(S)$ contains $\bigcup_{i \in S} \pi(S \setminus \{i\}) = \pi(S)$, where $m \geq 2$. Both sets have size m, so $B(S) = \pi(S)$.

Returning from subsets to sign vectors, $B_i(x) = x_{\pi(i)}$ after relabeling coordinates. Undoing the output sign change gives $A_i(x) = \varepsilon_i x_{\pi(i)}$ for signs $\varepsilon_i \in \{-1, 1\}$. □

Proposition 1 (Universal obstruction maps)**.** *For every $d \geq 3$ there is one non-coordinate cube bijection preserving simultaneously every even flag $\mathcal{F}^d_{\leq k}$. When d is odd, the same bijection also preserves $\mathcal{F}^d_{\leq d-1}$. In every dimension $d \geq 3$ there is a non-coordinate bijection preserving $\mathcal{F}^d_{\leq d-1}$.*

Proof of Proposition 1. Fix $d \geq 3$ and define

$$A_1(x) = x_1, \qquad A_i(x) = x_1 x_i \quad (2 \leq i \leq d).$$

This is a bijection, with inverse $x_1 = A_1$ and $x_i = A_1 A_i$ for $i \geq 2$, and it is non-coordinate. For a set S of size m,

$$\chi_S(A(x)) = x_1^m \prod_{i \in S \setminus \{1\}} x_i.$$

If m is even, this character has degree at most m. If m is odd, it has degree m when $1 \in S$ and degree $m+1$ when $1 \notin S$. Hence, for every even k and every $m \leq k$, the image of a pure level m character has degree at most k. Hence

$$U_A \mathcal{F}^d_{\leq k} \subseteq \mathcal{F}^d_{\leq k} \qquad \text{for every even } k.$$

Bijectivity and equality of dimensions turn every inclusion into equality, so one shear preserves all even flags simultaneously.

For the same map,

$$\prod_{i=1}^{d} A_i(x) = x_1^d \prod_{i=2}^{d} x_i.$$

When d is odd this is the top-parity character. Since $\mathcal{F}^d_{\leq d-1}$ is its orthogonal complement, the same shear also preserves $\mathcal{F}^d_{\leq d-1}$ in odd dimension.

For a top-flag obstruction in every dimension, define

$$B_1(x) = x_1 x_2, \quad B_2(x) = x_2,$$
$$B_i(x) = x_i \quad (3 \leq i \leq d-1), \quad B_d(x) = x_2 x_d.$$

Its inverse is $x_2 = B_2$, $x_1 = B_1 B_2$, $x_i = B_i$ for $3 \leq i \leq d-1$, and $x_d = B_2 B_d$, so B is a non-coordinate bijection. Its top-parity remains fixed because

$$\prod_{i=1}^{d} B_i(x) = x_1 x_2 \cdots x_d.$$

Hence unitarity again gives $U_B \mathcal{F}^d_{\leq d-1} = \mathcal{F}^d_{\leq d-1}$, completing the proposition. □

Theorem 5 (Edge-separating flags force coordinate maps)**.** *Let $T : \Omega_d \to \Omega_d$ be a bijection, and let $A \subseteq \{1, \ldots, d-1\}$ be edge-separating. If $U_T \mathcal{F}^d_{\leq k} = \mathcal{F}^d_{\leq k}$ for every $k \in A$, then T is a signed coordinate permutation. In particular, every coordinate function T_i has Fourier degree one.*

Proof of Theorem 5. Fix adjacent $x, y \in \Omega_d$, so $\mathrm{dist}_H(x, y) = 1$, and let

$$r' = \mathrm{dist}_H(Tx, Ty).$$

Since T is injective, $r' \neq 0$. For each $k \in A$, preservation of $\mathcal{F}^d_{\leq k}$ and the kernel-invariance identity equation 1 give

$$\kappa_k(r') = K_k(Tx, Ty) = K_k(x, y) = \kappa_k(1).$$

Because A is edge-separating, no distance $r' \in \{2, \ldots, d\}$ has all these kernel values. Therefore $r' = 1$. Every Hamming-edge is mapped to a Hamming-edge, and Lemma 3 identifies T as a signed coordinate permutation. □

C From Task Complexity to Space Preservation

The optimization domain consists of the bijective representations in Section 3. If the shared tasks separate all latent states, zero-error prediction requires an injective representation; a map into the equally sized cube is then bijective. A finite task panel need not satisfy this separation condition.

The uniform reference inner product defines the Fourier spaces used in this algebraic classification. It imposes no uniformity requirement on state visitation frequencies. Estimating these kernels from nonuniform observations requires appropriate coverage and weighting; an empirical projection using nonuniform state weights need not have a radial kernel.

This appendix spells out the finite-dimensional reduction behind equation 3. Let W be a nonzero linear subspace of functions on Ω_d, and define its degree height by

$$h(W) = \min\{m : W \subseteq \mathcal{F}^d_{\leq m}\}.$$

A probability law on W is *subspace-generic* if it gives zero mass to every proper linear subspace of W. Any law absolutely continuous with respect to Lebesgue measure on W is subspace-generic. Under such a law, a task has Fourier degree $h(W)$ almost surely. When $h(W) = 0$, the only exceptional vector is zero, which is itself a proper subspace. When $h(W) \geq 1$, the vectors of degree below $h(W)$ form the proper linear subspace $W \cap \mathcal{F}^d_{\leq h(W)-1}$.

For the level k latent task family, take $W = \mathcal{F}^d_{\leq k}$ and then compose tasks with the inverse of the induced representation map. The represented task family is

$$U_{T^{-1}} \mathcal{F}^d_{\leq k}.$$

Its generic Fourier degree is therefore

$$h(U_{T^{-1}} \mathcal{F}^d_{\leq k}) = \min\{m : U_{T^{-1}} \mathcal{F}^d_{\leq k} \subseteq \mathcal{F}^d_{\leq m}\} = h_k(T),$$

which is the term used in equation 3. Its lower bound and equality condition in Lemma 4 are the level-wise condition in Lemma E.10 used to prove Theorem 4.8 of Zhang et al. (2025). The theorem below makes the bridge explicit through a finite output alphabet that formalizes the bounded-precision counting convention used there, while $\mathcal{Y} = \{-1, 1\}$ gives the Boolean-valued special case.

Theorem 6 (Task-law bridge)**.** *Fix a finite output alphabet $\mathcal{Y} \subset \mathbb{R}$ with $|\mathcal{Y}| \geq 2$. For $0 \leq k \leq d$, let*

$$\mathcal{Q}_k = \{f \in \mathcal{F}^d_{\leq k} : f(\Omega_d) \subseteq \mathcal{Y}\}.$$

For the finite sum below, use the harmless convention $\deg(0) = 0$; the zero task is fixed by every composition and therefore cannot affect minimizers. For a subspace-generic law μ_k on $\mathcal{F}^d_{\leq k}$, define

$$J^{\mathrm{gen}}_k(T) = \mathbb{E}_{f \sim \mu_k} \deg(f \circ T^{-1}),$$
$$J^{\mathrm{fin}}_k(T) = \frac{1}{|\mathcal{Q}_k|} \sum_{f \in \mathcal{Q}_k} \deg(f \circ T^{-1}).$$

Then

$$J^{\mathrm{gen}}_k(T) = h_k(T), \qquad J^{\mathrm{fin}}_k(T) \geq J^{\mathrm{fin}}_k(\mathrm{Id}),$$

and equality in the second relation holds if and only if $U_{T^{-1}}\mathcal{F}^d_{\le k} = \mathcal{F}^d_{\le k}$. Consequently, for any nonnegative weights $p_0, \dots, p_d$, the generic continuous mixture and the uniform finite-output mixture have the same global minimizers over cube bijections, satisfying

$$\arg\min_T \sum_{k=0}^{d} p_k J_k^{\mathrm{gen}}(T) = \mathcal{M}_p,$$
$$\arg\min_T \sum_{k=0}^{d} p_k J_k^{\mathrm{fin}}(T) = \mathcal{M}_p.$$

Here $A_p = \{1 \le k < d : p_k > 0\}$, and $\mathcal{M}_p$ denotes the cube bijections satisfying $U_{T^{-1}}\mathcal{F}^d_{\le k} = \mathcal{F}^d_{\le k}$ for every $k \in A_p$. The two objectives may differ numerically. Their common minimizers define the same identification problem.

Proof. The first identity follows from the preceding height calculation because the push-forward of a subspace-generic law by the invertible linear map $U_{T^{-1}}$ remains subspace-generic.

For the finite-output law, put $C = U_{T^{-1}}\mathcal{Q}_k$. Composition by a cube bijection permutes all $\mathcal{Y}$-valued functions, so $|C| = |\mathcal{Q}_k|$ and

$$J_k^{\mathrm{fin}}(T) = \frac{1}{|\mathcal{Q}_k|} \sum_{g \in C} \deg(g).$$

If $C \neq \mathcal{Q}_k$, the sets $A = C \setminus \mathcal{Q}_k$ and $D = \mathcal{Q}_k \setminus C$ have the same positive cardinality. Every member of A has degree at least $k+1$, whereas every member of D has degree at most k. Hence

$$\begin{aligned} &|\mathcal{Q}_k|\big(J_k^{\mathrm{fin}}(T) - J_k^{\mathrm{fin}}(\mathrm{Id})\big) \\ &\quad = \sum_{g \in A} \deg(g) - \sum_{g \in D} \deg(g) \\ &\quad \ge |A| > 0. \end{aligned}$$

Therefore equality holds exactly when $U_{T^{-1}}\mathcal{Q}_k = \mathcal{Q}_k$.

It remains to compare set and subspace preservation. Choose distinct $a, b \in \mathcal{Y}$, and set $c = (a+b)/2$ and $\delta = (a-b)/2 \neq 0$. At least one of a, b is nonzero, so the constant function $\mathbf{1}$ lies in $\operatorname{span}_{\mathbb{R}} \mathcal{Q}_k$. For every $|S| \le k$, the function $f_S = c\mathbf{1} + \delta\chi_S$ takes only the values a, b, hence belongs to $\mathcal{Q}_k$. Therefore $\chi_S = (f_S - c\mathbf{1})/\delta$ lies in the same span, and

$$\operatorname{span}_{\mathbb{R}} \mathcal{Q}_k = \mathcal{F}^d_{\le k}.$$

Thus preservation of $\mathcal{Q}_k$ implies preservation of $\mathcal{F}^d_{\le k}$. Conversely, if $U_{T^{-1}}\mathcal{F}^d_{\le k} = \mathcal{F}^d_{\le k}$, composition preserves $\mathcal{Y}$-valuedness and maps $\mathcal{Q}_k$ bijectively to itself. This proves the level-wise equivalence.

Finally, $h_k(T) \ge k$, with equality exactly when the k-th flag is preserved (the endpoints $k = 0, d$ are automatic), while the finite objective has the just-proved level-wise lower bound. The identity map attains all these bounds simultaneously. A nonnegative mixture is minimal exactly when every positive-weight, nontrivial level is minimal, which gives the displayed common argmin. □

For the exact-degree formulation with $k \ge 1$, let ν_k be Lebesgue measure on $\mathcal{F}^d_{\le k}$, set $E_k = \mathcal{F}^d_{\le k} \setminus \mathcal{F}^d_{\le k-1}$, and assume $\mu_k(E_k) = 1$ and $\mu_k \ll \nu_k|_{E_k}$. The exceptional set on which the represented degree drops below $h_k(T)$ is

$$N_{T,k} = \mathcal{F}^d_{\le k} \cap U_T \mathcal{F}^d_{\le h_k(T)-1}.$$

It is a proper linear subspace of $\mathcal{F}^d_{\le k}$, hence $\nu_k(N_{T,k}) = 0$ and $(\nu_k|_{E_k})(N_{T,k} \cap E_k) = 0$. Absolute continuity gives $\mu_k(N_{T,k} \cap E_k) = 0$. Hence the represented degree is again $h_k(T)$ almost surely. For the one-task result, the union of $N_{T,k}$ over the finite set of cube bijections remains null, so one draw realizes the height simultaneously for every candidate T.

C.1 Population minimizers and finite tasks

Lemma 4 (Level-wise lower bound and equality). *For every cube bijection T and every $0 \le k \le d$, $h_k(T) \ge k$. For $k < d$, equality holds if and only if $U_{T^{-1}}\mathcal{F}^d_{\le k} = \mathcal{F}^d_{\le k}$. For $k = d$, equality is automatic.*

Proof. For $k = 0$, constants compose to constants, so $h_0(T) = 0$ and $U_{T^{-1}}\mathcal{F}^d_{\le 0} = \mathcal{F}^d_{\le 0}$. Now let $k \ge 1$. The subspace $U_{T^{-1}}\mathcal{F}^d_{\le k}$ has the same dimension as $\mathcal{F}^d_{\le k}$, because $U_{T^{-1}}$ is unitary. If $h_k(T) < k$, then $U_{T^{-1}}\mathcal{F}^d_{\le k} \subseteq \mathcal{F}^d_{\le k-1}$, but $\dim \mathcal{F}^d_{\le k} > \dim \mathcal{F}^d_{\le k-1}$, a contradiction. Therefore $h_k(T) \ge k$.

If $h_k(T) = k$, then $U_{T^{-1}}\mathcal{F}^d_{\le k} \subseteq \mathcal{F}^d_{\le k}$, and the two spaces have equal finite dimension; therefore they are equal. The converse is immediate from the definition of $h_k(T)$. When $k = d$, the flag $\mathcal{F}^d_{\le d}$ is the whole function space, so $h_d(T) = d$ for every bijection. □

Proof of Corollary 1. Theorem 6 shows that the generic continuous and finite-output objectives have exactly the same minimizing bijections, namely those preserving every active flag. Theorem 1 makes all of these minimizers signed coordinate maps exactly when A_p is edge-separating. If it is not, Proposition 1 provides a non-coordinate minimizer preserving every active flag.

For the full higher-degree mixture, $A_p = \{2, \dots, d-1\}$. In dimension 3, this is the non-separating singleton $\{2\}$, and the even shear in Proposition 1 is a non-coordinate minimizer. In dimension 4, the pair $\{2, 3\}$ satisfies the second alternative in condition (3) of Theorem 1; for every $d \ge 5$, the support contains the odd index $3 \le d-2$ and satisfies the first alternative in condition (3). Hence recovery holds exactly for $d \ge 4$. □

This also gives a gap in the average degree objective. If the selected families identify coordinates, any mixed encoding must increase the height of at least one family. Since degrees are integers, this increase incurs a minimum cost set by the family's weight.

Corollary 2 (Population gap). *Let $A_0 \subseteq A_p$ be edge-separating and set $\gamma_p(A_0) = \min_{k \in A_0} p_k$. If $\mathcal{L}_p(T) < \mathcal{L}^\star_p + \gamma_p(A_0)$, then T is a signed coordinate permutation.*

Proof. Otherwise some $k \in A_0$ has the positive integer gap $h_k(T) - k \ge 1$, so $\mathcal{L}_p(T) - \mathcal{L}^\star_p \ge p_k \ge \gamma_p(A_0)$. □

For generic continuous tasks, finitely many tasks also suffice to identify the minimizers. For any candidate encoding, a sampled task almost surely attains the height of its encoded family. There are only finitely many candidate bijections, so this property holds simultaneously for all candidates with probability one.

Corollary 3 (Finite generic cumulative-task truth tables). *Fix an edge-separating set B and weights $q_k > 0$. For each $k \in B$, draw $n_k \ge 1$ tasks $f_{k,j}$ independently from laws absolutely continuous on $\mathcal{F}^d_{\le k}$. Define the empirical objective*

$$\widehat{\mathcal{L}}_{B,q}(T) = \sum_{k \in B} \frac{q_k}{n_k} \sum_{j=1}^{n_k} \deg(f_{k,j} \circ T^{-1}).$$

With probability one, its global minimizers over cube bijections are exactly the signed coordinate permutations. Thus for $d \ge 5$ one generic cumulative cubic truth table suffices; for $d = 4$, one cumulative quadratic and one cumulative cubic truth table suffice.

Proof. Fix T and k. The set of tasks $f \in \mathcal{F}^d_{\le k}$ for which

$$\deg(f \circ T^{-1}) < h_k(T)$$

is the proper linear subspace

$$\mathcal{F}^d_{\le k} \cap U_T \mathcal{F}^d_{\le h_k(T)-1}.$$

It therefore has probability zero under any absolutely continuous task law on $\mathcal{F}^d_{\le k}$. For exact-degree laws, the same null set intersected with $\mathcal{F}^d_{\le k} \setminus \mathcal{F}^d_{\le k-1}$ has zero restricted Lebesgue measure. There are only finitely many cube bijections T and finitely many sampled tasks. Taking the union of these null events, with probability one we have

$$\deg(f_{k,j} \circ T^{-1}) = h_k(T)$$

for every sampled pair (k, j) and every cube bijection T simultaneously. On this event,

$$\widehat{\mathcal{L}}_{B,q}(T) = \sum_{k \in B} q_k h_k(T)$$

for every T. The proof of Corollary 1, with B in place of A_p and weights q_k, then identifies the global minimizers as exactly the signed coordinate permutations whenever B is edge-separating. The cubic and four-dimensional finite-task claims follow from Theorem 1. □

D Response Perturbations and Kernel Checks

D.1 From response-noise to projector error

We prove the response-matrix implication in Theorem 10. The following elementary bound also displays the conditioning parameter that controls tasks close to a rank-deficient panel.

Lemma 5 (Column-space perturbation)**.** *Let $G \in \mathbb{R}^{n \times D}$ have full column rank with $\sigma = \sigma_{\min}(G)$, let $\widetilde{G} = G + E$, and let $\Pi, \widetilde{\Pi}$ be the orthogonal projectors onto their column spaces. If $e = \|E\|_{\mathrm{op}} < \sigma$, then*

$$\|\widetilde{\Pi} - \Pi\|_{\mathrm{op}} \le \frac{e}{\sigma - e}.$$

Proof. For every coefficient vector a,

$$\|\widetilde{G}a\|_2 \ge (\sigma - e)\|a\|_2,$$
$$\|(I - \Pi)\widetilde{G}a\|_2 = \|(I - \Pi)Ea\|_2 \le e\|a\|_2.$$

Hence every unit vector in $\mathrm{col}(\widetilde{G})$ lies within $e/(\sigma - e)$ of $\mathrm{col}(G)$, which gives $\|(I - \Pi)\widetilde{\Pi}\|_{\mathrm{op}} \le e/(\sigma - e)$. The two spaces have the same dimension D; the elementary principal angle identity for equal rank orthogonal projectors gives

$$\|\widetilde{\Pi} - \Pi\|_{\mathrm{op}} = \|(I - \Pi)\widetilde{\Pi}\|_{\mathrm{op}}.$$

This establishes the claim. □

For every $k < d$, the projection-kernel Cauchy–Schwarz inequality gives $|\kappa_k(r)| \le D_k < n$, and hence $\Delta_A < 2n$. Under equation 24, write $a = \|E_k\|_{\mathrm{op}}/\sigma_k < \Delta_A/(4n) < 1/2$. Lemma 5 yields

$$\|\widehat{\Pi}_k - \Pi_k\|_{\mathrm{op}} \le \frac{a}{1 - a} < 2a < \frac{\Delta_A}{2n}.$$

Every matrix entry is bounded by the operator norm, so equation 22 follows.

For a finite-observation specialization, assume each clean entry of every G_k is measured R times with independent, conditionally mean-zero errors in $[-b, b]$, and $\widetilde{G}_k$ contains the entrywise averages. Put

$$D_\Sigma = \sum_{k \in A} D_k, \qquad L_\delta = \log \frac{2nD_\Sigma}{\delta}.$$

Hoeffding's inequality and a union bound over all response entries give, simultaneously for every k, with probability at least $1 - \delta$,

$$\|E_k\|_{\mathrm{op}} \le \|E_k\|_{\mathrm{F}} \le b\sqrt{\frac{2nD_k L_\delta}{R}}.$$

Hence the explicit condition

$$R > \frac{32b^2n^3L_\delta}{\Delta_A^2} \max_{k\in A} \frac{D_k}{\sigma_k^2} \tag{8}$$

implies equation 24 for every active flag and therefore exact graph and coordinate recovery with probability at least $1-\delta$. The smallest singular values σ_k quantitatively record proximity to the exceptional rank-deficient task panels. For a complete Walsh basis panel, $G_k^\top G_k = nI_{D_k}$ and $\sigma_k^2 = n$. In this case, the sufficient repetition budget simplifies to $R > 32b^2n^2L_\delta \max_{k\in A} D_k/\Delta_A^2$.

D.2 Small dimensional kernel-audit

Table 1 lists every distance-kernel value used at the $d = 3, 4, 5$ boundaries. The edge value is the entry at $r = 1$. It shows directly that the quadratic flag collides with the antipode in all three dimensions, that the cubic flag in $d = 4$ has a different collision, and that their pair separates edges; in $d = 5$ the cubic edge value is unique.

Table 1: Exact values of $\kappa_k(r)$ for the boundary dimensions.

d	k	$(\kappa_k(0), \kappa_k(1), \ldots, \kappa_k(d))$
3	2	(7, 1, −1, 1)
4	2	(11, 3, −1, −1, 3)
4	3	(15, 1, −1, 1, −1)
5	2	(16, 6, 0, −2, 0, 6)
5	3	(26, 4, −2, 0, 2, −4)

An independent exact-arithmetic check enumerates Walsh characters and cube states for $d = 3, 4, 5$. It verifies the table, the closed forms equation 6 and equation 7, bijectivity of the displayed obstruction maps, preservation of the quadratic flag by A, and preservation of top-parity by B. These finite checks supplement the symbolic proofs in Section 4, which establish the results for all stated dimensions.

E Proofs and Additional Results for Quantitative Recovery

Sections E.1 and E.2 prove the leakage and stability guarantees used in Section 5.1. Sections E.8 and E.9 give the general leakage bounds and the sharper cubic result. The recovery algorithms of Section 5.2 are developed in Sections E.3 through E.6, followed by the lower bounds in Section E.7.

E.1 Leakage bounds and graph recovery

Technical lemmas

Lemma 6 (Balanced FKN rounding)**.** *There is a universal constant C_{FKN} such that every balanced Boolean function $f : \Omega_d \to \{-1, 1\}$ has a signed dictator $g(x) = \varepsilon x_j$ satisfying*

$$\Pr[f(X) \neq g(X)] \le C_{\mathrm{FKN}}\big(\mathsf{I}(f) - 1\big),$$

where X is uniform and $\mathsf{I}(f)$ is total influence.

Proof. The Friedgut-Kalai-Naor theorem bounds the distance to a Boolean affine function by a universal constant times the Fourier mass above level one (Friedgut et al., 2002). Parseval and balancedness give

$$\mathsf{I}(f) - 1 = \sum_{|S|\ge 2} (|S| - 1)\widehat{f}(S)^2 \ge \sum_{|S|\ge 2} \widehat{f}(S)^2.$$

A balanced Boolean function is at distance $1/2$ from either constant; enlarging the universal constant in that case makes the approximating affine function a signed dictator. This is the displayed statement. □

Lemma 7 (Boolean derivative quantization)**.** *Let $0 \le k < d$ and $f : \Omega_d \to \{-1, 1\}$. If $f \notin \mathcal{F}^d_{\le k}$, then*

$$\big\|(I - P_k)f\big\|_2^2 \ge \frac{2^{1-k}}{n}. \tag{9}$$

Proof. For $i \in [d]$, set $\partial_i f(x) = (f(x) - f(x^{\oplus i}))/2$, and for $J \subseteq [d]$ let $\partial_J = \prod_{i \in J} \partial_i$. Fourier expansion gives

$$\|\partial_J f\|_2^2 = \sum_{S \supseteq J} \widehat{f}(S)^2.$$

Because f has a nonzero coefficient above level k, some J of size $k+1$ makes $\partial_J f$ nonzero. At each point, $\partial_J f$ is $2^{-(k+1)}$ times an alternating sum of 2^{k+1} signs. Every nonzero value therefore has magnitude at least 2^{-k}. Also $\partial_J f(x^{\oplus i}) = -\partial_J f(x)$ for $i \in J$, so a nonzero value occurs on an entire 2^{k+1} point J subcube. Therefore

$$\sum_{|S|>k} \widehat{f}(S)^2 \ge \|\partial_J f\|_2^2 \ge \frac{2^{k+1}}{n}\, 2^{-2k} = \frac{2^{1-k}}{n},$$

which is equation 9. □

GEOMETRIC CONTROL FROM LEAKAGE

For an edge-separating set A and weights $\lambda_k > 0$, put $\mathcal{R}_{A,\lambda}(T) = \sum_{k \in A} \lambda_k \ell_k(T)$ and

$$\mu_{A,\lambda} = \min_{2 \le r \le d} \sum_{k \in A} \frac{\lambda_k}{D_k} \big(\kappa_k(r) - \kappa_k(1)\big)^2;$$

the margin $\mu_{A,\lambda} > 0$ measures the minimum weighted squared difference between the joint kernel values of adjacent and nonadjacent pairs. Every broken adjacency therefore incurs a quantifiable kernel change.

We measure errors in both adjacency and coordinates. Let $B(T)$ be the number of undirected Hamming edges $\{x, y\}$ whose inverse images are not adjacent, and let $\rho_0(T, \mathcal{G}_d) = \min_{S \in \mathcal{G}_d} n^{-1} |\{x : T(x) \ne S(x)\}|$ be the normalized vertex disagreement from the signed coordinate group.

Theorem 7 (Fourier-mass robustness and geometric control)**.** *For every cube bijection T,*

$$\mathcal{R}_{A,\lambda}(T) \ge \frac{\mu_{A,\lambda}}{n^2}\, B(T). \tag{10}$$

There is a dimension-free constant C_{FKN} such that

$$\rho_0(T, \mathcal{G}_d) \le \min\left\{1, \frac{4C_{\mathrm{FKN}}(d-1)n}{\mu_{A,\lambda}} \mathcal{R}_{A,\lambda}(T)\right\}. \tag{11}$$

Proof. Let $Q_{T,k} = U_{T^{-1}} P_k U_T$, the projector onto the represented flag. Unitary invariance of the Hilbert–Schmidt norm and the equal rank projection identity give

$$\begin{aligned} \ell_k(T) &= \frac{1}{D_k} \|(I - P_k) Q_{T,k}\|_{\mathrm{HS}}^2 = \frac{1}{2D_k} \|P_k - Q_{T,k}\|_{\mathrm{HS}}^2 \\ &= \frac{1}{2D_k n^2} \sum_{x,y \in \Omega_d} \big[K_k(x,y) - K_k(T^{-1}x, T^{-1}y)\big]^2. \end{aligned} \tag{12}$$

The last equality uses the orthonormal point masses $n^{1/2}\mathbf{1}_{\{x\}}$, for which the corresponding projection-matrix entry is $K_k(x, y)/n$. Each bad undirected edge contributes two ordered pairs to equation 12. Summing over k and using the definition of $\mu_{A,\lambda}$ proves equation 10.

To prove equation 11, put $R = T^{-1}$ and write $f_i(x) = R_i(x)$. Since R is a bijection, every f_i is balanced and the f_i are pairwise orthogonal. With the standard total influence $\mathsf{I}(f) = \sum_{j=1}^d \Pr[f(X) \ne f(X^{\oplus j})]$, Fourier expansion gives $\mathsf{I}(f_i) \ge 1$. The same expansion gives

$$\begin{aligned} \sum_{i=1}^d \big(\mathsf{I}(f_i) - 1\big) &= \frac{2}{n} \sum_{\{x,y\} \in E(\Omega_d)} \big(\mathrm{dist}_H(Rx, Ry) - 1\big) \\ &\le \frac{2(d-1)}{n} B(T). \end{aligned}$$

The balanced FKN conclusion in Lemma 6 gives signed dictators $g_i = \varepsilon_i x_{\pi(i)}$ with

$$\delta_i := \Pr[f_i \neq g_i] \leq C_{\mathrm{FKN}}\big(\mathrm{I}(f_i) - 1\big).$$

Put $\delta = \sum_i \delta_i$. If $\delta < 1/2$, the indices $\pi(i)$ are distinct. Otherwise, for some $i \neq j$ one has $\mathbb{E}[g_i g_j] \in \{-1, 1\}$ but $\mathbb{E}[f_i f_j] = 0$, whereas changing either factor bounds the difference of these expectations by $2(\delta_i + \delta_j) < 1$. Hence $(g_1, \ldots, g_d)$ is a signed coordinate map and $\rho_0(R, \mathcal{G}_d) \leq \delta$. If $\delta \geq 1/2$, the trivial bound $\rho_0 \leq 1 \leq 2\delta$ applies. Inversion preserves permutation disagreement, so in both cases

$$\rho_0(T, \mathcal{G}_d) = \rho_0(R, \mathcal{G}_d) \leq 2\delta \leq \frac{4C_{\mathrm{FKN}}(d-1)}{n} B(T).$$

Combining with equation 10 proves equation 11. □

THE LEAKAGE GAP

Theorem 8 (State-scale leakage gap)**.** *Define the explicit quantization constant* $\beta_{A,\lambda} = \min_{k \in A} \frac{\lambda_k 2^{1-k}}{D_k}$. *Then every non-coordinate cube bijection satisfies*

$$T \notin \mathcal{G}_d \quad \Longrightarrow \quad \mathcal{R}_{A,\lambda}(T) \geq \frac{\beta_{A,\lambda}}{n}. \tag{13}$$

Proof. Expanding the Hilbert–Schmidt norm in equation 4 in the Walsh basis gives

$$D_k \ell_k(T) = \sum_{|S| \leq k} \big\|(I - P_k) U_{T^{-1}} \chi_S\big\|_2^2.$$

If T is non-coordinate, edge separation implies at least one active flag is not preserved. For such a k, some Boolean function $f = U_{T^{-1}} \chi_S$ occurring in the sum is not in $\mathcal{F}^d_{\leq k}$. Lemma 7 gives

$$\big\|(I - P_k) f\big\|_2^2 \geq \frac{2^{1-k}}{n}.$$

We obtain

$$\mathcal{R}_{A,\lambda}(T) \geq \lambda_k \ell_k(T) \geq \frac{\lambda_k 2^{1-k}}{D_k n} \geq \frac{\beta_{A,\lambda}}{n}.$$

This establishes equation 13.

□

E.2 WELL-CONDITIONED CUMULATIVE TASK LAWS

For each $k \in A$, let F_k be a square-integrable random task taking values in $\mathcal{F}^d_{\leq k}$ almost surely, and let

$$C_k g = \mathbb{E}[\langle F_k, g \rangle F_k]$$

be its second-moment operator. Define

$$\mathcal{Q}_{A,\lambda}(T) = \sum_{k \in A} \lambda_k \mathbb{E}\big\|(I - P_k) U_{T^{-1}} F_k\big\|_2^2.$$

Theorem 9 (Well-conditioned cumulative task laws)**.** *Let A be edge-separating and $\lambda_k > 0$. Suppose that, for every $k \in A$, constants $0 < a_k \leq b_k < \infty$ satisfy*

$$a_k P_k \preceq C_k \preceq b_k P_k. \tag{14}$$

Define $\underline{\lambda}_k = \lambda_k a_k D_k$ and $\overline{\lambda}_k = \lambda_k b_k D_k$. Then every cube bijection T satisfies

$$\mathcal{R}_{A,\underline{\lambda}}(T) \leq \mathcal{Q}_{A,\lambda}(T) \leq \mathcal{R}_{A,\overline{\lambda}}(T). \tag{15}$$

Consequently, $\mathcal{Q}_{A,\lambda}(T) = 0$ if and only if $T \in \mathcal{G}_d$. Moreover,

$$\mathcal{Q}_{A,\lambda}(T) \geq \frac{\mu_{A,\underline{\lambda}}}{n^2} B(T), \tag{16}$$

$$\rho_0(T, \mathcal{G}_d) \leq \min\left\{1, \frac{4C_{\mathrm{FKN}}(d-1)n}{\mu_{A,\underline{\lambda}}} \mathcal{Q}_{A,\lambda}(T)\right\}. \tag{17}$$

Every non-coordinate map also obeys

$$\mathcal{Q}_{A,\lambda}(T) \geq \frac{\beta_{A,\underline{\lambda}}}{n}. \tag{18}$$

Proof. Set $A_{T,k} = U_T(I - P_k)U_{T^{-1}}$. This is an orthogonal projection, and

$$\mathbb{E}\big\|(I - P_k)U_{T^{-1}}F_k\big\|_2^2 = \operatorname{tr}(A_{T,k}C_k).$$

Cyclicity of trace gives

$$\operatorname{tr}(A_{T,k}P_k) = \big\|(I - P_k)U_{T^{-1}}P_k\big\|_{\mathrm{HS}}^2 = D_k\ell_k(T).$$

Since $A_{T,k}$ is positive semidefinite, equation 14 implies

$$a_k D_k \ell_k(T) \le \operatorname{tr}(A_{T,k}C_k) \le b_k D_k \ell_k(T).$$

Multiplying by λ_k and summing proves equation 15. Its lower side and Theorems 7 and 8 give equation 16 through equation 18. Edge separation and positivity of $\underline{\lambda}_k$ identify the zero set as $\mathcal{G}_d$. □

For example, if $F_k = \varepsilon\chi_S$, where ε is an independent Rademacher sign and $\Pr(S = R) = \pi_{k,R} > 0$ for every $|R| \le k$, then the theorem applies with $a_k = \min_R \pi_{k,R}$ and $b_k = \max_R \pi_{k,R}$. For the uniform cumulative Walsh law, $a_k = b_k = D_k^{-1}$ and $\mathcal{Q}_{A,\lambda} = \mathcal{R}_{A,\lambda}$ exactly.

Corollary 4 (State-resolved task-law stability)**.** *Write $\|M\|_*$ for the nuclear norm, the sum of the singular values of M. Let C_k satisfy Theorem 9, and let another second-moment operator $\widetilde{C}_k$ satisfy*

$$\|\widetilde{C}_k - C_k\|_* \le \varepsilon_k, \qquad \varepsilon_{A,\lambda} = \sum_{k\in A} \lambda_k \varepsilon_k.$$

Define $\widetilde{\mathcal{Q}}_{A,\lambda}(T) = \sum_{k\in A} \lambda_k \operatorname{tr}(A_{T,k}\widetilde{C}_k)$, where $A_{T,k} = U_T(I - P_k)U_{T^{-1}}$. If T is an η-approximate minimizer of $\widetilde{\mathcal{Q}}_{A,\lambda}$ and $e_{\mathrm{state}} = n(2\varepsilon_{A,\lambda} + \eta)$, then

$$\frac{B(T)}{dn/2} \le \min\left\{1, \frac{2e_{\mathrm{state}}}{d\mu_{A,\underline{\lambda}}}\right\}, \tag{19}$$

$$\rho_0(T, \mathcal{G}_d) \le \min\left\{1, \frac{4C_{\mathrm{FKN}}(d-1)e_{\mathrm{state}}}{\mu_{A,\underline{\lambda}}}\right\}. \tag{20}$$

If $e_{\mathrm{state}} < \beta_{A,\underline{\lambda}}$, then $T \in \mathcal{G}_d$.

Proof. Trace duality and $\|A_{T,k}\|_{\mathrm{op}} = 1$ give, uniformly in T,

$$|\widetilde{\mathcal{Q}}_{A,\lambda}(T) - \mathcal{Q}_{A,\lambda}(T)| \le \varepsilon_{A,\lambda}.$$

For any $S \in \mathcal{G}_d$, Theorem 9 gives $\mathcal{Q}_{A,\lambda}(S) = 0$. Approximate optimality therefore gives

$$\mathcal{Q}_{A,\lambda}(T) \le 2\varepsilon_{A,\lambda} + \eta = \frac{e_{\mathrm{state}}}{n}.$$

The lower sandwich in equation 15 yields $\mathcal{R}_{A,\underline{\lambda}}(T) \le e_{\mathrm{state}}/n$. Theorem 7 now gives both displayed error bounds, and Theorem 8 gives the exact condition. □

Corollary 5 (State-resolved stability)**.** *Suppose an objective $\widehat{\mathcal{R}}$ satisfies $\sup_T |\widehat{\mathcal{R}}(T) - \mathcal{R}_{A,\lambda}(T)| \le \varepsilon$, and T is an η-approximate minimizer of $\widehat{\mathcal{R}}$. Put $e = n(2\varepsilon + \eta)$. Then*

$$\frac{B(T)}{dn/2} \le \min\left\{1, \frac{2e}{d\mu_{A,\lambda}}\right\},$$

$$\rho_0(T, \mathcal{G}_d) \le \min\left\{1, \frac{4C_{\mathrm{FKN}}(d-1)e}{\mu_{A,\lambda}}\right\}.$$

If $e < \beta_{A,\lambda}$, then $T \in \mathcal{G}_d$.

Proof of Corollary 5. For $S \in \mathcal{G}_d$, $\mathcal{R}_{A,\lambda}(S) = 0$. Hence

$$\mathcal{R}_{A,\lambda}(T) \le \widehat{\mathcal{R}}(T) + \varepsilon \le \widehat{\mathcal{R}}(S) + \eta + \varepsilon \le 2\varepsilon + \eta = \frac{e}{n}.$$

The claims follow from Theorems 7 and 8. □

E.3 Algorithm details and noisy projectors

Recovery from complete task spans

Proof of Theorem 2. Because the observed tasks span the transported flag, Π_k is its projection operator. Hence

$$(\Pi_k)_{xy} = n^{-1}\kappa_k\big(\mathrm{dist}_H(\psi^{-1}x, \psi^{-1}y)\big).$$

Injectivity and edge separation make equation 5 equivalent to latent Hamming adjacency. To label the recovered cube graph, choose a root o and assign its d neighbors to the d coordinate directions. If $m = \mathrm{dist}_H(o, x)$, coordinate i is flipped at x exactly when the distance from x to the neighbor assigned to coordinate i is $m - 1$. Therefore $d + 1$ breadth-first distance computations produce a graph isomorphism Φ. Any two such labelings differ by a cube automorphism, so $\Phi \circ \psi \in \mathcal{G}_d$. Thin QR factorizations compute the bases Q_k in $O(n\sum_{k\in A} D_k^2)$ operations. Forming all entries of $\Pi_k = Q_k Q_k^\top$ costs $O(n^2\sum_{k\in A} D_k)$; comparing them with the edge signatures costs $O(|A|n^2)$. The distance computations cost $O(d^2 n)$ on the cube graph with $dn/2$ edges. Since $D_k \geq 1$, the total is $O(n\sum_{k\in A} D_k^2 + n^2\sum_{k\in A} D_k + d^2 n)$. □

Kernel signature recovery procedure. The following procedure implements the constructive statements and sets the finite-precision convention.

1. From a complete panel, compute a thin QR factorization $G_k = Q_k R_k$ and use $(\widehat{\Pi}_k)_{xy} = \langle (Q_k)_{x,:}, (Q_k)_{y,:}\rangle$. From partial responses, accumulate the co-observed products in Theorem 11.
2. For every unordered pair x, y, apply either the exact test equation 5 or the strict finite-precision test equation 23. We report a value on the threshold as an insufficient-precision certificate and reject the reconstruction.
3. Verify that the resulting graph is connected, d regular, has 2^d vertices, and passes the distance-label consistency check in the next step. Choose a root o, order its d neighbors, and run breadth-first search from o and from each ordered neighbor.
4. Label x by the sign vector whose ith coordinate differs from the root exactly when $\mathrm{dist}(x, o_i) = \mathrm{dist}(x, o) - 1$. Reject if either $\mathrm{dist}(x, o_i)$ is not one of $\mathrm{dist}(x, o) \pm 1$ or two vertices receive the same label. Finally check that every recovered edge joins labels at Hamming-distance-one and that every Hamming-distance-one label pair is an edge.

Under the deterministic hypotheses of Theorems 2 and 10, the verification succeeds and the output agrees with the latent coordinates up to $\mathcal{G}_d$. Under Theorem 11, the same conclusion holds on its recovery event, which has probability at least $1 - \delta$. These checks require $O(d^2 n)$ time after the edge set is formed.

Closed-form margins

Define the unnormalized edge-signature margin

$$\Delta_A = \min_{2\leq r\leq d} \max_{k\in A} |\kappa_k(r) - \kappa_k(1)| > 0.$$

Proposition 2 (Closed-form Boolean recovery margins)**.** *Every Boolean edge-separating support satisfies $\Delta_A \geq 2$. In the two smallest constructive cases,*

$$\Delta_{\{2,3\}} = 2 \quad (d = 4),$$

$$\Delta_{\{3\}} = \begin{cases} 2, & d = 5, \\ 8, & d = 6, \\ 16, & d = 7, \\ (d-2)(d-3), & d \geq 8. \end{cases}$$

For a singleton cubic objective, $\mu_{\{3\},\lambda} = \lambda_3 \Delta_{\{3\}}^2 / D_3$. In dimension four,

$$\mu_{\{2,3\},\lambda} = \min\left\{\frac{16\lambda_2}{11}, \frac{4\lambda_3}{15}\right\}.$$

Proof. By definition, each $\kappa_k(r)$ is a sum of D_k signs. Hence every difference $\kappa_k(r) - \kappa_k(1)$ is an even integer, so $\Delta_A \geq 2$. For $d = 4$, formulas equation 6 and equation 7 give the absolute gap vectors, indexed by $r = 2, 3, 4$,

$$(4, 4, 0) \quad \text{and} \quad (2, 0, 2).$$

This establishes the stated Δ and, after squaring and dividing by $D_2 = 11, D_3 = 15$, the stated μ.

For the cubic singleton, the factorization in equation 7 gives, for $2 \leq r \leq d$,

$$\begin{aligned} g_d(r) &:= \kappa_3(1) - \kappa_3(r) \\ &= \frac{r-1}{3}\left(3d^2 - 6dr - 3d + 4r^2 - 2r + 6\right) > 0. \end{aligned} \tag{21}$$

For $d \geq 9$,

$$g_d(r) - g_d(2) = \frac{r-2}{3}\Big(3d^2 - 6dr - 9d + 4r^2 + 2r + 12\Big).$$

The quadratic in parentheses has real minimum $(3d^2 - 30d + 47)/4 > 0$, so the minimum is attained at $r = 2$ and equals $g_d(2) = (d-2)(d-3)$. Direct substitution in equation 21 for $d = 5, 6, 7, 8$ gives respectively

$$(6, 4, 2, 8), \quad (12, 12, 8, 8, 20),$$
$$(20, 24, 20, 16, 20, 40),$$
$$(30, 40, 38, 32, 30, 40, 70),$$

which proves all remaining cases. The formulas for μ now follow directly from its definition. □

NOISY PROJECTORS

Theorem 10 (Finite-precision graph recovery)**.** *In the setting of Theorem 2, suppose matrices $\widehat{\Pi}_k$ satisfy*

$$\max_{x,y} |(\widehat{\Pi}_k - \Pi_k)_{xy}| < \frac{\Delta_A}{2n} \qquad (k \in A). \tag{22}$$

Then declaring x, y adjacent when

$$\max_{k \in A} \left| (\widehat{\Pi}_k)_{xy} - \frac{\kappa_k(1)}{n} \right| < \frac{\Delta_A}{2n} \tag{23}$$

recovers the exact latent Hamming graph.

Proof. For a true edge, the left side of equation 23 is below the threshold by equation 22. For a nonedge at distance $r \geq 2$, some k has exact signature gap at least Δ_A/n, so the reverse triangle inequality puts the estimated gap strictly above the threshold. □

Response-matrix error. Let $G_k \in \mathbb{R}^{n \times D_k}$ be the clean response matrix, let $\sigma_k = \sigma_{\min}(G_k) > 0$, and form $\widehat{\Pi}_k$ from the column span of $\widetilde{G}_k = G_k + E_k$. The deterministic response-noise condition

$$\|E_k\|_{\mathrm{op}} < \frac{\sigma_k \Delta_A}{4n} \qquad (k \in A) \tag{24}$$

implies equation 22, hence exact coordinate recovery. The response-matrix implication follows from the singular-subspace perturbation bound in Appendix D.

E.4 PARTIAL AND NOISY OBSERVATIONS

Observation model. Let $A \subseteq \{1, \ldots, d-1\}$ be edge-separating and suppose the n observation-state identifiers in $\mathcal{X}$ are known. For every $k \in A$ and $t \in [m_k]$, independently draw $S_{k,t}$ uniformly from $\{S \subseteq [d] : |S| \leq k\}$ and let the latent task be $F_{k,t} = \chi_{S_{k,t}}$.

For every (k, t), independently across tasks draw a random mask $M_{k,t} \subseteq \mathcal{X}$, independent of the task values, but allow arbitrary dependence and nonuniformity among states within the same mask. Assume that known constants $q_k \in (0, 1]$ satisfy

$$\Pr(x, y \in M_{k,t}) \geq q_k \qquad (x \neq y,\ t \in [m_k]). \tag{25}$$

When $x \in M_{k,t}$, the revealed response is

$$Y_{k,t,x} = F_{k,t}(\psi^{-1}x) + \xi_{k,t,x},$$

where the noise vectors are independent across tasks conditional on the tasks and masks, may be heteroskedastic and dependent across states within one task, and satisfy, for $x \neq y$,

$$\begin{aligned} \mathbb{E}[\xi_{k,t,x} \mid F_{k,t}, M_{k,t}] &= 0, \\ \mathbb{E}[\xi_{k,t,x}\xi_{k,t,y} \mid F_{k,t}, M_{k,t}] &= 0, \\ |\xi_{k,t,x}| &\leq b. \end{aligned} \tag{26}$$

For $x \neq y$, let $N_{k,xy}$ be the number of level-k tasks observed at both states. When $N_{k,xy} > 0$, define

$$\widehat{K}_k(x,y) = \frac{D_k}{N_{k,xy}} \sum_{t:\, x,y \text{ observed}} Y_{k,t,x}Y_{k,t,y}.$$

Set $\widehat{K}_k(x,y) = 0$ when $N_{k,xy} = 0$; the theorem's success event excludes this convention. For a confidence level $\delta \in (0,1)$, set

$$L = \log \frac{4|A|n^2}{\delta}, \qquad c_b = (1+b)^2.$$

Theorem 11 (Partial-response graph recovery)**.** *If, for every $k \in A$,*

$$m_k q_k \geq L \max\left\{8, \frac{64 c_b^2 D_k^2}{\Delta_A^2}\right\}, \tag{27}$$

then, with probability at least $1-\delta$, every $N_{k,xy}$ is positive and

$$\max_{k\in A} \max_{x\neq y} |\widehat{K}_k(x,y) - K_k(\psi^{-1}x, \psi^{-1}y)| \leq \frac{\Delta_A}{4}.$$

On this event, declaring x, y adjacent exactly when

$$\max_{k\in A} |\widehat{K}_k(x,y) - \kappa_k(1)| < \frac{\Delta_A}{2}$$

recovers the latent Hamming graph and hence the coordinates up to $\mathcal{G}_d$.

Proof. Fix k, x, y. The joint observation count is $N_{k,xy} = \sum_t \mathbf{1}\{x, y \in M_{k,t}\}$, a sum of independent Bernoulli variables with mean at least $m_k q_k$. The multiplicative Chernoff bound gives

$$\Pr\left(N_{k,xy} < \tfrac{1}{2} m_k q_k\right) \leq e^{-m_k q_k/8} \leq e^{-L}.$$

Condition on all masks satisfying the complementary event. The co-observed products are independent, bounded in absolute value by c_b, and have mean

$$\begin{aligned} \mathbb{E}[Y_{k,t,x}Y_{k,t,y}] &= \frac{1}{D_k} \sum_{|S|\leq k} \chi_S(\psi^{-1}x)\chi_S(\psi^{-1}y) \\ &= \frac{1}{D_k} K_k(\psi^{-1}x, \psi^{-1}y). \end{aligned}$$

Hoeffding's inequality and equation 27 therefore give

$$\begin{aligned} &\Pr\left(|\widehat{K}_k(x,y) - K_k(\psi^{-1}x, \psi^{-1}y)| > \frac{\Delta_A}{4} \;\middle|\; \text{masks}\right) \\ &\leq 2\exp\left(-\frac{N_{k,xy}\Delta_A^2}{32 c_b^2 D_k^2}\right) \leq 2e^{-L}. \end{aligned}$$

A union bound over fewer than $|A|n^2$ level pair choices makes the combined mask and estimation failure probability at most $3|A|n^2 e^{-L} < \delta$. For a true edge all estimated gaps from the edge-signature are at most $\Delta_A/4$. For a nonedge, one exact gap is at least Δ_A, so its estimated gap is at least $3\Delta_A/4$. The threshold therefore reconstructs exact adjacency. Coordinate labeling is Theorem 2. □

No task is required to contain a complete truth table. If $s_{k,t} = |M_{k,t}|$, the data contain $Q = \sum_{k,t} s_{k,t}$ scalar responses. Accumulating all co-observed products and labeling the recovered graph costs

$$O\left(\sum_{k,t} s_{k,t}^2 + |A|n^2 + d^2 n\right)$$

arithmetic operations.

MCAR specialization. Suppose each state is revealed independently with probability $p_k \in (0, 1]$ in every level-k task. Then $q_k = p_k^2$, and Theorem 11 gives the sufficient integer task budget

$$m_k \geq \left\lceil \frac{L}{p_k^2} \max\left\{8, \frac{64(1+b)^4 D_k^2}{\Delta_A^2}\right\}\right\rceil \qquad (k \in A). \tag{28}$$

Moreover,

$$\mathbb{E}Q = n \sum_{k \in A} m_k p_k,$$
$$\mathbb{E}\sum_{k,t} s_{k,t}^2 = \sum_k m_k \{np_k(1-p_k) + n^2 p_k^2\}.$$

For the cubic support and $d \geq 8$, Proposition 2 gives $D_3 = \Theta(d^3)$ and $\Delta_{\{3\}} = \Theta(d^2)$. Consequently, one may take

$$m_3 q_3 = O\left((1+b)^4 d^2 \big(d + \log(1/\delta)\big)\right).$$

In the MCAR case $q_3 = p_3^2$. Consequently, a constant-observation rate and fixed confidence under the general bounded-noise guarantee uses $O(d^3)$ partially observed task functions, while retaining exact-recovery of all 2^d state coordinates.

E.5 Variance-sensitive recovery

Theorem 12 (Variance-sensitive noiseless cubic recovery)**.** *Let $d \geq 8$. Draw S_t uniformly from $\{S : |S| \leq 3\}$ and independent Rademacher signs ε_t, but do not reveal S_t or its degree. Let $F_t = \varepsilon_t \chi_{S_t}$, let the responses be noiseless, and let task masks be independent across t and independent of the task draws, with $q \in (0, 1]$ and $\Pr(x, y \in M_t) \geq q$ for every $x \neq y$. Fix $\delta \in (0, 1)$. Put $D = D_3$, $\Delta = (d-2)(d-3)$, and estimate $\theta_r = \kappa_3(r)/D$ by the co-observed product average $\widehat{\theta}(x, y)$. If*

$$mq \geq 128 d \log \frac{4n^2}{\delta}, \tag{29}$$

then, with probability at least $1 - \delta$, the rule

$$x \sim y \quad \Longleftrightarrow \quad \widehat{\theta}(x, y) > \theta_1 - \frac{\Delta}{2D}$$

recovers every Hamming edge. Under MCAR, $q = p^2$; at constant p and fixed confidence the task count is $O(d^2)$ because $\log n = \Theta(d)$.

Proof. Let $g_r = \kappa_3(1) - \kappa_3(r)$. Proposition 2 gives $g_r \geq \Delta$ for $r \geq 2$. A product $W_t = F_t(\psi^{-1}x)F_t(\psi^{-1}y)$ is in $\{-1, 1\}$, has mean θ_r, and has variance $v_r = 1 - \theta_r^2$. Relabel the N co-observed tasks as $t = 1, \ldots, N$ and define $\overline{W} = N^{-1}\sum_{t=1}^N W_t = \widehat{\theta}(x, y)$. The one-sided Bernstein bound is

$$\begin{aligned} &\Pr(\overline{W} - \theta_r \geq t) \quad \text{or} \quad \Pr(\overline{W} - \theta_r \leq -t) \\ &\qquad \leq \exp\left[-\frac{Nt^2}{2(v_r + 2t/3)}\right]. \end{aligned} \tag{30}$$

For an edge take $t_1 = \Delta/(2D)$; at distance $r \geq 2$ take $t_r = (g_r - \Delta/2)/D \geq g_r/(2D)$.

Here

$$D = \frac{d^3 + 5d + 6}{6}, \qquad a := D - \kappa_3(1) = d^2 - d + 2.$$

For $d \geq 8$, direct polynomial comparison gives

$$a \leq 2\Delta, \qquad D \leq d\Delta, \qquad 1 - \frac{a + \Delta/2}{D} \geq \frac{1}{5}.$$

For an edge, $v_1 \leq 2a/D \leq 4\Delta/D$, and therefore

$$\frac{2v_1}{t_1^2} + \frac{4}{3t_1} \leq 35d.$$

For a nonedge with $\theta_r \geq 0$, $v_r \leq 2(1 - \theta_r) = 2(a + g_r)/D \leq 6g_r/D$, so

$$\frac{2v_r}{t_r^2} + \frac{4}{3t_r} \leq 51d.$$

If $\theta_r < 0$, then $g_r > D - a$ and hence $t_r > 1/5$; the same expression is at most $50 + 20/3 < 51d$. Hence every one-sided error probability in equation 30 is at most $e^{-N/(51d)}$.

We now let $L = \log(4n^2/\delta)$. Chernoff gives $N_{xy} \geq mq/2 \geq 64dL$ for every pair except on an event of probability at most $n^2 e^{-mq/8}$. Conditional on masks satisfying these count bounds, every classification error has probability at most $e^{-64L/51} < e^{-L}$. A union bound over the mask and classification events for fewer than n^2 pairs is below δ. □

E.6 State-linear recovery

The decoder in this subsection uses atomic tasks $F_t = \varepsilon_t \chi_{S_t}$, whose responses form an affine binary code. For arbitrary real linear combinations of characters, Theorem 2 provides recovery through the projection kernels.

Complete-response model. Let $d \geq 8$ and $\delta \in (0, 1/4)$. Draw the noiseless signed-cubic tasks of Theorem 12, but observe every task at all $n = 2^d$ states. Neither S_t nor ε_t is revealed. Put

$$D = D_3 = \frac{d^3 + 5d + 6}{6}, \qquad \Delta = (d-2)(d-3),$$
$$u_1 = \frac{d^2 - d + 2}{2D},$$

and set the rational threshold

$$\alpha = u_1 + \frac{\Delta}{4D} = \frac{3d^2 - 7d + 10}{4D}. \tag{31}$$

Theorem 13 (State-linear complete cubic recovery)**.** *If*

$$m \geq 4096d \log \frac{d}{\delta}, \tag{32}$$

then, with probability at least $1 - \delta$*, the procedure below recovers the latent coordinates up to* $\mathcal{G}_d$*.*

Recovery procedure. Encode the response of state x as $c(x) \in \mathbb{F}_2^m$, with $c_t(x) = \mathbf{1}\{F_t(\psi^{-1}x) = -1\}$. Choose any root o and scan the other states once. Its d latent neighbors are exactly those $x \neq o$ satisfying

$$\frac{1}{m} \operatorname{wt}\big(c(x) + c(o)\big) < \alpha. \tag{33}$$

Order these neighbors as $o_1, \ldots, o_d$ and form the $m \times d$ binary matrix whose ith column is $c(o_i) + c(o)$. It has full column rank. For every state x, solve

$$c(x) + c(o) = \sum_{i=1}^{d} b_i(x)\big(c(o_i) + c(o)\big) \quad \text{over } \mathbb{F}_2. \tag{34}$$

Then $x \mapsto ((-1)^{b_1(x)}, \ldots, (-1)^{b_d(x)})$ is a valid recovered representation, up to the output sign change determined by the root.

The full input consists of mn response bits. Naive binary Gaussian elimination and application of a left inverse cost

$$O(mdn + md^2)$$

bit operations, and no all-pairs matrix is constructed. At the certified choice $m = \lceil 4096d\log(d/\delta)\rceil$, the input size is $O(nd\log(d/\delta))$ and the runtime is $O(nd^2\log(d/\delta)+d^3\log(d/\delta))$. The comparison in equation 33 can be evaluated in exact integer arithmetic as

$$4D\ \mathrm{wt}(c(x)+c(o)) < m(3d^2-7d+10).$$

Proof. Write a latent state as $z_i = (-1)^{b_i(z)}$ and let $R \in \mathbb{F}_2^{m\times d}$ have rows $\mathbf{1}_{S_t}$. If $e_t = \mathbf{1}\{\varepsilon_t = -1\}$, then the complete-response code obeys

$$c(\psi(z)) = e + Rb(z). \tag{35}$$

Therefore, for a difference vector $v \in \mathbb{F}_2^d$ of Hamming weight r,

$$\widehat{u}(v) := m^{-1}\,\mathrm{wt}(Rv)$$

is an average of independent Bernoulli variables with mean

$$u_r = \frac{D-\kappa_3(r)}{2D} = \frac{r\{1+(d-r)+\binom{d-r}{2}\}+\binom{r}{3}}{D}. \tag{36}$$

The last expression counts subsets S of size at most three having odd intersection with a fixed r set. Since $u_2 - u_1 = \Delta/(2D)$, the threshold α is their midpoint.

We begin by proving the simultaneous separation event

$$|v| = 1 \Longrightarrow \widehat{u}(v) < \alpha, \qquad |v| \geq 2 \Longrightarrow \widehat{u}(v) > \alpha. \tag{37}$$

The Bernstein calculations in the proof of Theorem 12, with $q = 1$, show that each one-sided error in equation 37 has probability at most $\exp[-m/(51d)]$. If $8 \leq d < 640$, a union bound over the fewer than 2^d nonzero difference vectors and equation 32 is below δ.

For $d \geq 640$, stratification removes the exponential union penalty from the hard, low-weight differences. Directly from the displayed formulas,

$$D \leq \frac{d^3}{3}, \qquad \alpha \leq \frac{5}{d}.$$

For $36 \leq r \leq d/2$, the first term in equation 36 gives

$$u_r \geq \frac{r\binom{d-r}{2}}{D} \geq \frac{9r}{32d} \geq 2\alpha.$$

For $r \geq d/2$, its last term gives $u_r \geq \binom{r}{3}/D \geq 1/64 \geq 2\alpha$. The multiplicative Chernoff bound therefore gives

$$\Pr(\widehat{u}(v) \leq \alpha) \leq \begin{cases} \exp[-9mr/(256d)], & 36 \leq r \leq d/2, \\ \exp[-m/512], & r \geq d/2. \end{cases}$$

Together with the Bernstein bound for weights $1 \leq r \leq 35$, the failure probability is at most

$$36d^{35}e^{-m/(51d)} + \sum_{r=36}^{\lfloor d/2\rfloor}\left(\frac{ed}{r}\right)^r e^{-9mr/(256d)} + 2^d e^{-m/512}.$$

Substitution of equation 32 makes the first and third terms at most $\delta/3$ because their exponents dominate $35\log d + \log(108/\delta)$ and $d\log 2 + \log(3/\delta)$, respectively. For every middle summand, the exponent dominates $r\log(ed/r) + \log(3d/\delta)$, so each is at most $\delta/(3d)$; there are fewer than d such terms. Hence the displayed failure probability is below δ, establishing equation 37.

On this event, equation 35 makes the one-pass root test exactly a test of whether the latent difference has weight one, so it returns the d neighbors. Their code differences are the columns of R, in some order. These columns are linearly independent. Indeed, a kernel vector of weight at least two contradicts the second implication in equation 37. If a column were zero, adding its unit vector to any other unit vector would give a nonedge with the same syndrome as an edge, contradicting both implications. The identity in equation 35 now proves that equation 34 has the unique solution equal to the latent state relative to the root and the chosen ordering of coordinate directions. The input and operation counts follow from one scan, one $m \times d$ elimination, and n applications of its left inverse. □

E.7 LOWER BOUNDS

The logarithmic task factor above is necessary for global coordinate recovery itself.

Proposition 3 (End-to-end complete cubic task lower bound)**.** *In the complete noiseless task model of Theorem 13, let an arbitrary, possibly randomized procedure be required to recover coordinates up to $\mathcal{G}_d$ with probability at least $1-\delta$ for every observation bijection, where $d \geq 8$ and $\delta \in (0, 1/4)$. Necessarily,*

$$m \geq \frac{\log(d(1-2\delta)/(2\delta))}{-\log(1-s_d)}, \qquad s_d = \Pr(i \in S_t) = \frac{d^2-d+2}{2D}. \tag{38}$$

The lower bound holds even if every S_t and ε_t is revealed to the learner. Since $s_d = 3/d + O(d^{-2})$, it is $\Omega(d\log(d/\delta))$. Hence Theorem 13 gives the matching $\Theta(d\log(d/\delta))$ end-to-end task complexity, up to universal constants, for this complete noiseless atomic cubic model.

Proof. Choose vertex-disjoint cube edges $e_1, \ldots, e_d$, with e_i in coordinate direction i, and let τ_i transpose the endpoints of e_i. Such edges can be chosen greedily. Before choosing direction i, at most $2(i-1)$ edges of its perfect matching meet a used vertex, whereas $2^{d-1} > 2(d-1)$ for $d \geq 5$. For $\vartheta \in \mathbb{F}_2^d$, define

$$\psi_\vartheta = \prod_{i=1}^{d} \tau_i^{\vartheta_i}.$$

These observation instances have pairwise disjoint correct representation classes. Indeed, the relative permutation for two different parameters is a nonempty product of at most d selected transpositions. It fixes a vertex because $2d < 2^d$ and sends another vertex to an adjacent one. A cube automorphism fixing a vertex preserves distance parity from that vertex and cannot have the latter property, so the relative permutation is not in $\mathcal{G}_d$.

Put a uniform prior on ϑ. Conditional on the task subsets, let U be the number of coordinates absent from every S_t. If coordinate i is absent, every complete task truth table is invariant under τ_i; hence the entire data set is independent of the bit ϑ_i. Even after revealing the subsets and signs, conditional Bayes success is therefore at most 2^{-U}. Uniform success $1-\delta$ would imply

$$1-\delta \leq \mathbb{E}2^{-U} \leq 1 - \tfrac{1}{2}\Pr(U>0), \qquad \Pr(U>0) \leq 2\delta.$$

For a fixed coordinate, the probability of being absent from all tasks is $(1-s_d)^m$. For two coordinates, the one-task probability that both occur is $(d-1)/D \leq s_d^2$; the inequality is equivalent to

$$d^4 - 4d^3 + 5d^2 - 14d + 24 \geq 0,$$

which holds for $d \geq 5$. Hence the two absence indicators are negatively correlated. Taking expectations gives

$$\mathbb{E}U = d(1-s_d)^m, \qquad \mathbb{E}U^2 \leq \mathbb{E}U + (\mathbb{E}U)^2,$$

and the second-moment inequality gives $\Pr(U>0) \geq \mathbb{E}U/(1+\mathbb{E}U)$. Combining this with $\Pr(U>0) \leq 2\delta$ yields

$$d(1-s_d)^m \leq \frac{2\delta}{1-2\delta},$$

which is exactly equation 38. Finally, $-\log(1-s_d) \leq s_d/(1-s_d)$ and $s_d = 3/d + O(d^{-2})$ give the stated order. □

Proposition 4 (Hard-pair co-observation lower bound)**.** *In the noiseless signed-cubic model of Theorem 12, suppose two named states are promised to have latent distance either one or two and each state is independently revealed with probability p in each of m tasks. Any test based on these masked two-state responses whose error is at most $\delta < 1/4$ under both hypotheses must satisfy*

$$mp^2 \geq \frac{d}{6}\log\frac{1}{4\delta}.$$

Thus the p^{-2} dependence and the $\Theta(d)$ hard-pair co-observation scale in Theorem 12 are information-theoretically necessary for the pair-signature experiment; the additional $\log n$ in equation 29 enforces simultaneous correctness over all state pairs.

Proof. Choose one latent endpoint to be $\mathbf{1}$. The independent task sign makes every singly observed response uniform under both hypotheses. When we observe both states, all information is in their product W. At latent distance r, the probability that $W = -1$ is $u_r = (1 - \kappa_3(r)/D)/2$. It follows that

$$u_2 - u_1 = \frac{\Delta}{2D}.$$

For $d \geq 8$, $\kappa_3(2) = \binom{d-1}{3} - \Delta > 0$, so $u_2 < 1/2$. The chi-square divergence bounds the Bernoulli KL divergence by

$$\begin{aligned}\mathrm{KL}(\mathrm{Ber}(u_1) \| \mathrm{Ber}(u_2)) &\leq \frac{(u_2 - u_1)^2}{u_2(1 - u_2)} \\ &\leq 2(u_2 - u_1) = \frac{\Delta}{D} \leq \frac{6}{d}.\end{aligned}$$

Only the mask event revealing both states contributes KL, so the full m task divergence is at most $6mp^2/d$. The Bretagnolle–Huber inequality (Bretagnolle & Huber, 1979) lower bounds the sum of the two testing errors by one half the exponential of minus this divergence. If both errors are at most δ, rearranging gives the displayed bound. □

Proposition 5 (Sharp state-coverage obstruction)**.** *Fix $d \geq 3$. Suppose a possibly adaptive randomized procedure queries at most s distinct observation states almost surely and recovers the representation up to $\mathcal{G}_d$ with probability at least $1 - \delta$ for every observation bijection. Then*

$$\binom{n - s}{2} \leq 2\delta \binom{n}{2}. \tag{39}$$

Thus zero-error recovery requires at least $n - 1$ queried states; in particular, a deterministic procedure querying at most $n - 2$ states has two indistinguishable instances with disjoint correct representation classes.

Proof. For the deterministic claim, run the procedure on any instance and let $a, b \in \mathcal{X}$ be two states outside its queried set. Let $\tau_{\mathcal{X}}$ transpose a, b and fix every other observation-state, and replace ψ by $\psi' = \tau_{\mathcal{X}} \circ \psi$, leaving every latent task unchanged. At every queried state x, one has $\psi'^{-1}x = \psi^{-1}x$, so induction over the adaptive queries shows the entire transcript is identical.

The two correct output classes cannot intersect. Indeed, an intersection would give $S, S' \in \mathcal{G}_d$ with $S \circ \psi^{-1} = S' \circ \psi'^{-1}$. After composition with ψ, this says that the latent transposition

$$\psi^{-1} \circ \tau_{\mathcal{X}} \circ \psi = S'^{-1} S$$

belongs to $\mathcal{G}_d$. However, a permutation fixing all but two cube vertices is not a Hamming graph automorphism when $d \geq 3$, by Lemma 3. Hence the classes are disjoint, and one of the two identical-transcript instances defeats the deterministic procedure.

For the randomized statement, fix a base instance and couple its task draws and internal randomness with every swapped instance. For an unordered pair $\{a, b\}$, let E_{ab} be the event that neither state is queried on the base instance. On E_{ab}, the coupled adaptive transcripts on the base and $\{a, b\}$ swapped instances remain identical. Their correct classes are disjoint by the preceding argument, so the sum of their two error probabilities is at least $\Pr(E_{ab})$. Uniform error at most δ therefore implies $\Pr(E_{ab}) \leq 2\delta$ for every pair. If U denotes the number of unqueried states on the base instance, summing over pairs gives

$$\mathbb{E}\binom{U}{2} = \sum_{\{a,b\}} \Pr(E_{ab}) \leq 2\delta \binom{n}{2}.$$

The almost-sure bound of s queries gives $U \geq n - s$ and proves equation 39. Conversely, averaging the events E_{ab} shows for some swapped pair their common-transcript probability is at least $\binom{n-s}{2}/\binom{n}{2}$; one of those two instances has at least half that error. This proves the equivalent success bound. □

E.8 Bounds on the leakage gap

Theorem 8 bounds the leakage of every non-coordinate map from below by $\beta_{A,\lambda}/n$. A single adjacent transposition gives an upper bound with the same explicit n^{-1} factor. We retain the dependence on dimension, degrees, and weights in the bracket below.

Let τ exchange two adjacent states u and v, let J be its permutation matrix, and write $\delta = e_u - e_v$ in the point-mass basis, so $J = I - \delta\delta^\top$ and $Q_{\tau,k} = JP_kJ$. Set

$$M_k = \sum_{j=0}^{k-1} \binom{d-1}{j}, \qquad a_k = \delta^\top P_k \delta.$$

A character χ_S with $|S| \le k$ takes different values at u and v exactly when S contains the coordinate in which they differ, which happens for M_k sets, and each contributes $4/n$; hence $a_k = 4M_k/n$. From $\operatorname{tr}(P_kQ_{\tau,k}) = D_k - 2a_k + a_k^2$ we obtain $\|P_k - Q_{\tau,k}\|_{\mathrm{HS}}^2 = 2a_k(2-a_k)$ and therefore

$$\ell_k(\tau) = \frac{a_k(2-a_k)}{D_k} = \frac{8M_k}{nD_k} - \frac{16M_k^2}{n^2 D_k}. \tag{40}$$

Since $M_k \le D_k$ for every k, equation 40 gives $\ell_k(\tau) \le 8M_k/(nD_k) \le 8/n$, and weighting by λ yields $\mathcal{R}_{A,\lambda}(\tau) \le 8\sum_{k\in A}\lambda_k/n$. Combining with Theorem 8,

$$\frac{\beta_{A,\lambda}}{n} \le \min_{T\notin\mathcal{G}_d} \mathcal{R}_{A,\lambda}(T) \le \frac{8\sum_{k\in A}\lambda_k}{n}.$$

The swapped edge is preserved, while the other $d-1$ edges incident to each endpoint become pairs at distance two. All remaining edges are preserved, so $B(\tau) = 2(d-1)$. The transposition differs from the identity at two states and is not a cube automorphism. Since distinct permutations differ at two or more states, $\rho_0(\tau, \mathcal{G}_d) = 2/n$.

The factor $\beta_{A,\lambda}$ depends on dimension through D_k, as well as on degrees and weights. For general supports, the bracket leaves a gap in these parameters. Appendix E.9 resolves the cubic case. We also record that $\mathcal{R}_{A,\lambda}$ is a squared normalized chordal error; a perturbation of size ε measured in the chordal norm itself corresponds to ε^2 here. Taking square roots gives the corresponding bracket in the chordal norm, including the square roots of the dimension, degree, and weight factors.

E.9 Cubic leakage and minimizing transpositions

For the cumulative cubic family, retaining the distance by which each edge is stretched yields a stronger geometric estimate. We combine it with concentration bounds for polynomials of degree at most three to determine the leakage gap. Throughout this subsection, let $d \ge 5$, $n = 2^d$, and $C = C_{\mathrm{FKN}}$ from Lemma 6. Define

$$\begin{aligned} D = D_3 &= \frac{d^3+5d+6}{6}, & a_0 &= D - \kappa_3(1) = d^2 - d + 2, \\ q_d &= \frac{3d^2-18d+23}{12}, & u_d &= \frac{4a_0}{nD}, \\ b_d &= \frac{108\sqrt{Ca_0}}{q_d}, & \phi(b) &= \frac{4}{(\sqrt{b^2+4}+b)^2}. \end{aligned} \tag{41}$$

Here $q_d > 0$ for $d \ge 5$.

Theorem 14 (Cubic leakage gap)**.** *Every cube bijection $T \notin \mathcal{G}_d$ satisfies*

$$\ell_3(T) \ge u_d\phi(b_d). \tag{42}$$

In particular, there are universal constants $c_-, c_+ > 0$ such that, for all $d \ge 5$,

$$\frac{c_-}{nd} \le \min_{T\notin\mathcal{G}_d} \ell_3(T) \le \frac{c_+}{nd}.$$

If $d \ge 8$ and $b_d < 1/3$, then

$$\min_{T\notin\mathcal{G}_d} \ell_3(T) = u_d\left(1 - \frac{a_0}{n}\right). \tag{43}$$

Under these conditions, the minimizers are exactly $S \circ \tau$, where $S \in \mathcal{G}_d$ and τ exchanges two adjacent vertices.

Since $q_d \geq d^2/8$ and $a_0 \leq d^2$ for $d \geq 11$, the explicit conditions $d \geq 11$ and $d > 2592\sqrt{C}$ suffice for the conclusion about exact minimizers.

Proof. **Control by edge stretch.** For $R = T^{-1}$, let

$$W(T) = \sum_{\{x,y\} \in E(\Omega_d)} \big(\mathrm{dist}_H(Rx, Ry) - 1\big).$$

Each summand is nonnegative because R is bijective. The factorization in equation 21, with $s = d - 2r$, gives

$$\kappa_3(1) - \kappa_3(r) = \frac{r-1}{3}\big(s^2 + (d+1)s + d^2 - 4d + 6\big).$$

The quadratic in parentheses has real minimum $(3d^2 - 18d + 23)/4$. For $r \geq 2$, this implies

$$\kappa_3(1) - \kappa_3(r) \geq q_d(r-1), \qquad [\kappa_3(1) - \kappa_3(r)]^2 \geq q_d^2(r-1).$$

Restricting the sum in equation 12 to ordered cube edges yields

$$\ell_3(T) \geq \frac{q_d^2}{Dn^2} W(T). \tag{44}$$

Write $R_i = f_i$. These functions are balanced and pairwise orthogonal, and their total influence excess satisfies

$$\sum_{i=1}^{d} \big(\mathsf{I}(f_i) - 1\big) = \frac{2W(T)}{n}.$$

Lemma 6 gives signed dictators g_i with errors $\delta_i = \Pr[f_i \neq g_i] \leq C(\mathsf{I}(f_i) - 1)$. Put $\delta = \sum_i \delta_i$. If $\delta < 1/2$, all rounded coordinate indices are distinct. A repeated index would give $|\mathbb{E}[g_i g_j]| = 1$, whereas orthogonality and the error bound give

$$|\mathbb{E}[g_i g_j] - \mathbb{E}[f_i f_j]| \leq 2(\delta_i + \delta_j) < 1.$$

The rounded map is therefore in $\mathcal{G}_d$, with disagreement at most δ. If $\delta \geq 1/2$, the bound $\rho_0 \leq 1 \leq 2\delta$ applies. Inversion preserves distance from $\mathcal{G}_d$, so both cases give

$$\rho_0(T, \mathcal{G}_d) \leq \frac{4C}{n} W(T).$$

Combining this with equation 44, we obtain

$$\alpha := \rho_0(T, \mathcal{G}_d) \leq \frac{4CDn}{q_d^2} \ell_3(T). \tag{45}$$

Concentration on the moved states. For a set $V \subseteq \Omega_d$ of density α, let $H = P_3[V, V]$ denote the principal submatrix in orthonormal point-mass coordinates. Since P_3 is an orthogonal projector, $0 \preceq H \preceq I$, and

$$\begin{aligned} h := \|H\|_{\mathrm{op}} &= \sup_{f \in \mathcal{F}^d_{\leq 3},\, \|f\|_2 = 1} \mathbb{E}[\mathbf{1}_V f^2] \\ &\leq \sqrt{\alpha} \sup_{f \in \mathcal{F}^d_{\leq 3},\, \|f\|_2 = 1} \|f\|_4^2 \leq 27\sqrt{\alpha}. \end{aligned} \tag{46}$$

The last inequality is the case of degree at most three in hypercontractivity (O'Donnell, 2014, Theorem 9.21). We include the needed proof. Let N_ρ multiply the Walsh coefficient at S by $\rho^{|S|}$, where $\rho = 1/\sqrt{3}$. Induction on the number of coordinates proves $\|\mathsf{N}_\rho g\|_4 \leq \|g\|_2$. The base case is a constant function. For the induction step, write $g(x', x_d) = v(x') + x_d w(x')$ and set $A = \mathsf{N}_\rho v$, $B = \mathsf{N}_\rho w$ on the first $d-1$ coordinates. Cauchy–Schwarz and the induction hypothesis give

$$\begin{aligned} \|\mathsf{N}_\rho g\|_4^4 &= \mathbb{E}[A^4 + 6\rho^2 A^2 B^2 + \rho^4 B^4] \\ &\leq \|A\|_4^4 + 2\|A\|_4^2 \|B\|_4^2 + \tfrac{1}{9}\|B\|_4^4 \\ &\leq (\|v\|_2^2 + \|w\|_2^2)^2 = \|g\|_2^4. \end{aligned}$$

For $f \in \mathcal{F}^d_{\le 3}$, divide its Walsh coefficients by $\rho^{|S|}$ to write $f = \mathsf{N}_\rho g$. Parseval gives $\|g\|_2 \le \rho^{-3}\|f\|_2 = 3^{3/2}\|f\|_2$, hence $\|f\|_4^2 \le 27\|f\|_2^2$.

Leakage on the moved states. Choose a nearest $S \in \mathcal{G}_d$ and put $J = S^{-1} \circ T$. Its support V has size $m = \alpha n$. Coordinate maps commute with P_3, so $\ell_3(J) = \ell_3(T)$. Because $T \notin \mathcal{G}_d$, the permutation J is nonidentity and $m \ge 2$.

Let Q be the permutation matrix of J on V, and put $\Delta = Q - I$. With $H = P_3[V, V]$, expansion of the projection trace gives

$$\begin{aligned} D\ell_3(J) &= -2\,\mathrm{tr}(\Delta H) - \mathrm{tr}(\Delta H \Delta^\top H) \\ &= A_0 - B_0. \end{aligned} \tag{47}$$

To verify the expansion, extend Q by the identity outside V and use $D\ell_3(J) = D - \mathrm{tr}(P_3 Q P_3 Q^\top)$. The linear terms both equal $\mathrm{tr}(\Delta H)$ because $P_3^2 = P_3$; the quadratic term is the second trace in equation 47. The identity $\Delta^\top \Delta = 2I - Q - Q^\top$ yields

$$A_0 = \mathrm{tr}(\Delta H \Delta^\top) \ge 0, \qquad 0 \le B_0 \le h A_0.$$

Every state in V moves, and the cubic kernel has its strictly largest value among distinct state pairs at distance one. It follows that

$$A_0 = \frac{2}{n} \sum_{x \in V} [D - \kappa_3(\mathrm{dist}_H(x, Jx))] \ge \frac{2m a_0}{n}.$$

Since $h \le 1$ and $m \ge 2$, we have

$$\ell_3(T) \ge \frac{2m a_0}{nD}(1 - h) \ge u_d(1 - h). \tag{48}$$

A uniform lower bound. Put $t = \ell_3(T)/u_d$. Equations 45 and 46 imply

$$h \le 27\sqrt{\frac{4CDn}{q_d^2} u_d t} = b_d \sqrt{t}.$$

Equation 48 gives $t \ge 1 - b_d\sqrt{t}$. Solving the quadratic inequality in $\sqrt{t}$ yields

$$t \ge \left(\frac{\sqrt{b_d^2 + 4} - b_d}{2} \right)^2 = \phi(b_d),$$

which proves equation 42. For all $d \ge 5$, direct polynomial comparison gives

$$q_d \ge \frac{2d^2}{75}, \qquad a_0 \le d^2, \qquad 16 \le nd\,u_d \le 24.$$

Thus $b_d \le 810\sqrt{C}$. Since ϕ is decreasing, the lower bound holds with the universal constant $c_- = 16\phi(810\sqrt{C})$. For an adjacent transposition, $M_3 = a_0/2$ in equation 40, so its leakage is $u_d(1 - a_0/n) \le 24/(nd)$. This gives $c_+ = 24$. The explicit lower bound is $u_d[1 - O(d^{-1})]$ as d grows, because C is universal and $b_d = O(d^{-1})$.

Exact minimizing maps. Assume $d \ge 8$ and $b_d < 1/3$, and suppose T has leakage at most that of an adjacent transposition. Then $t \le 1$ and $h \le b_d < 1/3$. Keeping m in equation 48 gives

$$\frac{m}{2}(1 - h) \le t \le 1, \qquad m \le \frac{2}{1 - h} < 3.$$

The integer m therefore equals two, so J is a single transposition. For a transposition at distance r, put $a(r) = 2[D - \kappa_3(r)]/n$. The same trace calculation gives

$$\ell_3(\tau_r) = \frac{a(r)[2 - a(r)]}{D}.$$

The cumulative-kernel identity expresses $\kappa_3(r)$ as a sum of $\binom{d-1}{3}$ products of three signs, so $|\kappa_3(r)| \le \binom{d-1}{3}$ for $r \ge 1$. Consequently,

$$a(r) \le \frac{2[D + \binom{d-1}{3}]}{n} = \frac{2d(d^2 - 3d + 8)}{3\,2^d} \le 1 \qquad (d \ge 8).$$

For the last inequality there is equality at $d = 8$. Writing $p(d) = 2d(d^2 - 3d + 8)/3$, the identity $2p(d) - p(d+1) = 2(d-1)(d-2)(d-3)/3 > 0$ propagates it by induction. The function $a(2-a)$ is strictly increasing for $0 < a \leq 1$. Since $\kappa_3(r) < \kappa_3(1)$ for $r \geq 2$, adjacent transpositions uniquely minimize leakage among transpositions. Coordinate composition preserves leakage, which proves equation 43 and the stated minimizer class. □

Multiplication by a fixed $\lambda_3 > 0$ gives the corresponding result for $\mathcal{R}_{\{3\},\lambda}$. The condition for exact minimizers matters in small dimensions. Substitution into the transposition formula shows that antipodal swaps have lower leakage than adjacent swaps at $d = 5, 6$; they tie at $d = 7$. The theorem gives a uniform gap bound for all $d \geq 5$ and classifies all minimizers under its stated sufficient condition.

F Automorphisms of Hamming Graphs with q Symbols

The cumulative ANOVA space in Section 6 has dimension

$$D_{q,k} = \dim \mathcal{F}^{q,d}_{\leq k} = \sum_{\ell=0}^{k} \binom{d}{\ell} (q-1)^{\ell}.$$

The one-coordinate mean-zero kernel is $q\mathbf{1}\{a = b\} - 1$. Thus its cumulative kernel depends only on $r = \mathrm{dist}_H(x, y)$ and is given by

$$\kappa_{q,k}(r) = \sum_{\ell=0}^{k} \sum_{a=0}^{\ell} \binom{r}{a} \binom{d-r}{\ell-a} (-1)^a (q-1)^{\ell-a}. \tag{49}$$

For $d \geq 2$ and $q \geq 3$, the following classification is given by Mirafzal & Ziaee (2021).

Lemma 8. *Every automorphism A of the Hamming graph $H(d, q)$ has the form*

$$A_i(x) = \tau_i\big(x_{\pi(i)}\big)$$

for a coordinate permutation $\pi \in S_d$ and symbol permutations $\tau_i \in S_q$.

Proof. The assertion is immediate for $d = 1$, so assume $d \geq 2$. Choose a reference symbol $0 \in [q]$ and let $\mathbf{0} = (0, \ldots, 0)$. Independent output-symbol permutations can first move $A(\mathbf{0})$ to $\mathbf{0}$, so it is enough to classify automorphisms fixing $\mathbf{0}$.

The neighbors of $\mathbf{0}$ split into d classes

$$C_i = \{x : \mathrm{supp}(x) = \{i\}\}, \qquad |C_i| = q - 1.$$

Two such neighbors are adjacent exactly when they belong to the same C_i. Hence an automorphism fixing $\mathbf{0}$ permutes the classes C_i (for $q = 2$, the classes are singletons and the same conclusion is simply a permutation of the d neighbors). Compose with a coordinate permutation and with symbol permutations fixing 0 so that the resulting automorphism B fixes $\mathbf{0}$ and every one-coordinate vertex.

We prove by induction on Hamming weight that B fixes every vertex. The weights zero and one are already fixed. Let x have support S of size $m \geq 2$, and for each $i \in S$ let $x^{(i)}$ be obtained by replacing x_i with 0. By induction all $x^{(i)}$ are fixed. The vertex x has weight m and is adjacent to every $x^{(i)}$. It is the unique vertex with these properties. For $m = 2$, the two predecessors have exactly two common neighbors, namely $\mathbf{0}$ and x, and the weight selects x. For $m \geq 3$, adjacency to any two predecessors forces agreement with x on their union of nonzero coordinates, and the remaining predecessors force every coordinate as well. Since graph automorphisms preserve distance from $\mathbf{0}$, $B(x)$ also has weight m and is adjacent to all $x^{(i)}$; hence $B(x) = x$.

Undoing the normalizing compositions yields a global coordinate permutation and an independent symbol permutation for each output coordinate. Every such map preserves Hamming distance, which proves the converse. □

Proof of Theorem 3. Choose $k \in A$. The generating polynomial for the pure ANOVA kernels is

$$(1 + (q-1)t)^{d-r}(1-t)^r,$$

so cumulative coefficient extraction gives

$$\kappa_{q,k}(r) = [t^k](1 + (q-1)t)^{d-r}(1-t)^{r-1}.$$

At $r = 1$ this equals $\binom{d-1}{k}(q-1)^k$. For $r \geq 2$, the coefficient sums all products of k distinct entries from a list containing $d-r$ copies of $q-1$ and $r-1$ copies of -1. Each product has absolute value at most $(q-1)^k$. Since $1 \leq k \leq d-1$, at least one product includes a -1 entry and has absolute value at most $(q-1)^{k-1} < (q-1)^k$. The sum is therefore strictly below $\binom{d-1}{k}(q-1)^k$, which proves $\kappa_{q,k}(r) < \kappa_{q,k}(1)$.

Theorem 4 therefore forces every flag preserver to preserve Hamming adjacency, and Lemma 8 identifies it as a coordinate and symbol permutation. Such maps preserve every ANOVA flag. If A is empty, an arbitrary non-graph permutation supplies the non-coordinate witness. □

G Experimental Protocols and Additional Results

This appendix gives the protocols, metrics, and additional results for Section 7. Intervals are means plus or minus 1.96 standard errors across runs with the stated seeds. Original and independent seed groups are reported separately. Frozen readout seeds resample tasks for each fixed encoder.

G.1 Independent arithmetic and implementation checks

The integer recurrence and an independent Walsh implementation check the kernel values. The latter enumerates characters χ_S and evaluates sampled coefficient tables; an equivalent fast Walsh transform accelerates the larger runs. The finite Walsh scan covers dimensions 6 through 12, and exact collision checks extend to $d = 80$.

G.2 Calibrated distances and complete response reconstruction

Tasks, supports, and budgets. For each cumulative family, independent standard Gaussian Walsh coefficients including the constant generate tasks scaled by $D_k^{-1/2}$. Observation identifiers are randomly permuted. Active families split the total task budget as evenly as possible, assigning any remainder to the first levels. Shared levels reuse coefficient tables across supports. The supports are $\{3,5\}$, $\{1\}$, $\{2,4\}$, and A_d, where $A_7 = \{2,3,4\}$ and $A_{11} = \{2,4,5\}$. Both A_d are identifying. The budget sweep uses 512, 2048, and 8192 tasks, all with Gaussian response noise of standard deviation 0.25. Seeds 2000 through 2029 define the original group. The method comparisons additionally use independent seeds 3000 through 3029.

Pair sampling and metrics. Canonical unordered pairs are sampled without replacement, first for calibration and then from the remaining pool for evaluation. Target counts are 24 and 48 per distance class. If a class has $N < 72$ available pairs, we use $c = \min(24, \max(1, \lfloor N/3 \rfloor))$ calibration pairs and $t = \min(48, N - c)$ test pairs; $N < 2$ raises an error. For example, the seven-dimensional antipodal class has 64 pairs, giving 21 calibration and 43 test pairs before state masking. Calibration and evaluation pairs are disjoint, although the two sets may share individual states.

The decoder predicts one of the d nonzero Hamming distances. Micro accuracy weights each test pair equally; macro accuracy weights each distance class equally. Witness accuracy restricts the same d-class predictor to pairs whose true distances are 2 and $d-1$. Its uniform randomized reference is $1/d$, even though the evaluation subset contains two classes. These are supervised calibration experiments on a fixed state space.

Distance decoders. For each active level, the mean product of the observed task responses at the two endpoints gives one feature. Nearest centroid uses class means and Euclidean distance. The robust rule uses class medians and a pooled residual covariance C, regularized as $0.85C + (0.15\,\mathrm{tr}(C)/|A| + 10^{-8})I$. We also evaluate five nearest neighbors, multinomial logistic regression

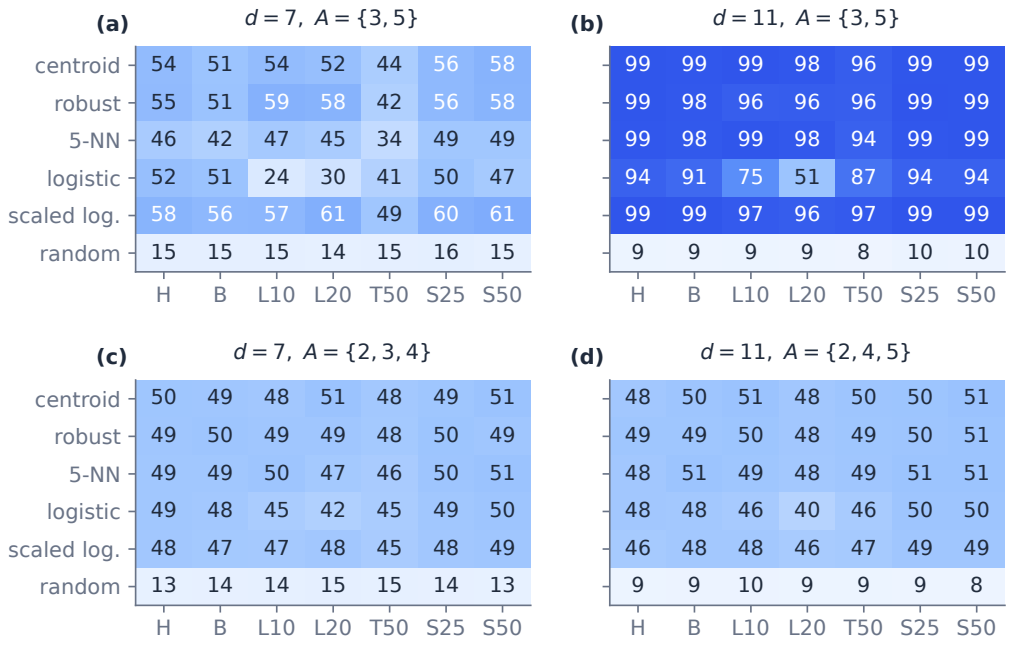


Figure 4: Witness multiclass accuracy (%) under all seven perturbations, averaged over 30 original seeds. Both displayed supports identify coordinates. H and B change response noise and bias; L10/L20 corrupt calibration labels; T50 masks task products per pair; S25/S50 evaluate pairs of visible states. The scaled logistic row fits its transformation on calibration data only. All rows retain the full d-class output space, with random reference $1/d$.

with $C = 1$ and 500 maximum iterations, and uniform random predictions. The scaled logistic variant fits feature means and standard deviations using calibration data only, then uses the same regression parameters. Test labels never select scaling or hyperparameters.

Perturbations. All seven perturbation conditions use 2048 tasks with baseline response noise $\sigma = 0.25$. H multiplies each task's noise scale by $\exp(0.7Z - 0.245)$, with Z standard normal. B adds a Gaussian task bias of standard deviation 0.35 shared by the two endpoints. Noise scales and biases are sampled per processing batch of at most 128 pairs. L10 and L20 independently replace 10% and 20% of calibration labels by a uniformly selected incorrect distance. T50 independently drops half the products of task responses per pair and averages the retained products; an empty mask retains the first product. S25 and S50 independently exclude each state with probabilities 0.25 and 0.50, fix this mask within the condition, and restrict both pair sets to visible states.

Endpoint noise is generated independently for each pair, so observations of a shared state can differ across pairs. S25 and S50 evaluate distance classification on visible pairs with one fixed state mask per condition; Theorem 11 samples masks independently across tasks.

Scaling comparison and distance collisions. Across two dimensions, two supports, and ten conditions, scaled logistic regression improves macro accuracy in all 40 cells in each seed group. On independent seeds the mean paired gain is 7.35 ± 0.17 percentage points, averaging conditions within each outer seed before computing the interval. Witness accuracy improves in 30 cells, declines in nine, and is unchanged in one. For $d = 7$, $\{3, 5\}$, and 8192 tasks, it falls from 86.25% to 82.08%. Scaling nearest centroid decreases macro accuracy in all 40 independent cells, by 3.34 points on average; we retain the unscaled centroid.

Low witness accuracy is compatible with coordinate identification. For $d = 7$, the joint kernels of A_7 at distances 2 and 6 both equal $(5, 0, -5)$. For $d = 11$, those of A_{11} at distances 2 and 10 both equal $(27, 42, 0)$. These supports still distinguish edges from nonedges. The complete graph can determine coordinates even when a local population signature does not distinguish every distance.

Complete response projection. This experiment supplies a coherent response matrix at all 2^7 states, with 8192 tasks and independent Gaussian entry noise $\sigma = 0.25$. For each family k, center the task columns and take the leading $D_k - 1$ left singular vectors as Q_k. Restore the known constant direction to form $\widehat{K}_k = nQ_kQ_k^\top + \mathbf{1}\mathbf{1}^\top$. Use the test at half the edge margin of equation 23, followed by the BFS labeling procedure in Appendix E. The graph must have 2^d vertices, degree d at every vertex, connectivity, unique labels, and agreement between its edges and labeled Hamming neighbors. Failure of any check rejects the reconstruction. Recovery uses the responses, the family indices, and their dimensions. Truncated SVD estimates the task space under noise.

Calibrated MDS/FastICA baseline. The baseline and projector receive the same complete response matrices. The baseline additionally uses the calibrated nearest-centroid decoder to predict all unordered pair distances in $\widehat{R}$. With $H = I - \mathbf{1}\mathbf{1}^\top/n$, it embeds the top d eigenvectors of $H(d\mathbf{1}\mathbf{1}^\top - 2\widehat{R})H$, scaled by the square roots of eigenvalues floored at 10^{-8}. FastICA (Hyvärinen, 1999) uses unit-variance whitening, 2000 maximum iterations, tolerance 2×10^{-4}, and four initializations. For each output U, divide column j by its median absolute value, using one when that median is below 10^{-8}. Select the smallest residual $\mathrm{mean}_{i,j}(|U_{ij}/s_j| - 1)^2$, where s_j denotes this scale. Binarization uses the signs of the selected output. The selection criterion uses no true coordinates. We record convergence and residuals.

Hungarian matching of absolute column correlations and independent sign alignment are applied after recovery solely for scoring. Coordinate accuracy is the fraction of correct bits; whole-state accuracy requires every bit of that state to be correct. Exact recovery requires every state to match. Table 2 reports the paired comparison. Every projector run passes the graph checks. For the non-identifying support $\{2, 4\}$, the theoretical edge margin is zero, so the certified procedure is inapplicable. Its calibrated baseline attains $57.67 \pm 0.67\%$ coordinate accuracy and $2.63 \pm 0.67\%$ whole-state accuracy on the original seeds.

Table 2: Seven-dimensional whole-state accuracy with the same complete responses. Entries are means and 95% interval half-widths in percent. Each row contains 30 runs; exact counts refer to the projector.

Seeds	Support	MDS/ICA (%)	Projector (%)	Exact
Original	$\{3, 5\}$	100.00 ± 0.00	100.00 ± 0.00	30/30
Original	$\{2, 3, 4\}$	98.62 ± 0.35	100.00 ± 0.00	30/30
Original	$\{1\}$	80.36 ± 10.49	100.00 ± 0.00	30/30
Independent	$\{3, 5\}$	100.00 ± 0.00	100.00 ± 0.00	30/30
Independent	$\{2, 3, 4\}$	98.59 ± 0.38	100.00 ± 0.00	30/30
Independent	$\{1\}$	83.70 ± 8.68	100.00 ± 0.00	30/30

G.3 q-ARY EDGE SEPARATION

The analytic sweep evaluates equation 49 for all 48 combinations of $q = 2, \ldots, 5$, $d = 4, 5, 6$, and $1 \le k < d$. Every proper nonconstant ternary or larger flag separates edges; the seven binary failures agree with the classification. Positive unnormalized margins range from 2 to 1240.

Task distribution and matched normalization. For $q = 2, 3$, $d = 5$, and each $k = 1, 2, 3, 4$, form real Gaussian combinations of the nonconstant complex characters with interaction order at most k. Normalize each task across states to unit standard deviation. Since the constant is excluded, write $R = D_{q,k} - 1$ and use the population kernel

$$\theta_k(r) = \frac{\kappa_{q,k}(r) - 1}{R}, \qquad \tau_k = \frac{\theta_k(1) + \max_{r \ge 2} \theta_k(r)}{2}.$$

The response-mean estimator is $FF^\top/m$. For the span estimator, center columns of F, compute the leading R left singular vectors Q, and use $nQQ^\top/R$. Both estimators use threshold τ_k. Random row permutations hide the state order from recovery; true states are used only to generate tasks and score the resulting edges. If fewer than R singular values exceed the numerical rank tolerance, we flag incomplete span and report the approximation using the available rank in the stress test.

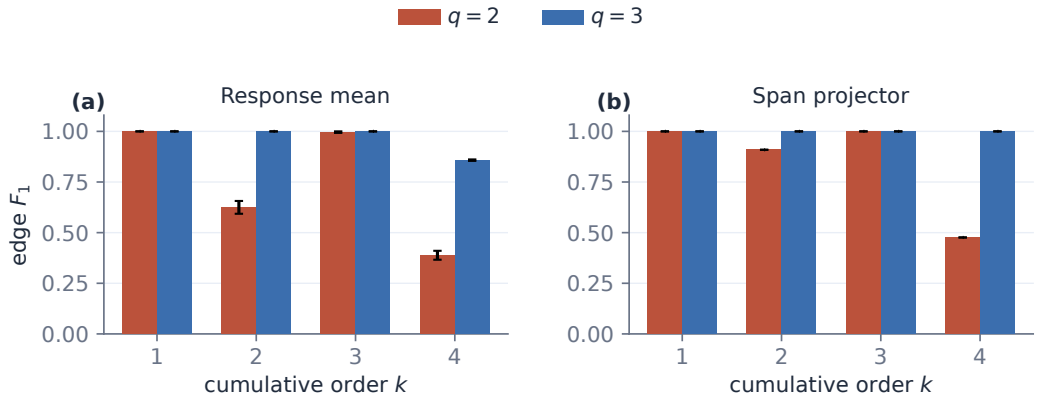


Figure 5: Graph recovery at $d = 5$ from the same 4000 complete task responses per run. Both response averaging and projection use the matched nonconstant kernel threshold. Means and 1.96 standard errors use ten original task seeds. All four ternary projector conditions recover the exact graph. Binary $k = 2, 4$ remain non-identifying despite positive edge F_1.

The projector uses floating-point tolerance $64\epsilon n \max(1, \|\widehat{K}\|_{\max})$ at the threshold. Pairs within that tolerance are marked uncertain and included in the candidate graph. In particular, binary collisions with zero margin are retained. For a clean complete span, adding the constant projector and rescaling gives the equivalent test with the constant included.

Complete panels and finite information. Original seeds 9000 through 9009 and independent seeds 19000 through 19029 cover all eight conditions. The projector recovers every edge in all 240 identifying runs; all 80 non-identifying runs fail exact graph recovery. Table 3 reports graph scores and compares the matched threshold with a threshold based on $\kappa_{q,k}/D_{q,k}$, applied to the same nonconstant response means. The latter normalization includes the constant component and is mismatched to this task generator. Its larger ternary fourth-order F_1 reflects a tradeoff between precision and recall.

Table 3: All q-ary main conditions. Edge F_1 uses the ten original seeds; exact graph counts are separate for the original and independent groups. The Unmatched column uses the threshold from $\kappa_{q,k}/D_{q,k}$ for responses without a constant component.

q	k	Unmatched	Matched mean	Projector	Exact orig.	Exact indep.
2	1	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000	10/10	30/30
2	2	0.0073 ± 0.0102	0.6248 ± 0.0316	0.9091 ± 0.0000	0/10	0/30
2	3	0.7592 ± 0.0158	0.9962 ± 0.0038	1.0000 ± 0.0000	10/10	30/30
2	4	0.0354 ± 0.0159	0.3883 ± 0.0220	0.4762 ± 0.0000	0/10	0/30
3	1	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000	10/10	30/30
3	2	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000	10/10	30/30
3	3	1.0000 ± 0.0000	1.0000 ± 0.0000	1.0000 ± 0.0000	10/10	30/30
3	4	0.9029 ± 0.0033	0.8579 ± 0.0036	1.0000 ± 0.0000	10/10	30/30

For $q = 3, d = 5, k = 4$, $R = 210$. A separate grid varies task count and additive Gaussian noise, using ten seeds 29000 through 29009 per cell. Table 4 retains every condition. Noise is added after the unit-variance normalization, and SVD truncation uses the known R. In the noiseless grid all runs with at least 210 tasks recover the graph. At 210 tasks and noise 0.25, mean F_1 is 0.387 with no exact recovery. At 4000 tasks all tested noise levels recover the graph.

An auxiliary frozen $q = 3$ diagnostic uses saved activations at $d = 4, 5$, singleton supports $\{1\}, \{2\}$, 3000 tasks, ridge coefficient 0.1, and seeds 5000 through 5009. All 40 evaluations of fitted responses and paired true responses attain edge $F_1 = 1$ with the matched threshold.

Table 4: Ternary fourth-order projector stress test. Entries give mean edge F_1, rounded to three decimals, and exact graph counts out of ten runs. Exact recovery requires every edge and nonedge to be correct.

Tasks	$\sigma = 0$	$\sigma = 0.05$	$\sigma = 0.1$	$\sigma = 0.25$
64	0.161 (0/10)	0.161 (0/10)	0.160 (0/10)	0.150 (0/10)
128	0.275 (0/10)	0.273 (0/10)	0.268 (0/10)	0.235 (0/10)
209	0.999 (1/10)	0.930 (0/10)	0.764 (0/10)	0.390 (0/10)
210	1.000 (10/10)	0.929 (0/10)	0.768 (0/10)	0.387 (0/10)
256	1.000 (10/10)	1.000 (10/10)	0.999 (0/10)	0.671 (0/10)
512	1.000 (10/10)	1.000 (10/10)	1.000 (10/10)	1.000 (1/10)
1024	1.000 (10/10)	1.000 (10/10)	1.000 (10/10)	1.000 (10/10)
4000	1.000 (10/10)	1.000 (10/10)	1.000 (10/10)	1.000 (10/10)

G.4 Frozen encoder task readouts

Training and saved representations. The synthetic Boolean system starts at the all-zero state, and named action a_i toggles factor i. A causal Transformer predicts a fixed bank of distinct exact cubic parities at every trajectory position. The bank sizes and model widths appear in Table 5; validation parity accuracy is at least 0.9989. Final normalized activations are averaged across observed occurrences of each labeled state, producing one fixed table per dimension. The stored activation tables are the inputs to all readout comparisons below.

Table 5: Boolean encoder training configurations. Bank sizes count actual distinct cubic targets.

d	Targets	Sequence length	Steps	Width
6	20	16	8000	192
7	35	16	8000	192
8	56	16	8000	192
9	84	18	8000	192
10	120	20	12000	256
11	165	22	12000	256

The training code uses binary cross-entropy on parity logits, AdamW with weight decay 0.01, gradient clipping at norm one, and a constant learning rate.

Supervised cumulative readouts. Standardize activation columns across all states, append an intercept, and fit ridge regression with penalty 0.1 on every coefficient. Each singleton support $\{3\}$, $\{2\}$, or $\{d-1\}$ receives 3000 Gaussian tasks from its cumulative Walsh space, including the constant. Target columns are scaled to unit standard deviation. The fitted responses are subsequently normalized to unit RMS before kernel estimation. Targets are constructed from true states, and all states enter fitting and scoring.

For these tasks including the constant, define $\theta_k^{\mathrm{full}}(r) = \kappa_k(r)/D_k$. Estimate the kernel by averaging response products and declare candidate edges at values at least $\theta_k^{\mathrm{full}}(1) - [\theta_k^{\mathrm{full}}(1) - \max_{r\geq 2}\theta_k^{\mathrm{full}}(r)]/2$. Reject if more than 5% of vertices have candidate degree different from d. Otherwise keep pairs for which each endpoint lies among the other's d partners with highest scores. This mutual top-d rule has degree at most d.

The labeler rejects isolated vertices, requires exactly d parallel edge classes obtained from opposite sides of four-cycles, and propagates their bit flips from a root. It rejects if any vertex remains unlabeled. Verification checks connectivity, degrees, label uniqueness, and edge consistency. The frozen score uses greedy column matching with sign alignment after output; the complete-response experiment above uses Hungarian alignment. All reported exact frozen successes also pass the strict cube checks.

Errors, oracle controls, and rejection. Seeds 7000 through 7009 resample tasks for each fixed encoder and support. The paired oracle uses the identical target functions directly, with the same

RMS normalization, threshold, and graph processing. We record raw and corrected edge F_1, kernel RMSE and maximum off-diagonal error, degree defects, rejection, and conditional coordinate and state accuracy. Coordinate and state accuracies are conditioned on a returned estimate. Rejection counts are reported separately. Edge F_1 in Figure 3(f) is measured before graph correction.

Table 6: Frozen readout diagnostics. Cumulative R^2, raw edge F_1, and kernel errors are means over ten task seeds per condition. Kernel RMSE is computed over off-diagonal state pairs within each run; Max. error averages the largest absolute off-diagonal error in each run. Errors are relative to the population kernel. Level R^2 uses all degree-k characters on the fixed encoder. Both R^2 measures precede RMS normalization. Reject and Oracle count rejected runs using fitted and oracle responses. All accepted runs recover the complete code.

d	k	Cumulative R^2	Level R^2	F_1	Kernel RMSE	Max. error	Reject	Oracle
6	3	0.997	1.000	1.0000	0.0217	0.1155	0/10	0/10
6	2	0.995	0.992	0.7834	0.0261	0.1207	10/10	10/10
6	$d-1$	0.995	0.999	0.3104	0.0224	0.1423	10/10	10/10
7	3	0.999	1.000	1.0000	0.0190	0.0777	0/10	0/10
7	2	0.997	0.997	0.6650	0.0219	0.0855	10/10	10/10
7	$d-1$	0.976	0.999	0.2011	0.0279	0.1769	10/10	10/10
8	3	0.984	0.999	1.0000	0.0257	0.1176	0/10	0/10
8	2	0.960	0.951	0.4845	0.0483	0.2309	10/10	10/10
8	$d-1$	0.736	0.933	0.1197	0.0429	0.2254	10/10	10/10
9	3	0.990	0.997	1.0000	0.0218	0.1343	0/10	0/10
9	2	0.977	0.972	0.5569	0.0355	0.2397	10/10	10/10
9	$d-1$	0.372	0.265	0.0712	0.0602	0.4504	10/10	10/10
10	3	0.990	0.995	1.0000	0.0211	0.1729	0/10	0/10
10	2	0.978	0.973	0.5869	0.0329	0.2556	10/10	10/10
10	$d-1$	0.248	0.148	0.0400	0.0572	0.5013	10/10	10/10
11	3	0.953	0.991	0.9997	0.0255	0.2682	0/10	0/10
11	2	0.859	0.831	0.4610	0.0616	0.4610	10/10	10/10
11	$d-1$	0.124	0.061	0.0219	0.0613	0.8710	10/10	10/10

Cubic readouts and their oracles recover every state in all 60 original runs. Both non-identifying controls are rejected in all 120 runs for each response source. Both supports have zero theoretical edge margin. Readout capacity also affects the fitted responses. At $d = 11$, the highest cumulative space has dimension 2047, whereas width 256 with an intercept gives at most 257 linear output dimensions. Its mean cumulative R^2 is 0.124, compared with 0.953 for cubic readouts.

For cubic readouts, an additional ten seeds 17000 through 17009 compare mean and span kernels using the same threshold and mutual top-d processing. Both methods recover the full code in all 120 cases across the two seed groups and six dimensions. A separate strict two-sided test followed by BFS succeeds in all 100 cases at $d = 6$ through 10 but rejects every one of the 20 cases at $d = 11$. This comparison changes both the threshold rule and graph correction jointly. Successful corrected graphs can have kernel errors above the sufficient bound at half the edge margin before correction.

G.5 Leakage calibration

The leakage calculation instantiates equation 4 at $d = 7$ and $k = 3$, where $D_k = 64$. We sample 320 perturbed bijections at transposition counts $1, 2, 3, 5, 8, 12, 20, 32$. The minimum leakage of a sampled non-coordinate map is 1.41×10^{-2}, approximately 462 times the $\beta/n = 3.05 \times 10^{-5}$ threshold. Every sampled alternative lies above the exact-recovery certificate.

G.6 Classification boundary at larger scale

The calibrated pipeline also evaluates $d = 13, 14$ with 8192 and 16384 states, using the original 30 seeds, disjoint pair splits, and 8192 tasks. Table 7 reports the nearest-centroid decoder on sampled test pairs. Supports $\{3, 5\}$ and $\{2, 3, 4\}$ are identifying; $\{2, 4\}$ is non-identifying. The separate complete-response coordinate evaluation runs at $d = 13$ only and uses the calibrated MDS/FastICA baseline.

Table 7: Larger dimensions with disjoint calibration and test pairs. Entries are means and 95% interval half-widths across 30 seeds. Distance metrics evaluate sampled pairs; coordinate and whole-state metrics evaluate all states.

d	Metric	$\{3,5\}$	$\{2,3,4\}$	$\{2,4\}$
13	Witness	1.0000 ± 0.0000	1.0000 ± 0.0000	0.5076 ± 0.0202
13	All distances	0.6784 ± 0.0040	0.8504 ± 0.0054	0.4968 ± 0.0061
13	Coordinate	1.0000 ± 0.0000	1.0000 ± 0.0000	0.5428 ± 0.0097
13	Whole state	1.0000 ± 0.0000	1.0000 ± 0.0000	0.0005 ± 0.0002
14	Witness	1.0000 ± 0.0000	1.0000 ± 0.0000	0.5198 ± 0.0196
14	All distances	0.6408 ± 0.0038	0.8690 ± 0.0038	0.4812 ± 0.0077

Both identifying supports recover every state in every run at $d = 13$. The non-identifying support has mean whole-state accuracy 0.00050. Accuracy across all distances and witness accuracy assess local distance prediction and remain distinct from the global reconstruction result.

Numerical runs use Python 3.12.11, NumPy 1.26.4, SciPy 1.16.0, and scikit-learn 1.8.0. Larger MDS runs use the leading d eigenpairs, with residuals below 2.5×10^{-10}.