GeoRARE: Certified Rarity Profiles for Structural Causal Effects

Anonymous authors
Paper under double-blind review

Abstract

Rare treatment–context regimes create two distinct problems: deciding how much they matter and determining what the data can establish about them. GeoRARE defines a two-dimensional rarity-indexed structural causal profile $(d, r) \mapsto \Theta_\Psi(d, r)$ that separately indexes treatment and context rarity. Sampling allocation governs information; reference weights govern the scientific question. One law-confidence event certifies the entire profile and its signed contrasts, with width controlled by unresolved reference mass. A finite-sample upper bound and a rare-cell lower bound identify the same local information scale, up to constants and logarithmic terms. Consequently, profile statements can be certified before all individual structural signatures are resolved. Categorical certification experiments, a stochastic oscillator, signed profile reversals, and an ACIC-covariate benchmark examine this distinction and its extension under measured confounding. The same rarity family also provides a secondary targeting interpretation as an interpretable training-priority family under finite-sample, regularization, and capacity trade-offs; the experiments show no universal predictive advantage. The contribution is a rarity-indexed structural causal profile whose entire surface and signed contrasts can be certified simultaneously, together with upper and lower results that characterize its information limits.

1 Introduction

An intervention can redistribute probability without changing a scientific regime of interest. Increased fluctuations around an equilibrium and the appearance of a low-occupancy gap may both change an outcome distribution, while only the latter changes a prespecified connectivity signature. A distributional equality test may correctly detect both. A structural audit asks whether a change, equality, or insufficient information can be established for the declared signature.

Rarity introduces a second distinction. The frequency with which a treatment–context regime is observed need not equal its scientific importance. A researcher may care about uncommon treatment responses or underrepresented contexts. Increasing their target weight changes the question being answered; it does not increase their observed sample size. Confusing these roles can produce large estimated effects without reliable structural evidence, or apparent learning gains obtained by changing the evaluation objective.

GeoRARE makes this separation explicit through a treatment exponent d and a context exponent r. In *profiling*, the scientific object is the complete map $(d, r) \mapsto \Theta_\Psi(d, r)$ for a fixed structural descriptor Ψ. The reference weights $q_d(a \mid z)$ and $\nu_r(z)$ determine which structural contrasts matter; the sampling probabilities π_{az} determine how much information is available. For nonnegative exponents, decreasing an exponent increases rare-to-common relative weight. No single (d, r) pair is privileged: the full surface is the inferential object, indexing sensitivity to the scientific reference population. The axes change the question, not the available information.

Integrated contribution. GeoRARE provides simultaneous, information-limited certification of a two-dimensional rarity-indexed structural causal profile, separating sampling information from scientific reference weights and allowing signed profile conclusions before all cell signatures are resolved. The central statistical quantity is $n\pi_{az}\rho^2_{\Psi,az}$: Theorem 1 and Corollary 2 give a sufficient certification regime, while Theorem 3 gives a rare-cell lower-bound construction showing that

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.

reliable certification is impossible when the corresponding information vanishes, up to constants and logarithmic terms at fixed geometry. Theorem 1 further translates cellwise information into a bound on whole-profile uncertainty through unresolved reference mass. Proposition 4 uses the same confidence event to certify signed increments, mixed contrasts, and reversals across rarity degrees, including data-selected contrasts. These are interactions in the rarity profile, not signed treatment effects or evidence of an interaction mechanism in a structural causal model.

The contribution concerns what can be certified jointly and how its uncertainty depends on the scientific reference population. GeoRARE builds on escort/power weights (Abe, 2003), confidence-set projection, structural summaries, and concentration inequalities. Related statistical-topology work develops confidence sets and rate analyses (Fasy et al., 2014; Balakrishnan et al., 2012; Rigollet & Vert, 2009), while structural causal contrasts and target-specific identification are studied in Faghihi & Saki (2026). The new inferential object is the full rarity-indexed structural causal profile $(d, r) \mapsto \Theta_\Psi(d, r)$, together with simultaneous whole-profile certification, signed profile contrasts, and achievable and impossible information regimes.

Recent work on topological causal effects addresses a related problem: Kim & Lee (2026) use power-weighted persistence-diagram silhouettes with doubly robust estimation and testing. Their power weights act within the topological representation. GeoRARE instead varies treatment and context reference weights, yielding the full rarity-indexed structural causal profile $(d, r) \mapsto \Theta_\Psi(d, r)$ while separating scientific weighting from the sampling information available for certification.

Distributional causal methods also address effects beyond means (Park et al., 2021; Fawkes et al., 2023; Martinez-Taboada et al., 2023). A valid full-law estimator can supply the same input confidence event, so the certification framework is not tied to a particular estimator of the outcome law.

The experiments follow this scientific question. A categorical factorial design tests the certification scale; a stochastic oscillator separates distributional changes from a prespecified structural change; a four-context design tests signed rarity-profile reversals; and an ACIC-covariate benchmark tests standardized causal-law certificates under measured confounding. Matched distributional, Wald, and finite-panel comparisons clarify which claims depend on the target, calibration, or representation. These studies report unresolved cases and negative findings alongside successful certification. Ordinary prediction metrics and generic distributional rejections do not by themselves certify statements about the rarity-indexed structural causal profile.

Secondary interpretation: targeting. The same d, r family can index importance-weighted training risks selected against a fixed external validation objective, but the fitted loss contains no Ψ. We use this as a secondary interpretation: an interpretable priority family for reallocating limited shared-model capacity under finite-sample, regularization, overlap, or model-capacity trade-offs. Its predictive behavior is evaluated empirically below.

2 Structural targets and the information available

Let $A \in \{1, \dots, J\}$ be treatment and $Z \in \{1, \dots, S\}$ a pre-treatment context. Independent units satisfy consistency and $Y(a) \perp A \mid Z$, with sampling probabilities $\pi_{az} = \mathbb{P}(Z = z)\mathbb{P}(A = a \mid Z = z) > 0$. The categorical causal law is $u_{az} = \mathbb{P}(Y(a) = \cdot \mid Z = z)$ on M fixed vertices. Conditional randomization is one sufficient design. This assumption is not a claim that geometry can reveal or remove hidden confounding.

The reference masses $p_{a|z}, h_z > 0$ are fixed independently of evaluation outcomes and can differ from the sampling allocation. With ordered pairs, the GeoRARE target is

$$q_d(a \mid z) = \frac{p_{a|z}^d}{\sum_b p_{b|z}^d}, \qquad \nu_r(z) = \frac{h_z^r}{\sum_u h_u^r}, \tag{1}$$

$$\Theta_\Psi(d, r) = \sum_z \nu_r(z) \sum_{a,b} q_d(a \mid z) q_d(b \mid z) \|\Psi(u_{az}) - \Psi(u_{bz})\|^2. \tag{2}$$

At $d = r = 0$, treatments and contexts receive equal reference weights. At $d = r = 1$, the specified reference frequencies are used. Lowering a weight exponent emphasizes rare cells relative to larger exponents; it does not necessarily give them greater absolute weight than frequent cells. For profiling,

the weights index different estimands. For learning, the same family can instead index training risks selected against a fixed external validation objective; the two uses must be distinguished.

Fix a finite complex K and an occupancy threshold $c \in (0,1)$. For a law u, retain vertices with $u_v \geq c$ and their induced simplices. In the experiments, $\Psi(u) = (\beta_0, \beta_1, \beta_2)$ over $\mathbb{R}$; an independent audit verifies agreement with $\mathbb{F}_2$ for every induced complex used. More generally, the theorem uses a specified finite signature table with diameter $D_\Psi = \max_{s,t} \|s - t\|$. Its decision radius is

$$\rho_\Psi(u) = \inf\{\|v - u\|_\infty : v \in \Delta_{M-1},\ \Psi(v) \neq \Psi(u)\}. \tag{3}$$

This radius can exceed the membership margin $\min_v |u_v - c|$: with an occupied cone center, changing outer vertices leaves the complex contractible. The radius is computable by distances to the polytopes of different-signature patterns. Exact enumeration is practical here for at most seven vertices and exponential in general. The target is the declared occupancy complex, not the topology of a continuous density or its support.

3 A CERTIFICATION THEOREM WITH AN IMPOSSIBILITY COUNTERPART

Let $C = JS$, $\ell_\alpha = \log(2MC/\alpha)$, and let N_{az} denote the random sample count in cell (a, z). For $N_{az} > 0$, intersect the empirical frequency box of radius $b_{az} = \sqrt{\ell_\alpha/(2N_{az})}$ with the simplex. Empty cells receive the whole simplex. Map each box to its feasible signature set, and minimize/maximize each squared signature distance over the corresponding pair of sets. Weighted sums give $L_\Theta(d, r)$ and $U_\Theta(d, r)$. Diagonal pair distances are identically zero.

Theorem 1 (Rarity, radius, and simultaneous certification)**.** *Under the preceding assumptions, these bands cover $\Theta_\Psi(d, r)$ simultaneously for every finite real d, r, with probability at least $1 - \alpha$. Set $\rho_{az} = \rho_\Psi(u_{az})$ and*

$$\mathcal{U}_n = \{(a, z) : n\pi_{az}\rho_{az}^2 \leq 4\ell_\alpha\}, \qquad \gamma_n = \sum_{a,z} \exp(-n\pi_{az}/8). \tag{4}$$

With probability at least $1 - \alpha - \gamma_n$, every cell outside $\mathcal{U}_n$ has its correct signature uniquely resolved, and, jointly for every d, r,

$$\begin{aligned} U_\Theta - L_\Theta &\leq D_\Psi^2 \sum_z \nu_r(z) \left[1 - \left\{ \sum_{a:(a,z)\notin\mathcal{U}_n} q_d(a \mid z) \right\}^2\right] =: \mathcal{B}_n(d, r) \\ &\leq 2D_\Psi^2 \sum_{(a,z)\in\mathcal{U}_n} \nu_r(z) q_d(a \mid z). \end{aligned} \tag{5}$$

On this event, $\Theta_\Psi(d, r) > \tau + \mathcal{B}_n(d, r)$ implies $L_\Theta(d, r) > \tau$. If $\mathcal{U}_n$ is empty, the whole profile is certified exactly. Negative probability lower bounds are read as zero.

The result has two different roles for rarity. Sampling rarity and poor overlap enter π_{az} and make signatures harder to resolve. Reference rarity enters $\nu_r q_d$ and determines how much unresolved structure matters to the target. The sufficient bound is conservative and may be vacuous. $\mathcal{U}_n$ involves unknown population radii and is a theoretical description of the regime, not an oracle quantity used by the estimator. Selecting d, r after viewing the simultaneous bands preserves coverage; it still changes the scientific question and cannot be relabeled as improved estimation of a fixed target.

Corollary 2 (Explicit sufficient sample size)**.** *If all radii are positive, a sufficient condition for complete correct profile certification with probability at least $1 - \delta$ is*

$$n > \max_{a,z} \left\{ \frac{8}{\pi_{az}} \log \frac{2C}{\delta},\quad \frac{4}{\pi_{az}\rho_{az}^2} \log \frac{4MC}{\delta} \right\}, \tag{6}$$

using the preceding construction with $\alpha = \delta/2$.

Theorem 3 (An unavoidable rare-cell uncertainty cost)**.** *There are two-context, two-arm models on a three-vertex path, with identical sampling allocation and all other cell laws identical, for which the rare cell has probability $\eta\epsilon$ and its two laws have center masses $c + \rho$ and $c - \rho$. For $c \in (0, 1/3)$*

and sufficiently small $\rho > 0$, both signature radii are ρ, but their GeoRARE effects are respectively 0 *and*

$$s(d, r) = 2\nu_r(z_*)q_d(0 \mid z_*)q_d(1 \mid z_*). \tag{7}$$

Define

$$K_n = n\eta\epsilon \ \mathrm{kl}(c + \rho, c - \rho), \qquad v_n = \min\{1, \sqrt{K_n/2}\}, \tag{8}$$

where kl *is binary relative entropy. Every procedure with false positive certification probability at most α in the zero-effect model has power at most $\min(1, \alpha + v_n)$ in the structural model. Every interval honest at level $1 - \alpha$ in both models satisfies, under the structural model,*

$$\mathbb{P}\{\mathrm{width}(I_{d,r}) \geq s(d, r)\} \geq \max\{0, 1 - 2\alpha - v_n\}. \tag{9}$$

For $0 < \rho \leq c/2$, $K_n \leq 8n\eta\epsilon\rho^2/[c(1-c)]$.

Thus reliable certification is impossible along local alternatives with $n\eta\epsilon\rho^2 \to 0$. The upper and lower results identify the same local information scale, up to logarithmic terms and constants at fixed M, J, S, c. They are not a sharp-constant minimax theorem for growing complexes. Positive reference weights with zero sampling overlap leave an irreducible uncertainty even at infinite total sample size. Hard topology at $\rho = 0$ also rules out uniform certification over arbitrarily close opposite-side laws. Proofs appear in Appendices A and B.

4 Two uses of rarity degrees

4.1 Profiling and certified changes across the surface

Write $s_{az} = \Psi(u_{az})$, $D_{abz} = \|s_{az} - s_{bz}\|^2$, and $\phi_d(z) = \sum_{a,b} q_d(a \mid z)q_d(b \mid z)D_{abz}$. In a Hilbert space, $\phi_d(z) = 2\,\mathrm{Var}_{q_d}(s_{Az})$. With $m_z(d) = \sum_a q_d(a \mid z) \log p_{a|z}$, direct differentiation gives

$$\begin{aligned} \partial_r \Theta_\Psi &= \mathrm{Cov}_{\nu_r}(\phi_d(Z), \log h_Z), \\ \partial_d \phi_d(z) &= \sum_{a,b} q_d(a \mid z)q_d(b \mid z)D_{abz}(\log p_{a|z} + \log p_{b|z} - 2m_z(d)). \end{aligned} \tag{10}$$

The mixed derivative is $\mathrm{Cov}_{\nu_r}(\partial_d\phi_d(Z), \log h_Z)$. Thus a positive context slope describes a positive weighted association between within-context contrasts and reference frequency at that value of d. It does not identify an individual rare arm. Flatness can arise from uniform reference masses despite strong effects, and a complete profile need not reconstruct its individual signatures. Since $\Theta_\Psi \geq 0$, the sign reversal of interest concerns a slope or increment.

Let $w_{abz}(d, r) = \nu_r(z)q_d(a \mid z)q_d(b \mid z)$ and $c_{abz} = \sum_j \lambda_j w_{abz}(d_j, r_j)$ for a signed combination of profile values. Put $c_+ = \max(c, 0)$ and $c_- = \min(c, 0)$ entrywise.

Proposition 4 (Simultaneous profile contrasts)**.** *On the law-confidence event of Theorem 1, every such contrast $T_c = \sum_j \lambda_j \Theta_\Psi(d_j, r_j)$ satisfies*

$$\sum (c_+ D^- + c_- D^+) \ \leq T_c \leq \sum (c_+ D^+ + c_- D^-). \tag{11}$$

The assertion is simultaneous over all finite coefficient choices and evaluation points, including choices made after examining the bands. Optimizing $\sum_{a,b} c_{abz}\|s_{az} - s_{bz}\|^2$ jointly over the feasible signatures in each context yields at least as tight valid bounds.

Negative coefficients must switch lower and upper distance bounds. Consequently, two oppositely signed context increments can be certified on one event without treating them as independent tests. Derivative coefficients are handled by the same argument. The finite-state exact optimization used below is practical because each context has four arms and at most four feasible component counts. General enumeration remains exponential. This is confidence-set projection, available also after any valid full-law estimation procedure.

4.2 Secondary use: targeting with an external objective

Let $\omega_{dr}(a, z) = \nu_r(z)q_d(a \mid z)$ and let ℓ be a prediction loss. If the sampling allocation π_{az} is known, define

$$R_{dr}(f) = \sum_{a,z} \omega_{dr}(a, z)\mathbb{E}\{\ell(f(a, z), Y(a)) \mid Z = z\} = \mathbb{E}\left[\frac{\omega_{dr}(A, Z)}{\pi_{AZ}}\ell(f(A, Z), Y)\right]. \quad (12)$$

The second equality uses the stated causal assumptions. Multiplying each observed loss by ω_{dr} alone would instead retain an extra factor of sampling prevalence. In the finite four-context design, $(1, 1)$ gives natural ERM and $(0, 0)$ gives equal-cell importance weighting.

With measured covariates X inside a context, assume exchangeability conditional on (X, Z) and first standardize the conditional arm laws across $X \mid Z$. The importance denominator becomes $h_z^{\text{sample}}e_a(X, z)$. Raw arm-cell intervals need not target this standardized causal law. Appendix D proves a simultaneous bounded-score construction, its two-fold projection rule, and the propensity-estimation bias boundary.

Fit candidates on training data. For a fixed external target mass ω_V, choose $(\hat{d}, \hat{r})$ and any regularization using an independent validation estimate of R_V. Evaluate the selected predictor once on independent test data, with the evaluation target held fixed. Maximizing the estimated structural effect is not this procedure. A finite-dictionary validation guarantee is proved in Appendix E.1; its variance factor is $\sum_{az} \omega_V(a, z)^2/\pi_{az}$, which grows under poor overlap.

This targeting rule is importance weighting restricted to an interpretable two-parameter priority family. In an unrestricted conditional model under a strictly proper loss, every strictly positive weighting has the same population optimum $f(a, z) = u_{az}$. In a saturated unregularized likelihood, the empirical conditional frequencies are also invariant to positive cell weights. Benefits therefore require a finite-sample, capacity, regularization, or optimization trade-off. Moreover, changing Ψ while holding the loss and weights fixed cannot change this learner. A claim that the training algorithm uniquely exploits topology would be incorrect.

5 Empirical evidence

5.1 Information and structural meaning

The factorial certification experiment contains 121,500 categorical datasets varying n, context prevalence, treatment propensity, structural radius, and $H_0/H_1/H_2$ geometry. It uses simultaneous Clopper–Pearson (CP) boxes (Clopper & Pearson, 1934), preserving the theorem's confidence event with tighter finite-sample intervals. At $\eta = \epsilon = .2$, $\rho = .02$, complete H_2 signature certification is $0\%, 2\%, 32\%, 100\%, 100\%$ for $n = 2{,}000, 8{,}000, 32{,}000, 128{,}000, 512{,}000$. All 9,600 datasets in the theorem's sufficient regime are correctly certified. Joint profile coverage is 99.9885%, including all 14 misses. The sufficient Hoeffding constants are conservative, not a fitted power curve.

The stochastic Hopf experiment uses independent stationary endpoint experiments with exact categorical probabilities. An increased-noise intervention preserves the prespecified occupied wheel complex; an annular alternative removes its center, creating one H_1 class. At $n = 80{,}000$, Gaussian MMD and categorical likelihood-ratio tests detect both alternatives in $200/200$ runs. GeoRARE certifies structural equality for the noise intervention and change for the annular alternative in $200/200$ each. At $n = 4{,}000$, the distributional tests detect the annular intervention while the structural certificate is always unresolved. Both findings are correct for their different scientific targets. The underlying continuous laws all have support $\mathbb{R}^2$. Detailed mechanisms, all outcomes, thresholds, and false certificates are reported in Appendices F and G.

5.2 Signed profile contrasts and rarity interactions

The profile experiment uses four contexts with masses $(.55, .25, .15, .05)$, four treatments with context-dependent allocation, and a seven-vertex path with threshold .10. Its true component-count

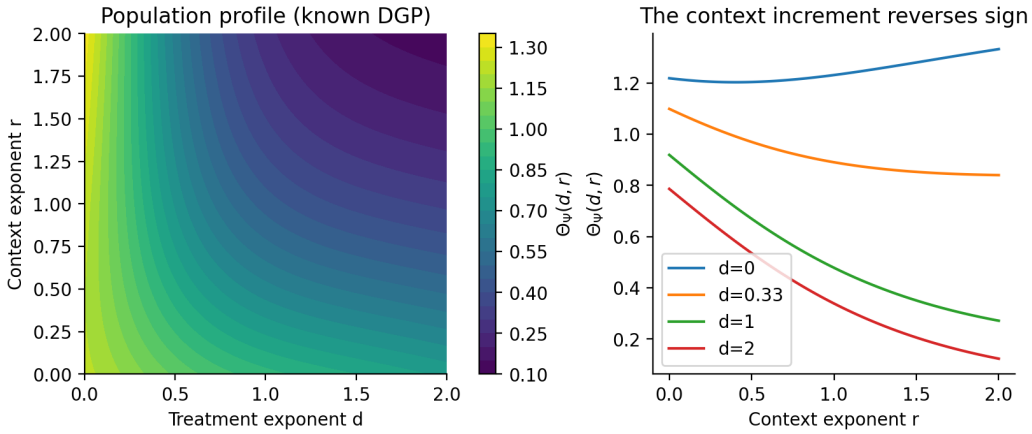


Figure 1: Known-DGP profile and cross-sections. The context increment $\Theta(d,2) - \Theta(d,0)$ is $+0.113782$ at $d = 0$ and -0.647836 at $d = 1$. These population calculations are separated from empirical certification. The DGP was designed for this pattern.

matrix is

$$[s_{az}]_{z,a} = \begin{pmatrix} 1 & 1 & 1 & 3 \\ 1 & 1 & 2 & 2 \\ 1 & 3 & 2 & 1 \\ 1 & 3 & 1 & 1 \end{pmatrix}. \tag{13}$$

The mechanism and threshold deliberately define a reversal in the population profile. Simulation seeds and analysis settings were fixed before evaluation; the reversal is a designed witness, not a discovery selected from simulated outcomes. The complete probability tables and the targeting protocol using this same mechanism are in Appendix H.

We sample 500 independent datasets at each $n \in \{5{,}000, 20{,}000, 80{,}000, 320{,}000\}$. CP boxes cover 112 coordinate probabilities jointly. They are projected onto component counts; exact joint signature optimization certifies the two signed context increments. All 2,000 datasets cover the true profile on the reported 21×21 grid and cover both increments. The continuum guarantee comes from the theorem, not this numerical grid check. Joint reversal certification is $0\%, 0\%, 0\%, 29\%$ across the four sample sizes; the final rate has an exact 95% interval $[25.1\%, 33.2\%]$. Complete resolution of all 16 signatures occurs in only $2/500$ of the largest datasets. This illustrates that task-specific projection can succeed before complete signature resolution while still requiring substantial information.

The comparison is between two inferential tasks on identical observations: certifying the requested signed profile pattern and identifying every cell signature. The former succeeds more often because joint projection uses the structural constraints relevant to the contrast. It does not recover missing information about every unresolved cell.

5.3 Structural certification under measured confounding

We evaluate a distributional extension of the ACIC 2016 benchmark (Dorie et al., 2019), using 4,802 real covariate rows with simulated four-arm interventions and seven-state outcomes. Fifty original-row splits cross two training sizes, randomized/confounded assignment, three propensity floors, three response complexities, and a no-effect control: 2,400 datasets. Nonlinear S- and T-learners, direct, equal-cell and overlap weighting, observed/standardized Group DRO, and estimated propensities share fixed external objectives and independent validation. All preprocessing uses training rows. This is semi-real validation, with simulated interventions and outcomes, not a clinical effect estimate.

Profiling uses independent samples and standardized causal laws before Ψ. For 450 confounded non-null cases with 20,000 training draws and threshold .10, the 320,000-draw profiling budget yields 100% joint law coverage for both known-propensity empirical-Bernstein constructions. IPW certifies a positive effect in 84.4% and two-fold AIPW in 57.1% of cases. Neither certifies the full direction reversal at this budget. Estimated-propensity Wald boxes have 85.8% joint law coverage; raw-cell

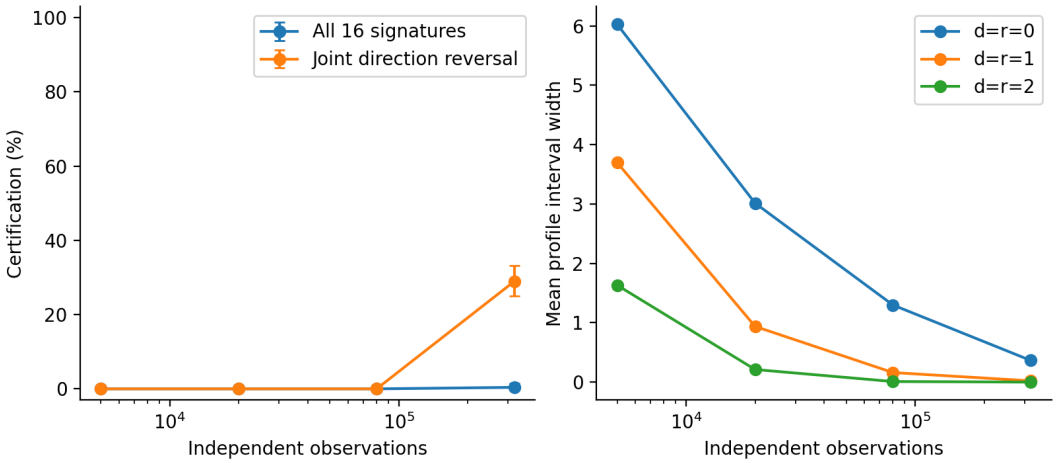


Figure 2: Signed-profile certification versus resolution of all 16 signatures, with exact binomial intervals over 500 datasets per sample size. Profile-band width depends on reference degrees at a fixed observation budget.

CP boxes have 31.8%. Their larger apparent certificate rates are not valid finite-sample evidence under the stated confounding. All failures, null controls and secondary thresholds are reported. The experiment makes causal adjustment and unresolved information substantive parts of the profile's interpretation.

Detailed protocols and condition-level results are in Appendix I.

5.4 Secondary use: targeting under restricted shared-model capacity

Each of 50 replications uses independent samples of 20,000 training, 10,000 validation, and 20,000 test observations. We fit a capacity-limited additive multinomial causal S-learner with context and treatment indicators but no treatment–context interaction. The method labelled *Escort / validation* is GeoRARE targeting: the priority weights q_d, ν_r in (12) are selected over a 7×7 degree grid and three regularization values by independent validation for four fixed external objectives: natural population, rare context, rare treatment, and balanced cells. Synthetic truth is used only for final scoring after all model choices are fixed.

Matched baselines include natural ERM, equal-cell weighting, generalized overlap weighting (Li & Li, 2019), finite-group DRO (Sagawa et al., 2020), direct external-objective weighting, and a saturated categorical T-learner. The saturated T-learner uses no GeoRARE weights and serves as a flexible capacity control; direct weighting uses the external objective itself rather than selecting within the GeoRARE priority family. An additional escort variant estimates the allocation from training counts. Table 1 reports population excess half-Brier risk; full protocols, validation selections, and secondary diagnostics are reported in Appendix H.

Compared with ERM, escort selection reduces rare-context, rare-treatment, and balanced risk by 27.7%, 15.9%, and 6.3%, respectively. These three paired differences remain significant after Holm correction, whereas the 0.47% increase in natural risk is not significant. Direct objective weighting has lower mean risk on every objective, and the saturated T-learner performs substantially better throughout. Thus this experiment supports d, r as interpretable training priorities under restricted shared-model capacity, rather than as an optimal weighting class or a substitute for model flexibility. The estimated-allocation variant is similar in this well-specified finite design, which does not establish robustness to arbitrary propensity misspecification.

Capacity and regularization in the ACIC benchmark. The initial penalty grid is inadequate: 93–100% of selections occur at its upper boundary across methods. We therefore conduct a regularization-sensitivity analysis using a wider grid and the uniform limiting predictor for every method. The analysis reuses the same training observations in all 2,400 cases but uses independent

Table 1: Four-context targeting experiment: mean population excess half-Brier risk $\times 1000$ after independent validation (50 replications; lower is better). All methods use the same declared evaluation objective. Escort / validation is GeoRARE targeting; direct weighting and the saturated T-learner are separate baselines.

Method	Natural	Rare context	Rare treatment	Balanced
Natural ERM	0.967	2.715	3.591	2.670
Equal-cell weighting	1.360	2.495	2.958	2.453
Overlap weighting	1.264	3.235	2.687	2.495
Group DRO	1.895	4.706	6.186	4.628
Escort / validation	0.972	1.962	3.021	2.501
Escort / estimated allocation	0.972	1.954	3.009	2.509
Direct objective weighting	0.967	1.701	2.447	2.453
Saturated T-learner	0.307	0.965	1.201	0.859

Method	Natural	Rare context	Rare treatment	Balanced
Natural ERM	2.918	5.227	5.050	4.404
Escort / known e	3.222	5.829	5.626	4.901
Direct objective	3.242	6.344	6.586	5.618
Flexible T-learner	3.377	5.997	6.156	5.126
Escort / estimated train + val	3.710	8.549	7.556	6.392
Uniform prediction	5.299	7.629	7.546	6.944

Table 2: Expanded regularization path with independent validation/test draws: excess half-Brier risk $\times 1000$, averaged over all 2,400 conditions. The short-path results appear in Table 8. All methods can select the uniform limiting predictor. Escort rows use the GeoRARE priority family; direct weighting and the T-learner are baselines.

validation and test outcomes. It should therefore be interpreted as a sensitivity analysis rather than an independent benchmark replication. Table 2 reports the resulting wider-path comparison. Escort selection has 10.4–11.5% higher average excess half-Brier risk than ERM across the four objectives, although it improves on direct weighting and the T-learner. Paired split-block intervals exclude zero after the four-comparison Holm adjustment. Thus the favorable finite additive-model result does not establish a general prediction advantage. Full condition tables, mean/CATE errors, and the short-/wide-path comparison are in Appendix I.

6 Discussion and conclusion

GeoRARE connects a two-dimensional scientific reference family to the information available for structural causal inference. The simultaneous event supports profile values and signed contrasts; unresolved reference mass bounds their uncertainty; and the rare-cell lower bound explains why changing priorities cannot compensate for insufficient observations. The signed-profile experiment demonstrates that a scientifically specified profile contrast can be certified even when individual cell signatures remain unresolved. The central contribution is the combination of a rarity-indexed structural causal profile, simultaneous whole-profile and signed-contrast certification, an uncertainty bound driven by unresolved reference mass, and a rare-cell lower bound on the same local information scale. The framework builds on structural causal contrasts, escort weighting, and concentration inequalities. The derivative identities describe the geometry of normalized reference weights.

The hard-signature theory assumes fixed finite geometry, declared reference masses and independent observational units. A learned representation must be fixed on independent training data; population reference-weight uncertainty must be accounted for separately. Enumeration does not scale to arbitrary complexes. Coarse signatures and the profile are generally noninjective. A characteristic-kernel or other full-law estimator can be followed by the same structural certificate. The upper and lower results match a local information scale at fixed geometry, not sharp minimax constants or growing-complexity rates.

The experiments support this inferential scope and expose its limits. Measured-confounding certificates require standardized causal laws and valid score uncertainty; estimated-propensity Wald boxes need additional nuisance-error control. Matched CDF and integrated-Euler statistics can be identical. The sci-Plex finite-panel study contains unfavorable topological results and uses fixed intervention profiles under imposed missing wells; it does not establish population biological identification or transfer to independent biological experiments. The ACIC analysis is conditional on one covariate release and finite-feature model classes. Its two regularization analyses share training data and mechanisms.

Targeting is a secondary application of the same rarity-priority family. Its mixed predictive results support an interpretable finite-sample priority mechanism rather than universal predictive superiority. The primary contribution is a principled way to define, profile, and certify rarity-sensitive structural causal questions while explicitly identifying when the available information is insufficient to resolve them.

Reproducibility Statement

The appendices provide complete mathematical proofs, experimental protocols, fixed seeds, data-generation mechanisms, evaluation criteria, and detailed condition-level results supporting the reported experiments. Appendices F–I describe the certification, oscillator, profile, targeting, and ACIC designs, while Appendix P documents computational validation. The experiments were conducted under fixed protocols and recorded computational checks. Anonymized implementation and reproduction materials can be made available to reviewers and area chairs during the discussion period.

AI use statement

Generative AI tools were used for wording, language editing, and literature-review assistance. They were not used to produce the mathematical derivations or theoretical results. The authors reviewed and verified the final manuscript and take responsibility for its content.

References

Sumiyoshi Abe. Geometry of escort distributions. *Physical Review E*, 68:031101, 2003. doi: 10.1103/PhysRevE.68.031101.

Sivaraman Balakrishnan, Alessandro Rinaldo, Don Sheehy, Aarti Singh, and Larry Wasserman. Minimax rates for homology inference. In *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*, volume 22, pp. 64–72, 2012. URL https://proceedings.mlr.press/v22/balakrishnan12a.html.

Roger L. Berger and Jason C. Hsu. Bioequivalence trials, intersection-union tests and equivalence confidence sets. *Statistical Science*, 11(4):283–319, 1996. doi: 10.1214/ss/1032280304.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 2018. doi: 10.1111/ectj.12097. URL https://arxiv.org/abs/1608.00060.

C. J. Clopper and E. S. Pearson. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26(4):404–413, 1934. doi: 10.1093/biomet/26.4.404.

Thai Son Doan, Maximilian Engel, Jeroen S. W. Lamb, and Martin Rasmussen. Hopf bifurcation with additive noise. *Nonlinearity*, 31(10):4567–4601, 2018. doi: 10.1088/1361-6544/aad208. URL https://arxiv.org/abs/1710.09649.

Vincent Dorie, Jennifer Hill, Uri Shalit, Marc Scott, and Dan Cervone. Automated versus do-it-yourself methods for causal inference: Lessons learned from a data analysis competition. *Statistical Science*, 34(1):43–68, 2019. URL https://arxiv.org/abs/1707.02641.

DoubleML contributors. DoubleML: Interactive regression models and distributional targets, version 0.11.2, 2026. URL https://docs.doubleml.org/stable/api/generated/doubleml.irm.DoubleMLIRM.html.

Usef Faghihi and Amir Saki. Topological ignorability for structural causal effects beyond means. *arXiv preprint arXiv:2603.14169v2*, 2026. URL https://arxiv.org/html/2603.14169v2.

Brittany Terese Fasy, Fabrizio Lecci, Alessandro Rinaldo, Larry Wasserman, Sivaraman Balakrishnan, and Aarti Singh. Confidence sets for persistence diagrams. *The Annals of Statistics*, 42(6):2301–2339, 2014. URL https://arxiv.org/abs/1303.7117.

Jake Fawkes, Robert Hu, Robin J. Evans, and Dino Sejdinovic. Doubly robust kernel statistics for testing distributional treatment effects. *arXiv preprint arXiv:2212.04922v2*, 2023. URL https://arxiv.org/html/2212.04922v2.

D. G. Horvitz and D. J. Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260):663–685, 1952. doi: 10.1080/01621459.1952.10483446.

Dominik Janzing, David Balduzzi, Moritz Grosse-Wentrup, and Bernhard Schölkopf. Quantifying causal influences. *The Annals of Statistics*, 41(5):2324–2358, 2013. doi: 10.1214/13-AOS1145. URL https://arxiv.org/abs/1203.6502.

Kwangho Kim and Hajin Lee. Topological causal effects. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/4be03be63b57362bb0344547490ee196-Abstract-Conference.html.

Fan Li and Fan Li. Propensity score weighting for causal inference with multiple treatments. *The Annals of Applied Statistics*, 2019. URL https://arxiv.org/abs/1808.05339.

Diego Martinez-Taboada, Aaditya Ramdas, and Edward H. Kennedy. An efficient doubly-robust test for the kernel treatment effect. *arXiv preprint arXiv:2304.13237v2*, 2023. URL https://arxiv.org/abs/2304.13237.

Andreas Maurer and Massimiliano Pontil. Empirical bernstein bounds and sample variance penalization. In *Proceedings of the Conference on Learning Theory*, 2009. URL https://arxiv.org/abs/0907.3740.

Facundo Mémoli, Zhengchao Wan, and Yusu Wang. Persistent laplacians: properties, algorithms and implications. *SIAM Journal on Mathematics of Data Science*, 4(2):858–884, 2022. doi: 10.1137/21M1435471. URL https://arxiv.org/abs/2012.02808.

Junhyung Park, Uri Shalit, Bernhard Schölkopf, and Krikamol Muandet. Conditional distributional treatment effect with kernel conditional mean embeddings and u-statistic regression. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 8401–8412, 2021. URL https://proceedings.mlr.press/v139/park21c.html.

PyWhy contributors. DoWhy: Causal inference library, version 0.14, 2026. URL https://www.pywhy.org/dowhy/v0.14/.

Philippe Rigollet and Régis Vert. Optimal rates for plug-in estimators of density level sets. *Bernoulli*, 15(4):1154–1178, 2009. doi: 10.3150/09-BEJ184. URL https://arxiv.org/abs/math/0611473.

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *ICLR*, 2020. URL https://arxiv.org/abs/1911.08731.

Sanjay R. Srivatsan et al. Massively multiplex chemical transcriptomics at single-cell resolution. *Science*, 367(6473):45–51, 2020. doi: 10.1126/science.aax6234. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC7289078/.

GUIDE TO THE APPENDICES

Appendices A–C give the central certification proofs and signed-contrast results. Appendices D–E develop measured-confounding and validation extensions. Appendices F–I specify the certification, oscillator, profile/targeting, and ACIC experiments. Appendices J–O provide identifying assumptions, structural geometry, matched comparisons, finite-panel results, and spectral extensions. Appendix P documents computation and audits.

A PROOF OF THEOREM 1 AND COROLLARY 2

This appendix proves Theorem 1 and Corollary 2. Appendix B proves the rare-cell lower bound, and Appendix C develops simultaneous signed profile contrasts.

A.1 CONDITIONAL SAMPLING, PROJECTION, AND THE UPPER REGIME

Conditional on the cell counts $\{N_{az}\}$, the categorical outcomes in each nonempty cell are iid from u_{az}. This follows from iid sampling and conditional exchangeability; conditional randomization is sufficient in the categorical certification experiments. Each coordinate empirical frequency is a binomial proportion. Hoeffding's inequality and the definition of b_{az} imply

$$\mathbb{P}\{|\widehat{u}_{az,v} - u_{az,v}| > b_{az} \mid \{N_{az}\}\} \le 2e^{-2N_{az}b_{az}^2} = \frac{\alpha}{MC}.$$

Empty cells use the entire simplex and incur no coverage error. Union over MC coordinates and integration over the random counts give an event E of probability at least $1-\alpha$ on which all true laws belong to their confidence sets. Dependence between categorical coordinates and between cell counts is immaterial to the union bound.

On E, every true signature belongs to its projected set. Thus each true pair distance lies between the minimum and maximum of that pair's feasible distances. Nonnegative escort weights preserve these inequalities under summation. The same event holds before d, r are chosen and applies to all of them; no union bound over the continuum of weights is required. The nonnegative-weight summation must be modified for signed coefficients, as in Proposition 4. It also requires fixed known reference masses, or appropriate joint uncertainty sets for estimated ones.

For each cell, the count N_{az} is marginally binomial with success probability π_{az}. A multiplicative Chernoff bound gives

$$\mathbb{P}\{N_{az} < n\pi_{az}/2\} \le \exp(-n\pi_{az}/8).$$

Let G be the simultaneous event $N_{az} \ge n\pi_{az}/2$. Then $\mathbb{P}(E \cap G) \ge 1-\alpha-\gamma_n$. On this event, every candidate law v in a cell confidence set obeys

$$\|v - u_{az}\|_\infty \le \|v - \widehat{u}_{az}\|_\infty + \|\widehat{u}_{az} - u_{az}\|_\infty \le 2b_{az} \le 2\sqrt{\frac{\ell_\alpha}{n\pi_{az}}}.$$

Outside $\mathcal{U}_n$, the last expression is strictly smaller than ρ_{az}, so every candidate law has the true signature. Strictness is essential for a hard threshold: the radius is an infimum, not a permitted perturbation size at equality. In particular, cells with radius zero are not certified by this argument.

Pairs of resolved cells have zero interval width. Any remaining pair has width at most D_Ψ^2. Let t_z be the total treatment escort mass of unresolved cells in context z. The ordered-pair mass touching an unresolved cell is $1-(1-t_z)^2 = 2t_z - t_z^2 \le 2t_z$. Including unresolved diagonal pairs in this upper bound is harmless, although their actual width is zero. This proves both inequalities in (5). Intersecting with coverage also gives $L_\Theta \ge \Theta_\Psi - (U_\Theta - L_\Theta)$, proving the practical-threshold claim. If every cell is resolved, every pair distance and hence the entire profile is exactly determined. Pairwise extrema need not be jointly attainable, so the bands are outer bounds, not generally sharp identified intervals.

For Corollary 2, set $\alpha = \delta/2$. The first sample-size condition bounds each count-failure term by $\delta/(2C)$, so $\gamma_n \le \delta/2$. The second gives $\mathcal{U}_n = \varnothing$. Applying Theorem 1 proves the claim. No knowledge of the population radius is needed to compute the confidence band; it enters only this sufficient sample-size statement.

B Proof of Theorem 3

B.1 The local two-point construction and lower bound

Use $Z \sim \text{Bernoulli}(\eta)$, with $\mathbb{P}(A = 1 \mid Z = 1) = \epsilon$, and identical allocations in two models. Conditional on Z, assign treatment independently of potential outcomes. All cells except $(Z, A) = (1, 1)$ have the same connected outcome law in both models. They can, for example, put center mass $c + \kappa$ on the path, where $0 < \kappa$ is fixed and outer masses remain above c. Potential outcomes with these marginals exist, for example under conditional independent coupling; no cross-world estimand is used.

For the rare cell, assign the two outer vertices equal mass and set

$$u_+ = \left(\frac{1-c-\rho}{2}, \frac{1-c-\rho}{2}, c+\rho\right), \qquad u_- = \left(\frac{1-c+\rho}{2}, \frac{1-c+\rho}{2}, c-\rho\right).$$

For $c \in (0, 1/3)$ and ρ sufficiently small, all outer vertices are above c in both laws. The plus complex is a connected path; the minus complex has two components. While the center stays active, any induced outer subcomplex coned to it is contractible. For sufficiently small ρ, losing an outer vertex in the minus law requires more than ρ coordinate perturbation. Moving the center across c while distributing the offset across outer vertices has infimum cost ρ. Both radii therefore equal ρ, and the squared signature difference is one. Only the two ordered treatment pairs in the rare context differ, which yields (7).

The same construction works on a cone over a four-cycle or an octahedral sphere, with uniform outer mass, when ρ is sufficiently small to keep the outer shell above threshold. The plus signature is $(1, 0, 0)$, and the minus signature acquires one H_1 or H_2 class. All declared experiment radii satisfy the required margins; independent linear programs verify them. This is a finite-complex extension of the explicit path witness, not a lower bound for every possible structural descriptor.

The conditional law of the outer vertex given "not center" is uniform in both models. Therefore $\mathrm{KL}(u_+, u_-) = \mathrm{kl}(c+\rho, c-\rho)$. Since all other cell laws and allocations coincide, the chain rule for relative entropy gives the *exact* one-unit divergence $\eta\epsilon\,\mathrm{kl}(c+\rho, c-\rho)$. Independence multiplies it by n, proving (8). Pinsker's inequality bounds the total variation of the full observed datasets by v_n. For any certification event T, $\mathbb{P}_-(T) \le \mathbb{P}_+(T) + v_n$, which proves the power bound when $\mathbb{P}_+(T) \le \alpha$. This argument permits randomized procedures and does not restrict their representation, nuisance models, or computation.

For the confidence-width statement, honesty gives $\mathbb{P}_+(0 \in I) \ge 1 - \alpha$ and $\mathbb{P}_-(s \in I) \ge 1 - \alpha$. Transfer the first event via total variation to obtain $\mathbb{P}_-(0 \in I) \ge 1 - \alpha - v_n$. Their intersection has probability at least $1 - 2\alpha - v_n$; an interval containing both endpoints has width at least s. The assertion is pointwise for each fixed (d, r), and thus also applies to any simultaneous band at each such point. It does not assert a joint lower probability for a data-selected continuum of points.

The elementary inequality $\log x \le x - 1$ implies

$$\mathrm{kl}(p, q) \le \frac{(p-q)^2}{q(1-q)}.$$

Substituting $p = c + \rho, q = c - \rho$ and using $c - \rho \ge c/2, 1 - c + \rho \ge 1 - c$ gives the stated upper bound on K_n. This local bound is used only when $\rho \le c/2$; the exact binary KL formula is used for every plotted experiment radius. If $n\eta\epsilon\rho^2 \to 0$, then $v_n \to 0$ and level-α certification has power at most $\alpha + o(1)$. In contrast, (6) supplies a sufficient regime at the same local scale, with logarithmic factors and conservative constants.

If a sampling cell has zero probability while its declared reference weight remains positive, the two observational laws agree exactly and the lower bound holds with $v_n = 0$. This is a limit/counterexample outside the positivity assumption of the upper theorem, not an application of that assumption at zero overlap. At $\rho = 0$ the displayed plus and minus laws coincide, so one must instead consider arbitrarily small positive ρ to demonstrate nonuniformity near the boundary. The code accordingly does not label the identical boundary laws as a distinct two-point pair.

B.2 A SIMULTANEOUS FORM OF THE LOWER BOUND

Theorem 3 assumes pointwise honest intervals. If the stronger assumption holds that a band contains each complete population surface with probability at least $1-\alpha$ in its respective model, then under the structural model, with probability at least $\max(0, 1-2\alpha-v_n)$, that band contains *both* complete surfaces. Indeed, transfer the event of containing the zero surface from the zero-effect model by total variation, then intersect with the event of containing the structural surface. On this intersection,

$$U(d,r) - L(d,r) \geq s(d,r) \quad \text{for every } (d,r).$$

This conclusion also applies at a data-selected point. It requires simultaneous honesty; it cannot be deduced by an uncountable union of pointwise lower bounds. It strengthens the interpretation of the two-point witness without claiming a general sharp minimax theorem.

B.3 BOUNDARY AND RARE-OBSERVATION LOWER BOUNDS

For the loop or void, take the center mass $c \pm \epsilon$ and distribute the residual mass equally over outer vertices. Near $c = 0.1$ all masses are bounded away from zero. Taylor expansion of finite multinomial KL divergence gives $O(\epsilon^2)$ between these laws. Alter only the rare arm/context cell, leaving the rest of the causal model unchanged. The resulting structural targets differ by a nonzero constant Δ at any fixed finite (d,r) with positive pair weights, while n-sample KL is $O(n\epsilon^2)$. Thus for $\epsilon = o(n^{-1/2})$, total variation between the two sample distributions tends to zero.

For any estimator T and two sample laws with densities f_0, f_1, integrate with respect to their common submeasure $\min(f_0, f_1)$:

$$\max_{j\in\{0,1\}} \mathbb{E}_j[(T-\theta_j)^2] \geq \frac{1}{2}\int \{(T-\theta_0)^2 + (T-\theta_1)^2\} \min(f_0, f_1) \tag{14}$$

$$\geq \frac{\Delta^2}{4}\{1 - \mathrm{TV}(f_0, f_1)\}. \tag{15}$$

The bound stays positive along this local-alternative sequence, ruling out uniform consistency. This is a standard two-point argument specialized to the structural discontinuity, not a novel minimax technique. If two models instead agree whenever a region of probability π is unobserved, the common submeasure of their n-sample laws has mass at least $(1-\pi)^n$. The same reasoning gives $\max_j \mathrm{MSE}_j \geq \Delta^2(1-\pi)^n/4$. Changing escort weights can change Δ but cannot remove the event that the rare region was never observed.

C SIMULTANEOUS SIGNED PROFILE CONTRASTS

C.1 VARIANCE, DERIVATIVES, AND LIMITS OF INTERPRETATION

For fixed z, let $q_a = q_d(a \mid z)$, $s_a = \Psi(u_{az})$, $m = \sum_a q_a s_a$. Expansion of the squared norm gives

$$\sum_{a,b} q_a q_b \|s_a - s_b\|^2 = 2\sum_a q_a \|s_a\|^2 - 2\|m\|^2 = 2\sum_a q_a \|s_a - m\|^2.$$

Both ordered off-diagonal terms contribute; every diagonal term vanishes. This is why the factor two is necessary. Differentiating the normalized exponential family gives

$$\partial_d q_a = q_a(\log p_{a|z} - m_z(d)), \qquad \partial_r \nu_z = \nu_z(\log h_z - \mathbb{E}_{\nu_r} \log h_Z).$$

Substituting in the finite sums proves (10) and the mixed derivative stated in the main text. Equivalently,

$$\partial_d \phi_d(z) = 2\,\mathrm{Cov}_{q_d}(\|s_A - m\|^2, \log p_{A|z}).$$

The derivative of m drops out after centering. This expression is an association between structural dispersion and log reference frequency; the dispersion itself changes with d. There is no universal sign.

For $p_a < p_b$, the relative weight $q_d(a)/q_d(b) = (p_a/p_b)^d$ decreases with d. This establishes the relative-rarity interpretation. It does not assert that every individual pair product $q_a q_b$ decreases

with d, or that smaller degrees increase the effect. If all reference masses on an axis are uniform, the corresponding profile is flat regardless of the magnitude of its structural contrasts. Moreover, replacing all s_{az} in a fixed context by $s_{az} + c_z$ preserves every pair distance and hence the entire surface. For scalar component counts, the two-arm patterns $(1, 2)$ and $(2, 3)$ give a realizable example on a seven-vertex path with threshold .10. Thus the complete profile need not identify the individual signatures or even their absolute levels. Mixed curvature describes sensitivity to the reference weights; it is not by itself a treatment–context interaction in a structural causal model.

C.2 Proof of simultaneous signed projection

Let E be the event that every true law belongs to its confidence set. On E, every true squared pair distance obeys $D^-_{abz} \le D_{abz} \le D^+_{abz}$. Multiplication by a positive coefficient preserves the inequality, while a negative coefficient reverses it. Summing proves (11) for any real coefficient array. This deterministic implication is valid for every array on the same event, so it also holds for data-chosen arrays without an additional multiplicity correction. Taking coefficients to be finite combinations of profile weights gives Proposition 4; using analytic derivative weights gives simultaneous derivative bounds.

For tighter bounds, write $\mathcal{S}_{az}$ for the projected signature sets. The feasible product contains the true signature tuple on E. Hence

$$L_c^{\text{joint}} = \sum_z \min_{s_a \in \mathcal{S}_{az},\, a=1,\dots,J} \sum_{a,b} c_{abz} \|s_a - s_b\|^2,$$

$$U_c^{\text{joint}} = \sum_z \max_{s_a \in \mathcal{S}_{az},\, a=1,\dots,J} \sum_{a,b} c_{abz} \|s_a - s_b\|^2$$

are valid and contained within the pairwise outer bounds. Separability holds across contexts because neither the target nor the confidence product couples their signatures. These are exact extrema over the enumerated signature product, which may itself conservatively include threshold-face closures. They are not automatically sharp confidence intervals over every statistically plausible full joint law.

The code computes the increment at $d = 0$ and the increment at $d = 1$ on this common product. A strictly positive lower bound for the former and a strictly negative upper bound for the latter certify the stated reversal. Subtracting two marginal intervals would remain valid on a simultaneous event but can discard their shared structural constraints.

C.3 A sharper unresolved-mass bound

Theorem 1 includes unresolved diagonal mass in a conservative upper bound. Since those pair distances are exactly zero, the same proof gives the optional refinement

$$U_\Theta - L_\Theta \le D_\Psi^2 \sum_z \nu_r(z) \left(2t_z - t_z^2 - \sum_{a:(a,z)\in\mathcal{U}_n} q_d(a \mid z)^2 \right), \tag{16}$$

where $t_z = \sum_{a:(a,z)\in\mathcal{U}_n} q_d(a \mid z)$. The total ordered-pair mass touching an unresolved cell is $2t_z - t_z^2$; its unresolved diagonal mass is the final sum. All remaining pairs have width at most D_Ψ^2. The expression is nonnegative because it sums nonnegative off-diagonal pair masses. The bound in Theorem 1 remains valid; this refinement subtracts only the unresolved diagonal mass.

C.4 Full-profile derivatives and reference geometry

Write $s_{abz} = \|\Psi(P^a_z) - \Psi(P^b_z)\|^2$, $\phi_d(z) = \sum_{a,b} q_a q_b s_{abz}$ and $\ell_z = \sum_a q_a \log p_{a|z}$. Differentiating normalized powers gives

$$\partial_d \Theta = \mathbb{E}_{\nu_r q_d q_d}[s_{ABZ}(\log p_{A|Z} + \log p_{B|Z} - 2\ell_Z)], \tag{17}$$

$$\partial_r \Theta = \mathrm{Cov}_{\nu_r}(\phi_d(Z), \log h_Z), \tag{18}$$

$$\partial_{rd} \Theta = \mathrm{Cov}_{\nu_r}(\partial_d \phi_d(Z), \log h_Z). \tag{19}$$

These are sensitivity derivatives of an estimand, not gradients that improve an estimator's sample size. Since all conditional d scores have mean zero, the Fisher information of $\nu_r(z) q_d(a \mid z) q_d(b \mid z)$

has zero off-diagonal entry, with $I_{dd} = 2\mathbb{E}_{\nu_r} \mathrm{Var}_{q_d}(\log p_{A|Z})$ and $I_{rr} = \mathrm{Var}_{\nu_r}(\log h_Z)$. These expressions vanish for the corresponding uniform reference. For continuous variables an escort density depends on a declared base measure and coordinate system; discretization is not a coordinate-invariance theorem.

D MEASURED-CONFOUNDING EXTENSIONS

The standardized law is the object of estimation; the structural transform is applied after adjustment. The ACIC notation uses X for measured covariates. The finite-confounder formulation below uses W for the same statistical role and supports the experiments in Appendix M.

D.1 VARIANCE-ADAPTIVE LAW CERTIFICATES UNDER MEASURED CONFOUNDING

Write $p_{ay}(x,z) = \mathbb{P}\{Y(a) = y \mid X = x, Z = z\}$ and $u_{azy} = \mathbb{E}\{p_{ay}(X,z) \mid Z = z\}$. Let $m_{ay}(x,z) \in [0,1]$ be an outcome prediction trained independently of the evaluation observations. For known $e_a(x,z) > 0$, define

$$S_{ay} = m_{ay}(X,Z) + \frac{\mathbf{1}\{A=a\}}{e_a(X,Z)}(\mathbf{1}\{Y=y\} - m_{ay}(X,Z)). \tag{20}$$

Its expectation within context is u_{azy} for every such prediction. The IPW estimator is the special case $m = 0$.

Proposition 5 (Simultaneous score projection)**.** *Conditional on independent nuisance training, suppose the score for a fixed (a,z,y) lies in a known interval of length R_{azy}. For $N_z > 1$ independent evaluation observations in context z, let $\bar{S}_{azy}$ and $\widehat{V}_{azy}$ be their mean and unbiased sample variance. With $K = 112$, set*

$$b_{azy} = \sqrt{\frac{2\widehat{V}_{azy}\log(4K/\alpha)}{N_z}} + \frac{7R_{azy}\log(4K/\alpha)}{3(N_z-1)}. \tag{21}$$

Intersect $[\bar{S}-b, \bar{S}+b]$ with $[0,1]$ coordinatewise and project the resulting boxes onto the simplexes and signatures. With probability at least $1-\alpha$, the entire collection of laws belongs to the boxes, and the profile and signed-contrast conclusions of Proposition 4 hold simultaneously. A context with at most one observation receives the whole simplex.

Proof. Conditional expectation of the residual in (20) is $p_{ay} - m_{ay}$, which proves unbiasedness. Rescale each score to $[0,1]$. Apply the empirical Bernstein inequality of Maurer & Pontil (2009) separately to its upper and lower tails, allocating $\alpha/(2K)$ to each. Rescaling back gives (21). Union over the K coordinates. Conditional on a context count, the observations inside that context have its conditional law. Integrating over counts and nuisance training preserves the bound. Simplex intersection and signature projection retain every true law on the joint event; the simultaneous profile result follows deterministically, including the three prespecified ACIC thresholds. □

On a finite covariate pool, score endpoints are computed from all possible covariate rows and responses using the independently fitted m and known e, without outcome truth. A looser universal interval is $[1-1/\epsilon, 1/\epsilon]$ for augmented scores and $[0, 1/\epsilon]$ for IPW. Empty numerical intersections are reported and replaced by vacuous signature sets; they are never relabelled as equality.

For two-fold distributional AIPW, original test-pool row identifiers are split into two disjoint folds. An outcome nuisance for a fold is trained on the original training observations and the other profiling fold only, using the independently selected T-learner penalty. Both original-row and outcome separation hold. Apply (21) within each fold with $\alpha/2$, then combine its coordinate endpoints using the known fractions of finite-pool covariate rows in that fold within each context. Conditional unbiasedness targets the corresponding fold population; their weighted average is the declared whole-pool law. A union bound over the folds requires no assertion that pooled cross-fit scores are iid. Dependencies between the two fitted nuisance functions therefore do not invalidate this construction. Predictive-test observations are separate from these profiling observations and never train the targeting predictors.

Proposition 6 (Variance and nuisance-error boundary)**.** *For a fixed arm, response and context, abbreviate* $p = p_{ay}(X, z)$, $m = m_{ay}(X, z)$, *and* $e = e_a(X, z)$. *The score variance is*

$$\mathrm{Var}(S \mid Z = z) = \mathrm{Var}(p \mid Z = z) + \mathbb{E}\left[\frac{p(1-p)}{e} + \left(\frac{1}{e} - 1\right)(p-m)^2 \,\middle|\, Z = z\right]. \tag{22}$$

It is minimized over independent outcome predictions by $m = p$ *pointwise where* $e < 1$. *Replacing* e *by an independent estimate* $\hat{e}$ *changes the conditional mean by*

$$\mathbb{E}\left[\left(1 - \frac{e}{\hat{e}}\right)(m-p) \,\middle|\, Z = z\right]. \tag{23}$$

If $\hat{e} \geq \underline{e} > 0$, *its absolute value is at most* $\|\hat{e} - e\|_{2,z}\|m - p\|_{2,z}/\underline{e}$.

Proof. Conditional on X, Z, the score mean is p. The conditional variance is $e^{-1}\{p(1-p) + (p-m)^2\} - (p-m)^2$. The law of total variance yields (22); its nonnegative final term proves the minimization claim. With $\hat{e}$, conditional expectation is $m + (e/\hat{e})(p-m)$. Subtract p, average over $X \mid Z$, and apply Cauchy–Schwarz to obtain (23) and its bound. □

Thus a good outcome model can reduce variance, but augmentation is not guaranteed to improve a poor prediction, and neither weighting nor topology removes the inverse-overlap factors. If a score box has radius b and $2b < \rho_\Psi$ for the true law on its coverage event, that law's signature is resolved, since every point in the box is within $2b$ of truth. The variance and range terms in (21) show both the quadratic radius scale and an additional finite-sample range penalty.

Estimated-propensity results use foldwise normal-score boxes with Bonferroni allocation. They are labelled nominal Wald results, not finite-sample certificates. Without independently justified nuisance-error bounds, (23) cannot be dropped from an exact guarantee. Training-estimated reference weights likewise change the reference target; their drift from the fixed-reference profile is reported separately.

D.2 Vector-score formulation and smooth targets

Adjusted law estimators. For observed W, write $e_{az}(w) = \mathbb{P}(A = a \mid W = w, Z = z)$ and let $V(Y)$ be the one-hot outcome vector, with conditional mean vector $m_{az}(w) = \mathbb{E}\{V(Y) \mid A = a, W = w, Z = z\}$. The standard augmented score is

$$T_{az}(O; \hat{m}, \hat{e}) = \hat{m}_{az}(W) + \frac{\mathbf{1}\{A = a\}}{\hat{e}_{az}(W)}\{V(Y) - \hat{m}_{az}(W)\}. \tag{24}$$

Its conditional expectation bias is exactly

$$\mathbb{E}\left[\left(1 - \frac{e_{az}(W)}{\hat{e}_{az}(W)}\right)(\hat{m}_{az}(W) - m_{az}(W)) \mid Z = z\right]. \tag{25}$$

We use two-fold cross-fitting and Euclidean projection onto the probability simplex. Correctly specifying either nuisance yields consistency in the fixed-dimensional setting with overlap; both-misspecified variants are reported. This is a finite-vector instance of established orthogonal and distributional estimation (Chernozhukov et al., 2018; Fawkes et al., 2023). A Gaussian embedding on distinct finite support points is injective: an exact embedding estimator can be inverted and passed through precisely the same transform. There is no information-theoretic dominance over such a comparator.

D.3 Double robustness and a smooth functional derivative

Condition on the fold used to train nuisance estimators. Under consistency, exchangeability and overlap, taking conditional expectations in (24) first given W, Z yields $\hat{m} + (e/\hat{e})(m - \hat{m})$. Subtracting m and integrating gives (25). If either nuisance is correct it vanishes. With a fixed finite outcome space, positive nuisance limits and consistent estimation of either model, the empirical fold averages converge to the correct vector. Simplex projection is nonexpansive and fixes the true law, so it preserves consistency. At a positive structural radius this also implies consistency of the hard signature. No such implication holds uniformly at radius zero.

For a differentiable Ψ, write $s_{az} = \Psi(\mu_{az})$, $\bar{s}_z = \sum_a q_a s_{az}$ and $J_{az} = D\Psi(\mu_{az})$. The derivative of the fixed-weight target with respect to the law vector is

$$g_{az} = 4\nu_r(z) q_d(a \mid z) J_{az}^\top (s_{az} - \bar{s}_z). \tag{26}$$

For a natural (rather than externally fixed) distribution of $W \mid Z$, an influence function for each law coordinate vector is

$$D_{az}(O) = \frac{\mathbf{1}\{Z = z\}}{\mathbb{P}(Z = z)} \left[\frac{\mathbf{1}\{A = a\}}{e_{az}(W)} \{V(Y) - m_{az}(W)\} + m_{az}(W) - \mu_{az} \right]. \tag{27}$$

The chain rule gives $\sum_{az} g_{az}^\top D_{az}(O)$ for the fixed-weight smooth target. Estimated reference masses add score terms. Fixed external standardization weights for W have a different tangent calculation. These formulas explain the smooth regime; no influence-function confidence interval is reported for the hard threshold target in this study.

E Validation selection and targeting priorities

E.1 Importance weighting and its invariance boundary

Under consistency and $Y(a) \perp A \mid Z$, conditional expectation of the observed weighted loss proves (12):

$$\sum_{a,z} \pi_{az} \frac{\omega_{dr}(a, z)}{\pi_{az}} \mathbb{E}\{\ell(f(a, z), Y) \mid A = a, Z = z\} = R_{dr}(f).$$

If an additional observed confounder X is needed, the appropriate density ratio must also standardize over its declared target distribution. The simple cell ratio is not valid after omitting an actual confounder.

For an unrestricted conditional model and a strictly proper loss, each conditional expected loss is uniquely minimized by the true cell law. A sum with strictly positive weights has exactly these same componentwise minimizers. For empirical categorical log likelihood, the objective in a cell is a positive multiplier times $-\sum_y N_{azy} \log f_{azy}$. Its simplex minimizer is $f_{azy} = N_{azy}/N_{az}$ when $N_{az} > 0$, independently of that multiplier. This finite-sample identity does not apply to an additive model with shared coefficients or to a fixed regularization penalty. The saturated baseline is therefore a meaningful capacity control. A separate direct-objective baseline may have zero weights in irrelevant cells, which removes uniqueness there without invalidating its external-risk objective.

The training objective contains no Ψ. Changing the descriptor while keeping counts, weights, loss and numerical procedure fixed leaves every candidate unchanged. Computing Ψ after prediction does not turn importance-weighted likelihood into a topology-specific learning algorithm.

Proposition 7 (Validation selection for a fixed objective)**.** *Condition on a training sample and let $f_1, \dots, f_K$ be the resulting fixed candidates. Validation observations are independent and follow a known positive allocation π. Let $0 \le \ell \le 1$, and let ω_V be a fixed probability mass over cells. Define*

$$B_V = \max_{az} \frac{\omega_V(a, z)}{\pi_{az}}, \qquad \Gamma_V = \sum_{az} \frac{\omega_V(a, z)^2}{\pi_{az}}.$$

For m validation observations, set $t = \log(2K/\delta)$ and $\epsilon_V = \sqrt{2\Gamma_V t/m} + 2B_V t/(3m)$. If $\hat{k}$ minimizes the validation importance-weighted mean loss, then with probability at least $1 - \delta$,

$$R_V(f_{\hat{k}}) \le \min_{k \le K} R_V(f_k) + 2\epsilon_V. \tag{28}$$

For Q declared objectives replace t by $\log(2KQ/\delta)$ to obtain their joint assertion.

Proof. Each validation summand is in $[0, B_V]$. Its second moment is at most $\sum_{az} \pi_{az}(\omega_V/\pi_{az})^2 = \Gamma_V$, and hence its variance is at most Γ_V. Bernstein's inequality and a union bound over candidates give $|\widehat{R}_V(f_k) - R_V(f_k)| \le \epsilon_V$ for all k. Compare the selected candidate with a population minimizer in the finite dictionary and apply this error bound twice. A further union bound gives the result for Q objectives. Integrating over the training sample preserves the guarantee. □

The half-Brier loss used for validation meets the boundedness assumption. The result does not claim optimality outside the candidate dictionary, nor validate selecting on test data. No refit follows selection in this experiment. If one refits on training plus validation data, independent test evaluation remains valid, but this exact fixed-dictionary bound does not automatically describe the refitted estimator. The factor Γ_V exposes the cost of prioritizing poorly observed cells; for equal-cell targets it equals $\sum_{az}(JS)^{-2}/\pi_{az}$.

E.2 Estimated reference masses

An independently fitted reference $(\hat{h}, \hat{p})$ defines a random scientific target. Conditional on that training sample, the joint confidence event covers its entire profile. This does not establish coverage for the profile defined by the unknown population (h, p). One valid alternative is joint confidence projection over both law and reference-mass sets. A useful deterministic error bound is as follows. Suppose all masses are positive and

$$\max_z |\log \hat{h}_z - \log h_z| \le e_h, \qquad \max_{az} |\log \hat{p}_{a|z} - \log p_{a|z}| \le e_p.$$

Integrating the softmax Jacobian along the segment between log masses gives $\|\hat{\nu}_r - \nu_r\|_1 \le 2re_h$ and $\|\hat{q}_d(\cdot \mid z) - q_d(\cdot \mid z)\|_1 \le 2de_p$ for $d, r \ge 0$. The product-measure triangle inequality therefore gives

$$|\Theta_{\hat{h},\hat{p}}(d, r) - \Theta_{h,p}(d, r)| \le D_\Psi^2 \min\{1,\ 2re_h + 4de_p\}. \tag{29}$$

The bound follows by first changing ν, then the two factors of q; the truncation uses the target range $[0, D_\Psi^2]$. Combine this event with law confidence using an allocated error budget. Near-zero probabilities may make it vacuous, correctly reflecting overlap and reference uncertainty. The implemented estimated-allocation ablation concerns training; the reported profile certificates use fixed design reference masses and need no estimated-weight correction.

E.3 Transfer between empirical covariate pools

Because validation and test use disjoint empirical covariate pools, the finite-dictionary guarantee from the iid finite-cell study does not automatically bound their population-risk difference. If $\sup_{f \in \mathcal{F}} |R_{\text{test}}(f) - R_{\text{val}}(f)| \le \Delta$, and validation errors are uniformly at most ε, then

$$R_{\text{test}}(\hat{f}) \le \min_{f \in \mathcal{F}} R_{\text{test}}(f) + 2\varepsilon + 2\Delta.$$

Indeed, compare test to validation at $\hat{f}$, apply the uniform validation-selection inequality, and compare validation to test at its test-risk minimizer. We assume no unproved small Δ for these pools. The benchmark measures transfer across their covariate rows empirically; resampling more virtual units does not eliminate this source of variation.

F Certification experiment protocol

F.1 Implementing certificates and testing the transition

For the power experiments we replace the Hoeffding boxes by simultaneous Clopper–Pearson boxes (Clopper & Pearson, 1934), with each one-sided tail $\alpha/(2MC)$. They retain the same $1 - \alpha$ coverage event. Their power need not wait until the conservative Hoeffding sufficient threshold. For each active pattern, box–simplex feasibility reduces to coordinate bound compatibility and lower/upper sums straddling one. We retain closures of inactive faces, which can only enlarge uncertainty at threshold ties. The confidence set is projected onto signatures rather than requiring a unique active pattern. A zero upper bound *certifies equality*; a positive lower bound certifies a change. Otherwise the result is inconclusive.

Conditional on the observed cell counts, the probability that a CP upper bound falls below c has an exact binomial expression. A union bound over a sufficient collection of such inequalities predicts a lower bound on complete signature-certification power (Appendix F.2). These predictions use the known simulation laws only for evaluation. An audit changes the oracle laws while holding counts fixed and verifies that the estimated bands do not change.

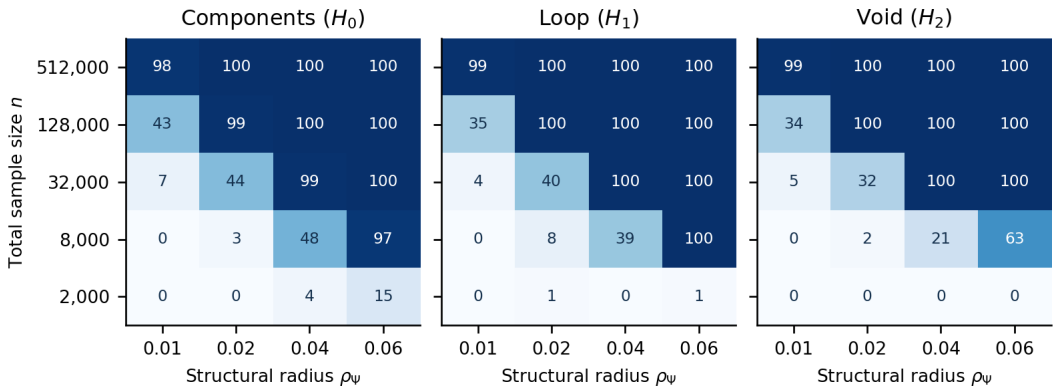


Figure 3: Measured complete-signature certification percentages at the fixed slice $\eta = \epsilon = .2$ (100 datasets per entry). Increasing sample size or structural radius moves the procedure from inconclusive to certified. The complete factorial grid, including rare cells with probability .0025, is shown in Figure 4 and supplied as condition-level CSVs.

Factorial design. The factorial design crosses three cone geometries (components, loop, void), five sample sizes $n \in \{2,000, 8,000, 32,000, 128,000, 512,000\}$, three context prevalences $\eta \in \{.05, .2, .5\}$, three rare-arm propensities $\epsilon \in \{.05, .2, .5\}$, four radii $\rho \in \{.01, .02, .04, .06\}$, and two sides of the threshold, with 100 repetitions each. Boundary controls add 13,500 datasets, giving 121,500 in total. The rare cell has center mass $c \pm \rho$, $c = .08$; all other cells have center mass .23. Outer mass is uniform. Linear programming independently verifies every declared radius. All sample sizes and conditions are reported; no optimal cell, threshold, or seed is selected from the results.

Results. At $\eta = \epsilon = .2$ and $\rho = .02$, the H_2 complete-certification rates are $0\%, 2\%, 32\%, 100\%, 100\%$ across increasing n; the corresponding binomial sufficient-power bounds average $0\%, 0\%, 32.8\%, 99.3\%, 100\%$. The H_0 and H_1 geometries exhibit the same qualitative transition. These are Monte Carlo rates, so a measured proportion can fall below its population lower bound by sampling variation. Over all 108,000 nonboundary datasets, observed center-certification probability minus its exact conditional prediction averages -0.000535, with Monte Carlo standard error 0.000678. No curve was fitted to obtain this agreement.

Joint profile coverage is 99.9885%; 14 profiles are not covered. There are 13 false positive structural certificates among 67,500 ordinary/boundary controls, all at the boundary. The very high coverage reflects conservative simultaneous boxes, not a claim of calibrated 100% certainty. The Hoeffding construction is also executed: at the same H_2 slice and $n = 128,000$, it certifies only 3%, compared with 100% for CP boxes. Thus the information scale is useful, but the generic sufficient constants are not an accurate finite-sample power curve. Both procedures estimate the same structural target.

The sufficient regime in (4) contains 9,600 datasets; both constructions certify every profile there correctly. At the opposite end, $n = 2,000$, $\eta = \epsilon = .05$, and $\rho = .01$ give $K_n = .01417$. Theorem 3 bounds every level-.05 procedure's structural power by 13.42%. None of the 300 structural datasets across the three geometries receives a positive certificate. The necessary interval-width probability is at least 81.58%; 299 of those 300 intervals are at least as wide as the two-point gap. That gap is .25 at $(d, r) = (0, 0)$ and .00475 at $(1, 1)$. These are different target uncertainty costs under the same lack of observations, not different amounts of statistical information.

F.2 Exact conditional power of the implemented certificate

Let $K \sim \text{Binomial}(N, p)$ and use the one-sided CP tail $t = \alpha/(2MC)$. Write $F_{N,c}$ for the binomial CDF and set

$$k_*(N, c, t) = \max\{k \in \{-1, 0, \ldots, N\} : F_{N,c}(k) < t\}, \qquad F_{N,c}(-1) = 0.$$

The CP upper limit is strictly below c exactly when $K \le k_*$. Consequently its exact conditional probability is

$$B_-(N, p; c, t) = F_{N,p}(k_*), \qquad B_+(N, p; c, t) = B_-(N, 1-p; 1-c, t). \tag{30}$$

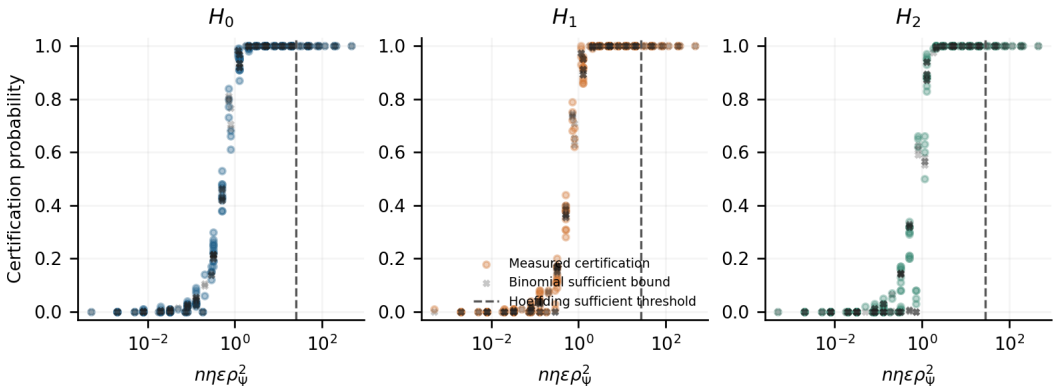


Figure 4: Every structural-side factorial condition, plotted against the prespecified information product $n\eta\epsilon\rho_\Psi^2$. Points show measured full-signature certification and the exact-binomial sufficient power lower bound averaged over observed counts. The dashed line is the generic Hoeffding sufficient threshold $4\ell_\alpha$, not a fitted transition or a sharp necessary threshold. Monte Carlo variation, finite probability variances, and outer-vertex constraints prevent exact collapse of all conditions onto one curve.

The second expression gives the probability of a CP lower limit strictly above c. Empty cells certify neither inequality. At boundary $p = c$, these one-sided false certificates have probability at most t, including the discreteness of the binomial distribution.

For a cone with center below threshold in the rare treated cell, one sufficient collection of inequalities is: the three reference-cell centers are above c, the rare treated center is below c, and each rare treated outer vertex is above c. Denote the latter center-test power by B_* and the remaining elementary powers by B_j. Conditional on the four sample counts, a union bound gives

$$\mathbb{P}\{\text{all required inequalities hold} \mid \{N_{az}\}\} \geq \max\{0, B_* - \sum_j (1 - B_j)\}. \tag{31}$$

This is a sufficient lower bound on complete correct signature certification, not an equality or a bound based on independent coordinates. For the ordinary side, it suffices that all four centers are above c; replace the rare center power by B_+. The algebra also defines a conservative sufficient event at the boundary, where complete power cannot be uniform in nearby laws. Averaging (31) over the random cell counts gives an unconditional lower bound. The study records its realized-count conditional value on every dataset, together with the observed center decision. The prediction is never used to narrow the fitted confidence set.

The practical CP implementation uses the same tail for every coordinate within a dataset, not a tail divided by the number of Monte Carlo repetitions. Counts within an outcome multinomial are dependent; the confidence union bound and (31) do not require otherwise. The positive-only intersection–union procedure in Appendix L.1 targets a different inferential task. This experiment uses full signature projection to certify equality as well as change; its coverage guarantee does not imply greater power than every positive-only test.

G Oscillator experiment protocol

G.1 A mechanistic structural example: stochastic oscillations

Consider the noisy Hopf normal form (Doan et al., 2018), with rotational matrix $J_0 = \left(\begin{smallmatrix} 0 & -1 \\ 1 & 0 \end{smallmatrix}\right)$:

$$dX_t = \{(\mu - \|X_t\|^2)X_t + J_0 X_t\}\, dt + \sqrt{2D}\, dB_t. \tag{32}$$

We use independent stationary endpoint experiments, not correlated points from one trajectory. The invariant density is proportional to $\exp\{(\mu\|x\|^2/2 - \|x\|^4/4)/D\}$, so $U = \|X\|^2 \sim N(\mu, 2D)$ truncated to $U \geq 0$, with independent uniform angle. This gives exact endpoint probabilities without

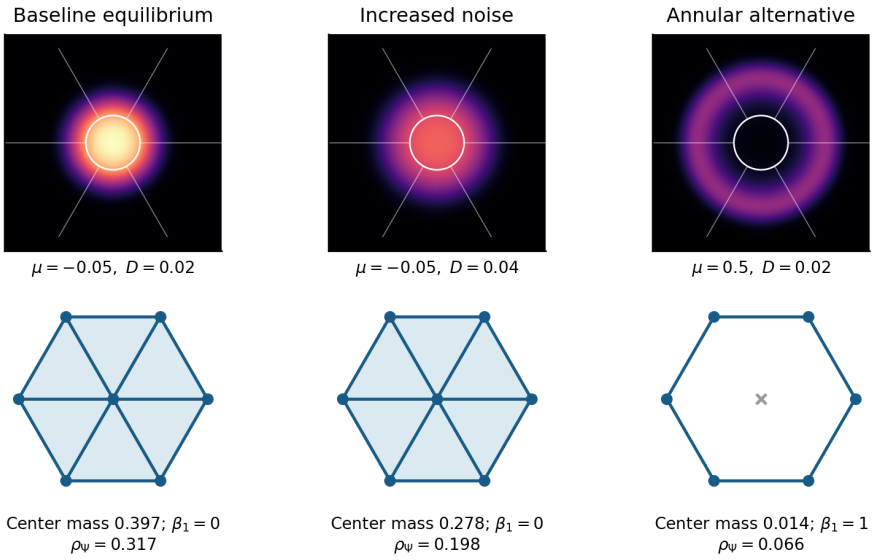


Figure 5: Exact model densities above and their prespecified occupied complexes below. Increased noise redistributes mass but retains the filled complex. The annular alternative removes its center and creates a hole. The plot is calculated from the model, not a selected empirical sample.

time-discretization error. The baseline has $(\mu, D) = (-.05, .02)$. The ordinary intervention doubles noise to $D = .04$ at the same negative μ; the structural alternative sets $\mu = .5$ and keeps $D = .02$. All coordinate means remain zero.

Before simulation, we fix a central disk of radius .3, six angular sectors outside it, their wheel complex, and occupancy threshold $c = .08$. Each sector is adjacent to the center and its two neighboring sectors; the corresponding triangles are filled. The exact center probabilities are .39704, .27809, and .01406 for baseline, ordinary, and structural settings. The first two occupied complexes are contractible; the third is a cycle with one H_1 class (Figure 5). This is a regime of the specified occupancy representation. Every continuous stationary law still has support $\mathbb{R}^2$, and the claim does not concern support topology, random attractors, or Lyapunov exponents.

The intervention acts only in a context of prevalence $\eta = .2$, with conditional treatment probability $\epsilon = .1$; the common context has allocation .5 and no effect. We use 200 independent datasets for each of three scenarios and $n \in \{4,000, 20,000, 80,000\}$. Counts are sampled directly from the exact multinomial endpoint law, an exact implementation of the iid categorical experiment. General distributional comparators are a Gaussian MMD on the same seven vertices, with fixed bandwidth one, and the full categorical likelihood-ratio statistic. They share 499 treatment permutations within the rare context. There is no change of data or information budget between representations. Neither test is labeled a test of structural equality.

At the prespecified primary sample size $n = 80,000$, both distributional tests detect both alternatives in all 200 datasets. GeoRARE certifies a zero profile for the ordinary intervention and a positive profile for the structural intervention in all 200. A proportion of $200/200$ has two-sided exact 95% lower limit 98.17%, not certainty of perfect future performance. At $n = 20,000$, structural certification succeeds in 92.5% of annular datasets; complete membership recovery succeeds in only 38.5%. This is the practical benefit of resolving the chosen signature rather than every occupied vertex. At $n = 4,000$, distributional detection of the annular alternative is already 100%, while structural certification is 0%: rejecting distributional equality supplies insufficient information for that stronger scientific claim.

All sensitivity thresholds .06, .08, .10 are reported in the appendix using the same probability event. The full escort grid is also reported. For the structural alternative, $\Theta(0,0) = .25$ and $\Theta(1,1) = .036$. Changing d, r changes these target magnitudes; it does not alter the rare-cell count or manufacture

Scenario, n	MMD	LR	Positive	Zero
Null, 4,000	3.0	3.0	0.0	100.0
Ordinary, 4,000	23.0	26.5	0.0	97.0
Structural, 4,000	100.0	100.0	0.0	0.0
Null, 20,000	7.0	3.5	0.0	100.0
Ordinary, 20,000	89.5	97.5	0.0	100.0
Structural, 20,000	100.0	100.0	92.5	0.0
Null, 80,000	6.0	5.5	0.0	100.0
Ordinary, 80,000	100.0	100.0	0.0	100.0
Structural, 80,000	100.0	100.0	100.0	0.0

Table 3: Oscillator experiment at the prespecified occupancy threshold $c = .08$ (200 independent datasets per row). MMD and LR reject equality of the full categorical distributions. Positive and Zero certify, respectively, a nonzero or exactly zero GeoRARE profile using the same probability confidence event. They test different scientific statements. All percentages are measured, not theoretical guarantees.

extra certification power. This example is a mechanistic simulation, with simulated rather than biological observations.

G.2 Exact stationary oscillator law and prespecified coarsening

Define $V(x) = \|x\|^4/4 - \mu\|x\|^2/2$. The drift in (32) is $-\nabla V(x) + J_0 x$, and the diffusion generator is $D\Delta$. For $f(x) \propto \exp\{-V(x)/D\}$, the gradient-drift and diffusion terms cancel in the stationary Fokker–Planck equation. Moreover, $\operatorname{div}(J_0 x) = 0$ and $J_0 x \cdot \nabla f(x) = 0$, so the rotational term also has zero divergence. Quartic confinement makes the density integrable. In polar coordinates, with $u = r^2$, the Jacobian gives $r\,dr = du/2$ and

$$\frac{\mu u/2 - u^2/4}{D} = -\frac{(u-\mu)^2}{4D} + \frac{\mu^2}{4D}.$$

After normalization, U has a $N(\mu, 2D)$ distribution truncated below zero, independent of a uniform angle. With disk radius $r_0 = .3$, its central probability is exactly

$$p_c(\mu, D) = \frac{\Phi\{(r_0^2 - \mu)/\sqrt{2D}\} - \Phi(-\mu/\sqrt{2D})}{1 - \Phi(-\mu/\sqrt{2D})}. \tag{33}$$

Each outer sector has probability $(1 - p_c)/6$. Numerical quadrature of the radial density agrees with (33) to below 10^{-10} for all declared interventions. The sample generator draws categorical counts with these probabilities; it does not substitute the population probabilities for empirical counts during estimation.

At $c = .08$, all six outer sectors are active in each population. The baseline and noise intervention also retain the center, making the wheel a cone and therefore contractible. The annular alternative removes the center and leaves precisely a six-cycle. Its squared H_{012} distance from baseline is one. In the two-arm, two-context design, only the rare context changes, so its GeoRARE profile is exactly $2\nu_r(1)q_d(0 \mid 1)q_d(1 \mid 1)$. A characteristic Gaussian kernel on the seven distinct representative vertices is injective on categorical laws, so its distributional test can detect both interventions. Applying the same confidence-set-to-signature map after a full-law estimator yields the same structural capability. The example establishes the value of the scientific target and certificate, not unique information in a topological embedding.

G.3 Threshold sensitivity, membership recovery, and rarity degrees

Scenario, n, c	MMD	LR	Positive	Zero	Signature	Membership
null, 4,000, 0.06	3.0	3.0	0.0	100.0	100.0	0.0
null, 4,000, 0.08	3.0	3.0	0.0	100.0	100.0	0.0
null, 4,000, 0.10	3.0	3.0	0.0	99.5	99.5	0.0
null, 20,000, 0.06	7.0	3.5	0.0	100.0	100.0	0.0
null, 20,000, 0.08	7.0	3.5	0.0	100.0	100.0	0.0
null, 20,000, 0.10	7.0	3.5	0.0	100.0	100.0	0.0
null, 80,000, 0.06	6.0	5.5	0.0	100.0	100.0	99.0
null, 80,000, 0.08	6.0	5.5	0.0	100.0	100.0	0.0
null, 80,000, 0.10	6.0	5.5	0.0	100.0	100.0	0.0
ordinary, 4,000, 0.06	23.0	26.5	0.0	99.5	99.5	0.0
ordinary, 4,000, 0.08	23.0	26.5	0.0	97.0	97.0	0.0
ordinary, 4,000, 0.10	23.0	26.5	0.0	88.0	88.0	0.0
ordinary, 20,000, 0.06	89.5	97.5	0.0	100.0	100.0	46.5
ordinary, 20,000, 0.08	89.5	97.5	0.0	100.0	100.0	0.0
ordinary, 20,000, 0.10	89.5	97.5	0.0	100.0	100.0	0.0
ordinary, 80,000, 0.06	100.0	100.0	0.0	100.0	100.0	100.0
ordinary, 80,000, 0.08	100.0	100.0	0.0	100.0	100.0	91.5
ordinary, 80,000, 0.10	100.0	100.0	0.0	100.0	100.0	0.0
structural, 4,000, 0.06	100.0	100.0	0.0	0.0	0.0	0.0
structural, 4,000, 0.08	100.0	100.0	0.0	0.0	0.0	0.0
structural, 4,000, 0.10	100.0	100.0	0.0	0.0	0.0	0.0
structural, 20,000, 0.06	100.0	100.0	95.0	0.0	95.0	95.0
structural, 20,000, 0.08	100.0	100.0	92.5	0.0	92.5	38.5
structural, 20,000, 0.10	100.0	100.0	10.0	0.0	10.0	0.0
structural, 80,000, 0.06	100.0	100.0	100.0	0.0	100.0	100.0
structural, 80,000, 0.08	100.0	100.0	100.0	0.0	100.0	100.0
structural, 80,000, 0.10	100.0	100.0	100.0	0.0	100.0	0.0

Table 4: All oscillator thresholds and the complete-membership ablation. All threshold bands use the same simultaneous probability event. Signature certification can succeed without resolving all occupied vertices.

d, r	True effect	Mean lower	Mean upper	$L > .01$ (%)
0, 0	0.250000	0.231250	0.250000	92.5
0, 0.5	0.166667	0.154167	0.166667	92.5
0, 1	0.100000	0.092500	0.100000	92.5
0, 2	0.029412	0.027206	0.029412	92.5
0.5, 0	0.187500	0.173437	0.187500	92.5
0.5, 0.5	0.125000	0.115625	0.125000	92.5
0.5, 1	0.075000	0.069375	0.075000	92.5
0.5, 2	0.022059	0.020404	0.022059	92.5
1, 0	0.090000	0.083250	0.090000	92.5
1, 0.5	0.060000	0.055500	0.060000	92.5
1, 1	0.036000	0.033300	0.036000	92.5
1, 2	0.010588	0.009794	0.010588	92.5
2, 0	0.012046	0.011143	0.012046	92.5
2, 0.5	0.008031	0.007429	0.008031	0.0
2, 1	0.004819	0.004457	0.004819	0.0
2, 2	0.001417	0.001311	0.001417	0.0

Table 5: All escort settings for the oscillator structural alternative at $n = 20,000$. Lower and upper are averaged endpoints, not a confidence interval for a mean across repetitions. Changing d,r changes the target, not the number of observed rare-cell outcomes.

H PROFILE AND TARGETING EXPERIMENT PROTOCOL

This design supports two separate questions: certification of signed profile contrasts and validation of training priorities. The profiling and learning experiments use independently sampled datasets from the same fixed mechanism.

Table 6: Paired escort-minus-ERM excess half-Brier differences $\times 1000$. Negative favors escort selection. Intervals are 95% paired t intervals over 50 independent replications.

Objective	Mean difference	95% interval	Holm-adjusted p
natural	0.005	[-0.001, 0.010]	0.0883
rare context	-0.753	[-0.833, -0.673]	$< .001$
rare treatment	-0.570	[-0.680, -0.460]	$< .001$
balanced	-0.169	[-0.215, -0.123]	$< .001$

H.1 Shared mechanism and sampling design

The reference context masses are $(.55, .25, .15, .05)$ and the propensity rows are

$$p_{a|z} = \begin{pmatrix} .65 & .25 & .08 & .02 \\ .50 & .30 & .15 & .05 \\ .45 & .30 & .20 & .05 \\ .50 & .30 & .15 & .05 \end{pmatrix}.$$

For signatures 1, 2 and 3, the outcome templates are respectively

$$v_1 = (.16, .15, .14, .13, .14, .14, .14),$$
$$v_2 = (.20, .18, .05, .05, .16, .17, .19),$$
$$v_3 = (.18, .17, .05, .17, .05, .18, .20).$$

With zero-based $z, a \in \{0, 1, 2, 3\}$, transfer $\varepsilon_{za} = .004(a - 1.5) + .002(z - 1.5)$ from the last coordinate to the first. This preserves normalization and the intended signatures. Outcome counts are sampled from the joint multinomial on 112 (z, a, y) categories, exactly equivalent to the iid unit-level experiment. The population reversal is a designed witness of the specified mechanism; its sign pattern is not chosen from the Monte Carlo outcomes.

The master seed is 202609130000. Replication $j = 0, \ldots, 49$ uses seeds $202609130000 + 100j + (1, 2, 3)$ for training, validation and test. Profiling uses $202609130000 + 10^6 + 10^4 i + j$, where $i = 0, \ldots, 3$ indexes the four sample sizes and $j = 0, \ldots, 499$. All 2,150 sample seeds are distinct. Source hashes and the protocol were recorded before evaluation.

H.2 Targeting experiment protocol

The additive model uses eight binary features (four context and four treatment indicators), a seven-class softmax, and penalty $\lambda \|B\|_F^2/2$ with $\lambda \in \{0, .001, .01\}$. Each of the known- and estimated-allocation escort families has 147 candidates. *Escort / validation* is the GeoRARE method: the q_d, ν_r priority family is selected by independent validation over (d, r) and regularization. Direct objective weighting is a separate baseline using the external target directly. Overlap weighting targets $\omega_{az}^{\mathrm{ov}} \propto h_z / \sum_b 1/p_{b|z}$ uniformly across arms within a context. Group DRO solves the convex epigraph problem $\min_{B,t} t + \lambda \|B\|_F^2/2$ subject to every empirical cell log loss being at most t. It is a finite additive-softmax comparator, not a reproduction of a neural benchmark. The saturated T-learner baseline, which uses no GeoRARE weights, uses $(N_{azy} + \kappa)/(N_{az} + 7\kappa)$, $\kappa \in \{.1, .5, 1, 5, 20\}$, selected by validation.

All four evaluation objectives are fixed probability masses. Natural weight is π; rare-context weight is $1/8$ on $z \in \{2, 3\}$; rare-treatment weight is $1/8$ on $a \in \{2, 3\}$; balanced weight is $1/16$ on every cell. The conditional ordering of treatment frequencies agrees with these declared labels in this DGP. The targets do not change after a method is selected. Observed validation/test scores use the known sampling ratio ω_V/π. The primary excess half-Brier score is $\frac{1}{2} \sum_{az} \omega_V(a, z) \|\hat{u}_{az} - u_{az}\|_2^2$. Truth enters this final evaluator after all candidate selections are fixed. This is not an oracle-selected risk minimum. TV and hard-signature error are secondary diagnostics, and the latter are not certificates.

H.3 Profile-certification protocol and uncertainty

Figure 2 compares complete-signature resolution with the signed-profile task on the same 500 datasets at each sample size.

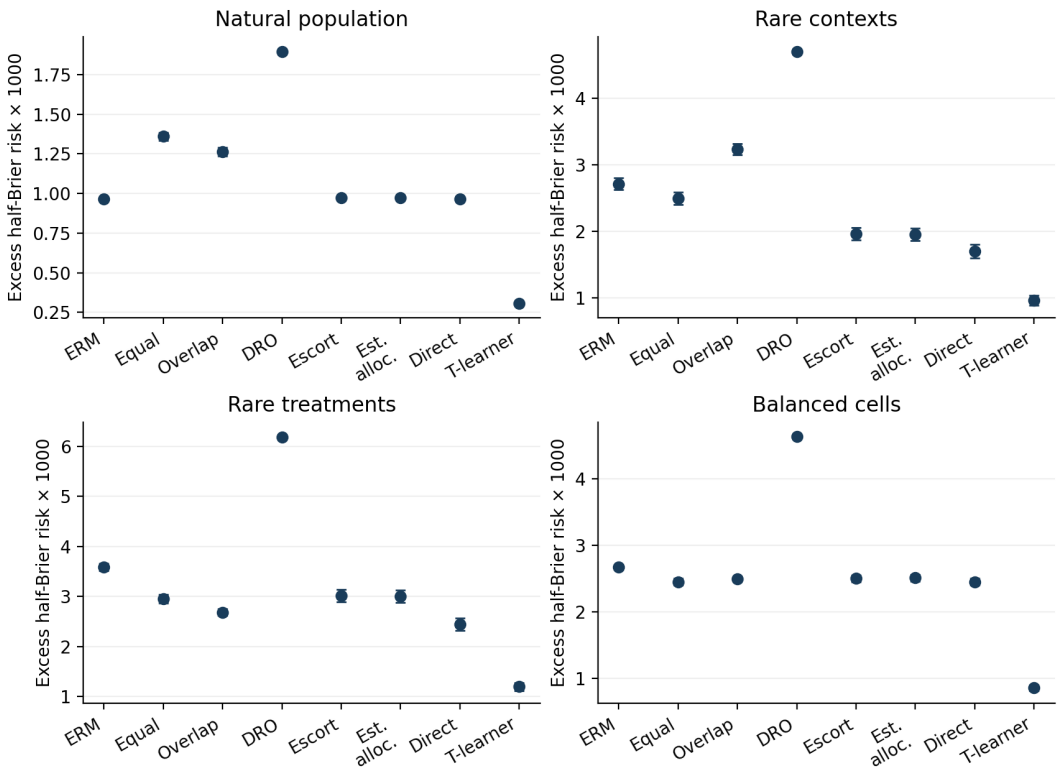


Figure 6: All methods, with 95% Monte Carlo intervals for mean excess half-Brier risk across 50 independent replications. The saturated T-learner and direct objective weighting are separate baselines; both outperform the escort family on the reported mean risks. Intervals describe variation across this DGP's replications.

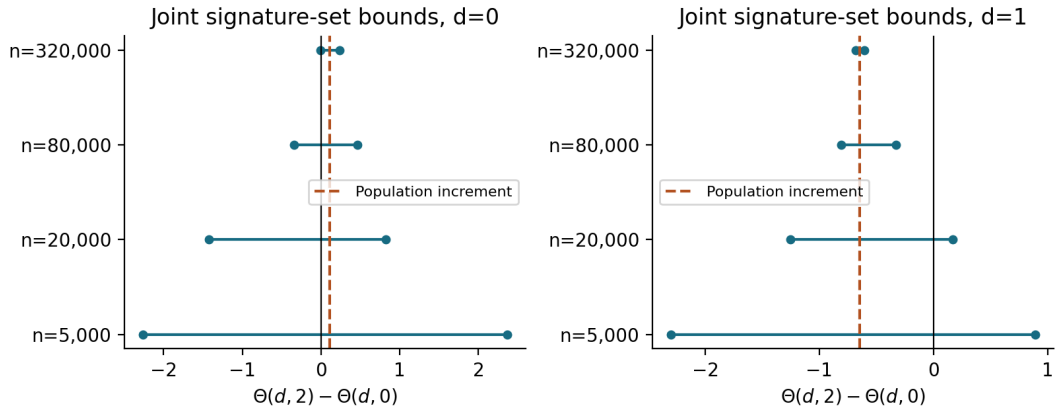


Figure 7: Signed increment intervals for replication zero at every sample size. The displayed replication is fixed in advance; no successful replication was selected for this figure.

The underlying all-coordinate box coverage rates are $97.6\%, 96.8\%, 95.6\%, 93.8\%$. The final empirical proportion is below the nominal lower guarantee, but its Monte Carlo exact 95% interval $[91.3\%, 95.7\%]$ includes 95%; a finite Monte Carlo proportion need not exceed its population lower bound. The coarser signature projection covers all true signatures in these runs. For $500/500$ coverage, the two-sided exact lower confidence limit is 99.265%, not certainty of future coverage. The profile study does not include an additional reversed or flat-surface null DGP, so its zero opposite-reversal count alone is not an empirical false-discovery guarantee over such alternatives. Coverage

Table 7: Profile-certification results, 500 datasets per row. Widths are mean outer profile-band widths.

n	Reversal (%)	All signatures (%)	Width $(0,0)$	$(1,1)$	$(2,2)$
5,000	0.0	0.0	6.026	3.697	1.6324
20,000	0.0	0.0	3.015	0.939	0.2124
80,000	0.0	0.0	1.299	0.161	0.0100
320,000	29.0	0.4	0.365	0.022	0.0006

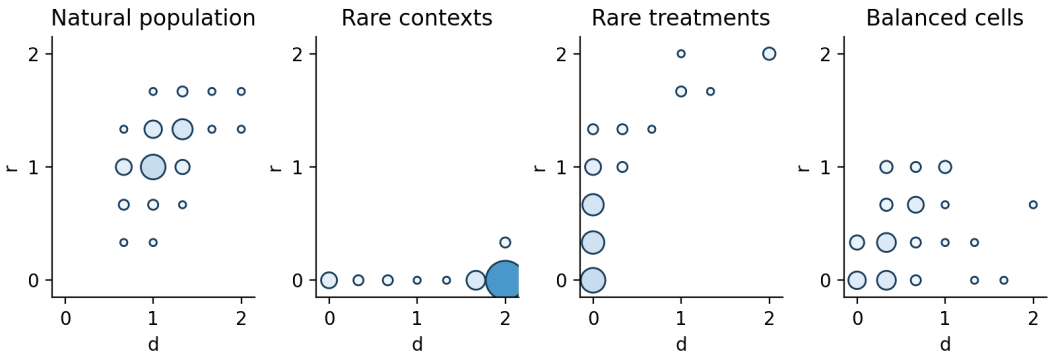


Figure 8: Validation selections across all 50 replications. Bubble area is proportional to selection frequency. Different fixed objectives favor different degree regions; no selected point uses oracle or test outcomes.

is justified by the probability construction; the factorial certification experiment supplies separate ordinary and boundary controls (Appendix F).

Mathematical checks compare the variance identity, analytic derivatives against finite differences, likelihood gradients, empty-cell behavior, point-box recovery and importance-risk identities. An independent audit replays sample counts and validation selections and verifies source hashes. Fitting took approximately 69 CPU seconds and profiling approximately 4 seconds on the recorded environment, excluding radius linear programs, audits, summaries and document generation. The small finite state space explains this speed.

I ACIC SEMI-REAL BENCHMARK

I.1 DATA, POPULATIONS, AND DESIGN

We downloaded the original `input_2016.RData` object from `vdorie/aciccomp`, commit `282d26659b2d3d6fd060dde6d32feeb8f1e8ab5a`. The authors' documentation describes 4,802 rows and 58 de-identified covariates from the Collaborative Perinatal Project. The release declares GPL version 2 or later; its source documentation, original object, CSV conversion, and byte hashes accompany the experiment. This is a four-arm, seven-state distributional extension using the real covariates, not a reproduction of the official binary-ATE competition (Dorie et al., 2019). We assign no clinical meaning to the simulated arms or outcomes and do not reconstruct the de-identified covariate identities.

For each of 50 independently seeded splits, original row identifiers are partitioned into 2,881 training, 960 validation, and 961 test rows before preprocessing. All generated copies of a row remain in its assigned pool. Numeric centering, scaling, categorical levels, and feature eligibility are learned on training rows. Values are clipped at five training standard deviations. There are no missing values in the release. The first five training-eligible numeric columns with at least 20 distinct values are `x_1`, `x_3`, `x_4`, `x_5`, and `x_6`. All column choices are recorded for every split.

Four contexts use training quantiles $.55, .80, .95$ of the standardized first chosen variable, with an outcome-independent 10^{-6} row-level tie break. Unequal context frequency is deliberately constructed. The four nonlinear coordinates are $\tanh(x_1)$, $\sin(x_2)$, $\mathrm{clip}(x_3, -3, 3)$ $\mathrm{clip}(x_4, -3, 3)$, and $\mathbf{1}\{x_5 >$

$0\}$, where the subscripts here index the five selected standardized columns. Together with four context indicators these form $g(x, z)$. The coefficient seeds are fixed across the entire experiment.

Treatment propensities satisfy

$$e_a(x, z) = \epsilon + (1 - 4\epsilon)\,\mathrm{softmax}_a\{\log p^{\mathrm{base}}_{a|z} + 1.25 B_a^\top g_{1:4}(x, z)\}. \tag{34}$$

Here $p^{\mathrm{base}}_{a|z}$ denotes the fixed baseline allocation matrix specified in Appendix H; it is distinct from the propensity $e_a(x, z)$ after the positivity-floor mixture and covariate-dependent modification. The randomized control omits the nonlinear term. The confounded condition includes it. Both obey measured-covariate exchangeability by construction; there is no hidden-confounding claim. We cross $\epsilon \in \{.005, .02, .05\}$ with

$$p_a(x, z) = (1 - \tau)v_{s(a,z)} + \tau\,\mathrm{softmax}\{C_a^\top g(x, z)\}, \qquad \tau \in \{0, .15, .30\}. \tag{35}$$

The v templates are specified in Appendix H; $s(a, z)$ is the component-count matrix in (13). A further no-effect control uses v_1 and the same C_0 for all arms, at $\tau = .30$. It therefore has identical arm laws within every context. No condition is removed when a population signature changes or disappears. The path and thresholds $.08, .10, .12$ are fixed, with $.10$ primary.

Each split crosses two assignment mechanisms, three positivity floors, four response conditions, and $n_{\mathrm{train}} \in \{5{,}000, 20{,}000\}$: 48 cases per split and 2,400 cases in total. Each case also uses 10,000 validation draws, 20,000 independent predictive-test draws, and independent profiling samples of 20,000 and 320,000 draws. Sampling is iid with replacement from each declared finite covariate pool. Multinomial count arrays over row, arm, and response are exact sufficient representations of those draws. These budgets count simulated observations, not additional real patients.

Within context, every arm is standardized over the same reference covariate distribution before applying Ψ. Training-pool covariates and the known assignment define fixed reference h_z and $p_{a|z}$. The external natural target is $h_z p_{a|z}$; rare-context, rare-treatment, and balanced masses follow the finite study. Evaluation within each split uses its own common $X \mid Z$ distribution, with context-density correction. The relevant importance ratio is $\omega(a, z)/(h_z^{\mathrm{eval}} e_a(x, z))$, not the 16-cell allocation ratio when e_a varies inside a context.

I.2 Learners, validation, and comparators

The S-learner uses all encoded covariates, the four nonlinear coordinates, 16 fixed random Fourier features, context-by-arm intercepts, and arm-by-nonlinear interactions. Its conditional multicategory scores minimize one-half Brier loss plus an ℓ_2 penalty. We solve the resulting ridge normal equations and project predictions onto the simplex, followed by a common 10^{-8} numerical probability floor. Projection cannot increase squared distance to a one-hot outcome. The model has explicit treatment interactions; it differs from the deliberately additive model in Appendix H. The T-learner fits separate coefficients for every arm using the full nonlinear feature basis. These are capable finite-feature models, not claims of state-of-the-art performance over arbitrary function classes.

The degree grid is $\{0, 1, 2\}^2$, and the short-path analysis uses the common penalty grid $\{10^{-4}, 10^{-3}, 10^{-2}\}$. Baselines are observational ERM, equal-cell weights, standardized natural weights, generalized overlap weights, direct external-objective weights, empirical Group DRO over observed (Z, A) groups, its inverse-propensity-standardized variant, and the flexible T-learner. Group DRO optimizes all nonempty declared groups and records any empty group. Its convex quadratic objective has a computable primal–dual gap. An objective-preserving numerical refinement is applied to every gap above 2×10^{-6}; outputs before and after numerical refinement are archived. No data, loss, penalty, scenario, or selection rule changes in that refinement.

Regularization sensitivity analysis. The short penalty grid binds at its upper endpoint for 93–100% of choices, depending on method. This validation diagnostic motivated a wider-grid analysis of the same complete 2,400-case matrix, including penalties $.1, 1, 10, 100$ for every method. The uniform $1/7$ predictor, which is the limiting ridge-plus-simplex predictor as the penalty tends to infinity, is also eligible for every method and reported separately. Both paths use identical training observations. The wider-path analysis uses 10,000 independently generated validation responses and 20,000 independently generated test responses per case, with seeds disjoint from the short-path

Method	Natural	Rare context	Rare treatment	Balanced
Natural ERM	4.977	8.700	7.902	6.983
Escort / known e	6.751	12.309	10.826	9.651
Equal cell	15.114	24.309	21.749	19.775
Generalized overlap	10.561	17.175	15.548	14.051
Observed Group DRO	5.728	9.819	9.030	8.131
Standardized Group DRO	7.054	12.011	10.863	9.889
Direct objective	7.007	43.757	38.192	19.775
Flexible T-learner	14.030	29.418	37.897	25.575
Escort / estimated e	8.094	16.520	14.630	12.535
Escort / estimated train + validation	8.394	19.739	16.942	14.406

Table 8: Short regularization grid in the ACIC-covariate benchmark: population excess half-Brier risk $\times 1000$, averaged over all 2,400 datasets. Selection uses independent validation. All four objectives are fixed across methods.

evaluation. It comprises 453,600 candidate fits and 4,800 evaluation sample tables. Both paths are reported. This analysis assesses regularization on the same mechanisms and covariate pools; it is not an independent cohort or an independent choice of benchmark design. Profiling and its nuisance fits use the short penalty path; the wider-path analysis concerns targeting only.

For estimated propensities, a training-only nonlinear multinomial-logit fit uses g, penalty 10^{-4}, and a fixed .001 numerical floor followed by normalization. The floor-mixture generator is not exactly a standard multinomial-logit model. We report both known-propensity validation of these training candidates and the more practical estimated-propensity validation ablation. The final predictive test uses the declared target for all methods. For each external objective, candidates and penalties are selected solely by independent validation half-Brier risk. Predictions are not refitted. The short-path analysis has 81 candidate fits per case and 194,400 in total.

Mean/CATE evaluation uses both the categorical predictors' implied means and direct numeric-response T-ridge predictions, selected by validation squared error. Linearity of ridge regression permits the numeric-response coefficients to be calculated from the category coefficients, but its range clipping and validation objective differ from simplex projection and proper-score selection. No mean estimator is judged as having failed because it does not output a structural estimand.

The complete categorical CDF and the delta-kernel mean embedding both contain the full seven-state law. CDF differencing followed by the identical Ψ is checked numerically on all profile estimates. Distributional AIPW in Appendix D uses custom orthogonal indicator scores, following the standard double/debiased-learning principle (Chernozhukov et al., 2018); it is not mislabelled as an official DoubleML package run. Package-based DoWhy and DoubleML comparisons are reported in Appendix K; their protocols are distinct from these custom score estimators.

The benchmark also reports the Janzing edge-removal KL (Janzing et al., 2013) for the DAG with $X \to A$, $X \to Y$, and $A \to Y$. The cut mechanism is $p^{\mathrm{cut}}(y \mid x) = \sum_b \bar{e}_b p_b(y \mid x)$, where $\bar{e}_b$ is the marginal treatment probability in the test covariate pool. We compute $\mathbb{E}_X \sum_{a,y} e_a(X) p_a(y \mid X) \log\{p_a(y \mid X)/p^{\mathrm{cut}}(y \mid X)\}$. Using $e_b(x)$ inside the cut mixture would instead give a conditional mutual-information quantity. Oracle values and T-learner plug-in values are reported separately from Θ_Ψ; they are not ranked as though the two statistics estimated the same target. A positive plug-in KL under a null is estimation error, not a calibrated rejection.

I.3 Structural certification and predictive results

The law-score construction and its assumptions are in Appendix D. Both regularization paths below use the same external objectives. Profiling results use the short-path nuisance fits. The short grid selected its largest penalty in 97.0% of ERM choices, 98.3% of escort choices, 99.8% of direct-weighting choices, and 99.8% of T-learner choices. Its mean excess half-Brier risk for natural, rare-context, rare-treatment and balanced objectives is, respectively, $(.004977, .008700, .007902, .006983)$ for ERM and $(.006751, .012309, .010826, .009651)$ for escort selection. The wider-path analysis addresses this validation-boundary limitation. The short-grid results describe that restricted penalty range, not conclusive rankings over regularization strength.

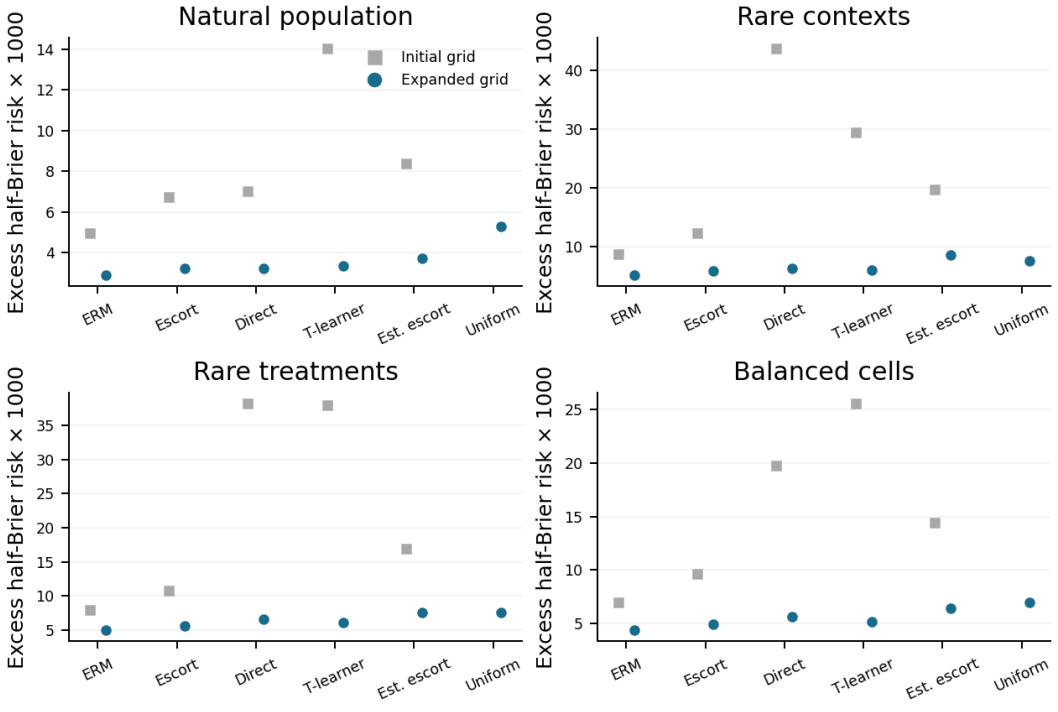


Figure 9: Short and wide regularization grids on the complete benchmark. The expanded path uses independent validation/test outcomes and the same training observations. Blue intervals describe 50 split-block means. The uniform predictor is an explicit control and a limiting candidate available to every method.

The wider path substantially lowers all methods' prediction risks, but does not recover a general escort advantage over ERM. Mean excess half-Brier risks for the four objectives are $(.002918, .005227, .005050, .004404)$ for ERM and $(.003222, .005829, .005626, .004901)$ for escort selection. Escort risk is 10.4%, 11.5%, 11.4%, and 11.3% higher, respectively. The corresponding paired differences, multiplied by 1,000, are 0.305 $[0.299, 0.310]$, 0.602 $[0.551, 0.654]$, 0.576 $[0.529, 0.623]$, and 0.498 $[0.471, 0.525]$, with 95% intervals over the 50 split blocks. All four comparisons remain nonzero after their declared Holm adjustment. These are conditional benchmark results, not population-independent superiority statements.

Escort selection is better on average than direct weighting and the flexible T-learner for these four objectives, reversing their ordering relative to the additive 16-cell example. Natural ERM is the strongest reported average predictor here. Estimated-propensity validation further increases rare-objective risk. Condition-specific and mean/CATE results are reported separately; no favorable subset replaces the complete matrix.

The result is consistent with finite-sample variance and parameter sharing: inverse-propensity weights can be large, and a shared conditional model can use common-regime observations to learn features relevant elsewhere. Directly targeting a small group discards some of that training information. These are explanations of the available comparisons, not an experimentally isolated causal decomposition of the risk difference. None of the fitted losses uses Ψ. The analysis supports cautious targeting interpretation and structural certification, not a topological prediction algorithm.

The wider-path audit replays all 4,800 independently generated evaluation sample tables and 115,200 validation choices, verifies that its training counts equal the short-path counts, and refits six predetermined complete cases. Every sample seed is distinct from the short-path validation, test and profiling seeds. Both sets of source hashes and the numerical refinement source are verified. The largest replayed score discrepancy is 1.12×10^{-16}.

Method	Natural	Rare context	Rare treatment	Balanced
Natural ERM	2.918	5.227	5.050	4.404
Escort / known e	3.222	5.829	5.626	4.901
Equal cell	4.335	6.100	6.200	5.618
Standardized natural	3.242	5.680	5.581	4.878
Generalized overlap	3.779	6.028	5.515	5.153
Direct objective	3.242	6.344	6.586	5.618
Observed Group DRO	3.328	6.072	6.034	5.227
Standardized Group DRO	3.515	6.283	6.283	5.458
Flexible T-learner	3.377	5.997	6.156	5.126
Escort / estimated e	3.464	6.191	6.054	5.279
Escort / estimated train + val	3.710	8.549	7.556	6.392
Uniform prediction	5.299	7.629	7.546	6.944

Table 9: Expanded regularization path with independent validation/test draws: excess half-Brier risk $\times 1000$, averaged over all 2,400 conditions. The short-path results appear in Table 8. All methods can select the uniform limiting predictor.

The known-propensity IPW and two-fold AIPW boxes cover all true law coordinates in every one of the 4,800 profiling datasets. Their induced profiles cover every point of the audited 11×11 grid at all three thresholds. The theoretical continuum guarantee comes from the common law-confidence event, not from checking a finite grid. At the primary threshold, confounded non-null cases with 20,000 training observations give 450 datasets per profiling budget. At 20,000 profiling observations, both valid methods certify a positive structural effect in zero cases. At 320,000 observations, IPW certifies a positive effect in 380/450 (84.4%) and cross-fit AIPW in 257/450 (57.1%). Neither resolves all 16 signatures or certifies the two-increment reversal in this stratum.

The lower AIPW certificate rate does not contradict Proposition 6. That identity is conditional on an outcome predictor and minimizes variance only at the true conditional law. The implemented cross-fit bound additionally pays for two smaller evaluation folds and a fold union bound; its range term can dominate any variance reduction. The experiment does not claim universally shorter intervals from augmentation.

For the same 450 cases and larger profiling budget, estimated-propensity Wald boxes cover all causal law coordinates in 386/450 (85.8%) cases; their pointwise asymptotic construction has no finite-sample guarantee here. Raw-cell CP boxes cover the full causal law in 143/450 (31.8%) cases despite their binomial label. In the 150 confounded no-effect controls at that budget, the corresponding raw-law coverage is zero. At threshold .10, all methods nevertheless produce zero false positive structural certificates in those controls. A coarse signature can mask law-level error; this coincidence does not validate a wrongly targeted confidence construction. Every threshold and every failed law-coverage event remains available in the output tables.

The short-path audit replayed all 12,000 count tables and all 105,600 validation choices, checked 57,600 complete profile grids, and refitted 12 predetermined full cases. The largest score replay difference was 4.45×10^{-16}. All 14,400 Group DRO fits meet the 2×10^{-6} dual-gap gate after the documented numerical refinement; the largest ridge normal-equation residual is 7.11×10^{-15}. The full-CDF reconstruction discrepancy is at most 4.45×10^{-16}. These checks verify implementation and simulation provenance; they do not establish an information-theoretic advantage of topology or guarantee performance outside the declared mechanisms.

All paired risk summaries first average the entire 48-condition matrix within a split, then form intervals over 50 split blocks. Holm adjustment is applied to the four stated escort-versus-ERM objectives within each analysis phase. Other comparisons are labelled descriptive or pointwise. The 50 blocks are independent randomizations of one fixed covariate release, not independent biological cohorts. Both training sizes, both assignment mechanisms, all positivity floors, all response settings, the no-effect controls, and all three structural thresholds are reported.

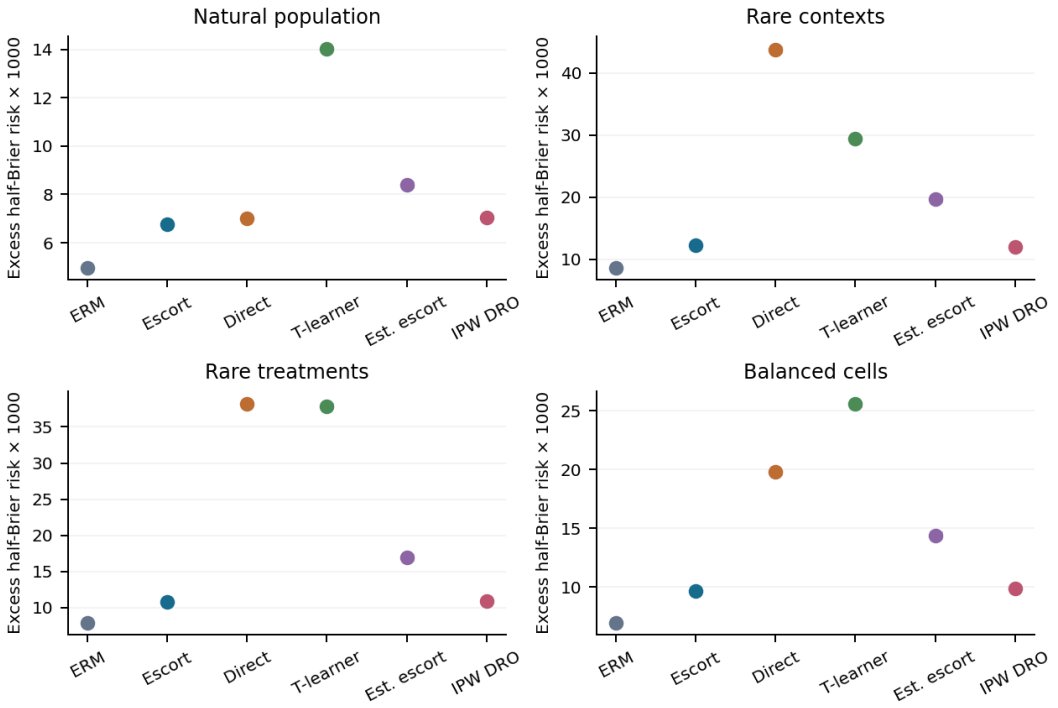


Figure 10: Short-path benchmark, averaging the complete condition matrix. Intervals describe variation over 50 independently randomized split blocks of the fixed covariate release. They are not intervals over 50 independent clinical cohorts. All methods use the same external objectives.

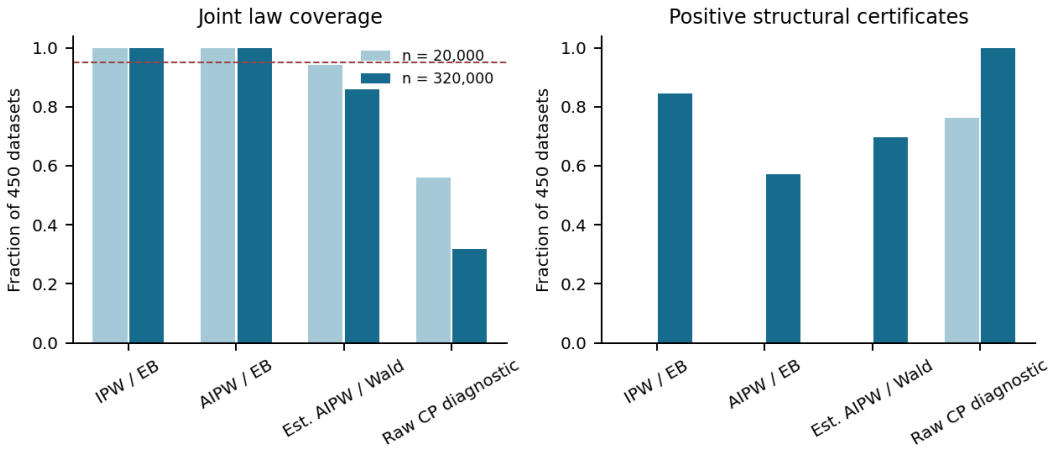


Figure 11: Confounded, non-null cases with 20,000 training observations, primary threshold .10: 450 datasets per profiling budget. Known-propensity empirical-Bernstein (EB) intervals have a finite-sample guarantee; estimated-propensity Wald and raw-cell CP diagnostics do not inherit it.

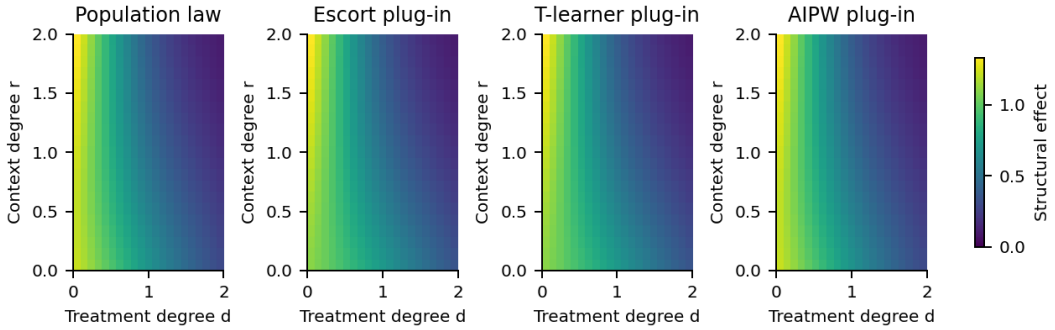


Figure 12: Fixed-index illustration: split 0, confounded assignment, $\epsilon = .02$, $\tau = .15$, and 20,000 training draws; the AIPW panel uses 320,000 independent profiling draws. The example is chosen by protocol index, not performance. Plug-in surfaces are not certificates.

J IDENTIFICATION, STRUCTURAL GEOMETRY, AND SENSITIVITY

J.1 STRUCTURAL ESTIMANDS AND IDENTIFYING ASSUMPTIONS

Let $A \in \{1, \dots, J\}$ denote treatment, $Z \in \{1, \dots, S\}$ a pre-treatment context, and $P_z^a = \mathcal{L}(Y(a) \mid Z = z)$. Let Ψ map an outcome law into a finite-dimensional Hilbert space. Specify strictly positive reference masses $p_{a|z}$ and h_z, independently of the evaluation outcomes. They describe the reference population, which need not equal the sampling allocation. Define

$$q_d(a \mid z) = \frac{p_{a|z}^d}{\sum_b p_{b|z}^d}, \qquad \nu_r(z) = \frac{h_z^r}{\sum_u h_u^r}, \tag{36}$$

$$\Theta_\Psi(d, r) = \sum_z \nu_r(z) \sum_{a,b} q_d(a \mid z) q_d(b \mid z) \|\Psi(P_z^a) - \Psi(P_z^b)\|^2. \tag{37}$$

The sum uses ordered pairs; its diagonal terms vanish. At $d = 0$ treatments receive equal weights, and at $d = 1$ their reference frequencies are used. Lower d emphasizes rare arms *relative to larger* d, without necessarily assigning a rare arm more absolute weight than a frequent arm. The same interpretation applies to r. If p and h are uniform, the surface is constant. The finite-confounder, oscillator, and finite-panel comparisons report $d, r \in \{0, 0.5, 1, 2\}$; the four-context and ACIC experiments use the grids specified in their own protocols. The finite confidence event applies to every finite real d, r.

Proposition 8 (Target-specific identification)**.** *Assume consistency and positivity. If, for every relevant* a, z*,*

$$\Psi(P_z^a) = \Psi\{\mathcal{L}(Y \mid A = a, Z = z)\}, \tag{38}$$

then (37) *is identified from the observational law and the specified reference masses, for all* d, r*.*

This is a multi-arm escort corollary of Proposition 6 of Faghihi & Saki (2026), with assumptions specific to the declared structural target. Their finite-cell no-switching result supplies one sufficient mechanism for (38). Compatible filtered homotopy equivalences also preserve the corresponding homological summaries. The converse, from isomorphic homology modules to filtered homotopy equivalence, is not assumed here. Nonlinear transforms must be applied after any required mixture of laws; averaging the topologies of confounder strata is generally incorrect.

When an observed confounder W is used to identify the *whole* law, we instead assume $Y(a) \perp A \mid (W, Z)$ and overlap. On a finite outcome space, write $m_{az}(w) = \mathbb{P}(Y = \cdot \mid A = a, W = w, Z = z)$. The law to transform is

$$\mu_{az} = \mathbb{E}\{m_{az}(W) \mid Z = z\}, \qquad \Psi(P_z^a) = \Psi(\mu_{az}). \tag{39}$$

The synthetic adjusted estimators below use this stronger assumption. The weak target-specific condition alone does not justify arbitrary distribution estimation followed by a different transform.

J.2 FINITE STRUCTURAL GEOMETRY AND DECISION RADIUS

Fix a finite simplicial complex $\mathcal{K}$ on M vertices and a probability threshold c. For $u \in \Delta_{M-1}$, form the induced subcomplex on $\{v : u_v \geq c\}$. A descriptor $\Psi(u)$ is computed from this complex. This is a probability-superlevel construction; c is not a metric radius. It defines a finite categorical target, not an estimate of a continuous density's topology without further approximation assumptions.

For each active pattern $b \in \{0, 1\}^M$, let s_b be its descriptor and C_b the corresponding region in the simplex, with inactive inequalities strict. Its closure is a polytope. Define

$$\rho_\Psi(u) = \inf\{\|v - u\|_\infty : v \in \Delta_{M-1},\ \Psi(v) \neq \Psi(u)\}. \tag{40}$$

Theorem 9 (Signature radius and computable feasibility)**.** *For a fixed finite descriptor table,* $\rho_\Psi(u)$ *equals the minimum distance from* u *to* $\overline{C}_b$ *over nonempty patterns with* $s_b \neq \Psi(u)$*. A probability perturbation strictly smaller than this radius preserves* Ψ*. Moreover* $\rho_\Psi(u) \geq \min_v |u_v - c|$*. If* $\Psi_2 = g \circ \Psi_1$*, then* $\rho_{\Psi_2}(u) \geq \rho_{\Psi_1}(u)$*. For a box* $[\ell, v]$*, define*

$$L_i(b) = \max\{\ell_i, c b_i\}, \qquad U_i(b) = \min\{v_i, c + (1 - c) b_i\}. \tag{41}$$

The closed pattern intersects the box and simplex exactly when $L_i(b) \leq U_i(b)$ *for all* i*,* $\sum_i L_i(b) \leq 1$*, and* $\sum_i U_i(b) \geq 1$*.*

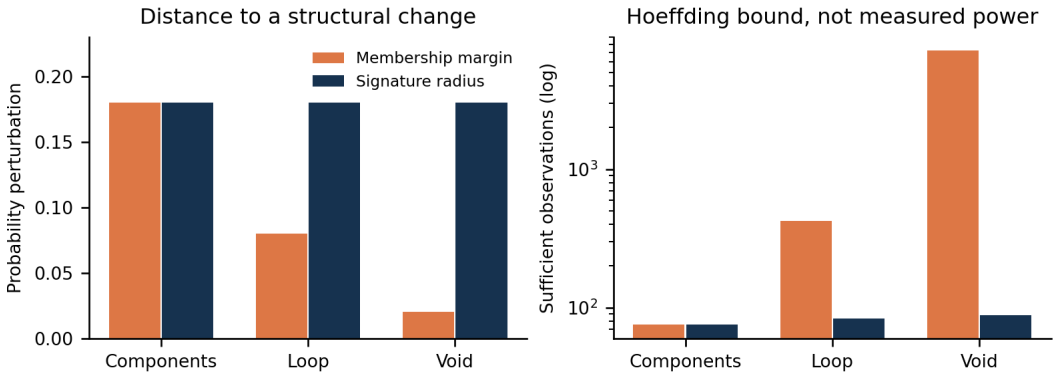


Figure 13: A fixed-complex example separates a membership margin from a structural radius. A center vertex of mass 0.28 fills the two-point boundary, cycle, or octahedral sphere; outer vertices share mass 0.72. At $c = 0.1$, the void example has margin 0.02 but signature radius 0.18. At box radius 0.04, 57 feasible active patterns have the same H_0, H_1, H_2 signature. These are analytic examples, not real-data performance results.

Thus radius computation needs monotone bisection over box feasibility, or one small linear program per different-signature pattern. We use 2^M pattern enumeration and 45 bisection steps, and independently verify the reported radii with linear programming. This is practical for $M \leq 7$ in the experiment, and is exponential in general. Inactive closures can overinclude possibilities at ties in confidence sets, which is conservative.

Corollary 10 (A sufficient sample bound)**.** *For n independent draws from a finite law u, its empirical frequency $\hat{u}$ satisfies $\mathbb{P}\{\Psi(\hat{u}) \neq \Psi(u)\} \leq 2M \exp(-2n\epsilon^2)$ for every $0 < \epsilon < \rho_\Psi(u)$. Consequently $n \geq \log(2M/\alpha)/(2\epsilon^2)$ suffices for error at most α.*

This is Hoeffding's inequality applied with the structural decision radius. Relative to the membership margin, the limiting sufficient-bound ratio in Figure 13 is 81 for the void example. It is not a measured 81-fold data saving. Under rarity, n is the number of relevant observations, not the total dataset size.

Let simultaneous law confidence sets be $\mathcal{C}_{az} = \{u \in \Delta_{M-1} : \ell_{az} \leq u \leq v_{az}\}$, with joint coverage at least $1 - \alpha$. For a specified sensitivity parameter $\delta \geq 0$, expand their bounds by δ and intersect with the simplex. Let $\mathcal{S}_{az}(\delta)$ contain every feasible signature in this expanded set, and let D^-_{abz} and D^+_{abz} be the minimum and maximum squared distances between signatures in $\mathcal{S}_{az}$ and $\mathcal{S}_{bz}$. Set $D^-_{aaz} = D^+_{aaz} = 0$.

Theorem 11 (One event covers the entire rarity profile)**.** *Suppose the law estimated by $\mathcal{C}_{az}$ is u_{az} and the target law satisfies $\|\mu_{az} - u_{az}\|_\infty \leq \delta_\star$ for all a, z. On the simultaneous confidence event, for every $\delta \geq \delta_\star$ and every d, r,*

$$\sum_{z,a,b} \nu_r(z) q_d(a \mid z) q_d(b \mid z) D^-_{abz}(\delta) \;\leq \Theta_\Psi(d, r) \leq \sum_{z,a,b} \nu_r(z) q_d(a \mid z) q_d(b \mid z) D^+_{abz}(\delta). \quad (42)$$

These bounds widen monotonically with δ. They remain valid when (d, r) is inspected or selected using the same data.

Shared signatures across pairs can make the pairwise extrema incompatible, so these are outer bounds, not generally sharp extrema of the joint target. The guarantee needs fixed reference masses. Estimated weights require their own joint uncertainty accounting. A sensitivity radius is an assumption about the discrepancy between laws, not a data-derived estimate of hidden confounding. Total variation at most δ implies the stated coordinate bound, but the optimized radius is sharp for the stated ℓ_∞ class, not automatically for a total-variation class.

J.3 Proof of target-specific identification

For Proposition 8, substitute (38) separately in every summand of (37). Positivity makes each observed conditional law defined, and consistency supplies its factual interpretation. All remaining

operations use fixed known masses and deterministic transforms, proving joint identification of the family. This proof does not identify a full law when Ψ is noninjective.

J.4 Proof of the signature-radius theorem

The finite simplex is partitioned by the active-pattern regions C_b, with deterministic descriptor s_b on each region. The set on the right of (40) is their finite union over different descriptors. The distance to a set equals the distance to its closure, and the distance to a finite union equals the minimum of distances. Empty regions are excluded; the nonempty closed regions used at $c = 0.1$ in the experiments are closures of feasible active regions. This proves the optimization formula. Strictly below its infimum there is no law with a different signature, proving the robustness statement.

If $\|v - u\|_\infty < \min_i |u_i - c|$, no coordinate crosses a threshold, so the entire active set and hence its signature are unchanged. This proves the margin bound. A coarsening $\Psi_2 = g \circ \Psi_1$ can change only if Ψ_1 changes, so its adverse set is a subset, proving radius monotonicity. This partial order formalizes the tradeoff: a less informative descriptor can be easier to certify. It does not establish that the coarser descriptor is better for the scientific question.

For the feasibility characterization, necessity follows by summing lower and upper coordinate bounds. For sufficiency, start at L and distribute the remaining mass $1 - \sum_i L_i$ among coordinates with capacities $U_i - L_i$. The total capacity is at least the remaining mass, so the procedure yields a probability vector in the box and closed pattern. This also proves monotonicity of feasibility as a probability ball is enlarged. Inactive closures may admit an unattainable exact pattern on a threshold boundary; retaining it enlarges uncertainty and cannot invalidate coverage. For general thresholds, closed patterns whose strict region is empty must be removed when claiming the exact radius, although they can remain in an outer confidence set. The implemented examples at $c = 0.1$, $M \in \{3, 5, 7\}$ have no such extraneous feasible pattern.

The void witness is a cone over the octahedral sphere. If its center remains active, every induced outer subcomplex coned to that center is contractible. It therefore has Betti vector $(1, 0, 0)$ even when many outer vertices disappear. To obtain $(1, 0, 1)$, take the center just below 0.1 while retaining the outer shell; the required center change has infimum 0.18 and outer masses can absorb it. A different topology at smaller distance is impossible while the center stays active. The loop and connected-path examples have the analogous argument. The generic membership margins are respectively 0.08 and 0.18.

For Corollary 10, each empirical frequency is an average of independent Bernoulli variables, so $\mathbb{P}(|\hat{u}_i - u_i| > \epsilon) \le 2e^{-2n\epsilon^2}$. A union bound gives the sup-norm event; on it, the radius theorem applies. The statement uses $\epsilon < \rho$ rather than asserting invariance at a possibly changing boundary. If the relevant stratum occurs with probability π, the conditional bound uses its realized count N_π; the unconditional failure bound is at most $\mathbb{E}[\min\{1, 2Me^{-2N_\pi\epsilon^2}\}]$. Small π can dominate the nominal total sample size.

J.5 Proof and construction of simultaneous bands

On the joint law-confidence event, all u_{az} lie in their respective boxes. A target law at coordinatewise distance at most $\delta_\star$ lies in each box expanded by any $\delta \ge \delta_\star$. Feasibility enumeration consequently includes its true signature. Every true pair distance is then between its computed extrema. All weights in (37) are nonnegative, so summation preserves the inequalities. The event never depends on d, r, proving simultaneity even under data-dependent inspection. Larger sensitivity balls give supersets of signatures, so minima decrease and maxima increase. No union penalty over the escort grid is needed. If additional transforms are computed from the same valid law sets, their image sets are also jointly covered; causal assumptions must still identify each chosen transform.

In the experiment, for each multinomial conditional count vector of total n, a coordinate count k has a binomial marginal. With Q coordinates over all (Z, W, A) tables, the interval endpoints are

$$\ell(k,n) = \begin{cases} 0 & k = 0, \\ B^{-1}_{k,n-k+1}(\alpha/(2Q)) & k > 0, \end{cases} \tag{43}$$

$$v(k,n) = \begin{cases} 1 & k = n, \\ B^{-1}_{k+1,n-k}(1-\alpha/(2Q)) & k < n. \end{cases} \tag{44}$$

Here $B^{-1}_{a,b}$ is the beta quantile. For $n = 0$ the code explicitly returns $[0, 1]$. Binomial coverage and a union bound conditionally on table sizes imply unconditional simultaneous coverage as well (Clopper & Pearson, 1934). Averaging lower and upper bounds over the *known* W weights preserves coverage for the standardized law. This procedure is unrelated to estimating propensities; the AIPW nuisance fit does not enter its coverage proof. Probability-simplex feasibility is then enforced. The averaging and union bound explain some of its conservatism.

Uncertainty near a structural boundary. If a vertex probability approaches c from opposite sides, arbitrarily close laws may have distinct structural signatures. For the finite examples here, positive-probability laws can be chosen with per-observation KL divergence $O(\epsilon^2)$ but a fixed structural effect difference. For $\epsilon = o(n^{-1/2})$, their n-sample laws become indistinguishable. The two-point lower bound in Appendix B.3 rules out uniform consistency over such a neighborhood. Cross-fitting does not repair this nonregularity. A normal interval for a smooth law functional cannot be relabeled as a normal interval for a hard Betti effect at a threshold.

K Causal capabilities: a matched comparison

The scientific question is whether a structural target reveals a causal phenomenon absent from a conventional summary, rather than whether a geometric predictor wins an unrelated regression task. We separate *mean invisibility*, *identification*, and *testing power*. If all conditional means agree but one pair of structural signatures differs, every coordinate-mean contrast is zero while $\Theta_\Psi(d, r) > 0$ for positive reference masses. This elementary separation motivates GeoRARE but is not exclusive to it: distributional kernel methods already detect unequal laws (Fawkes et al., 2023).

A stronger distinction concerns what is identifiable. In an explicit hidden-confounding SCM class, two free latent reweighting parameters vary while the entire observed law remains fixed. The ATE identified set is exactly $[-.4, .4]$, whereas the declared component contrast remains $\Theta_\Psi(0) = .5$ throughout the class. The proof, sharp endpoint linear programs, and exhaustive 441-grid-model checks are in Appendix K.1. This is a quantitative extension of Faghihi & Saki (2026), conditional on the stated signature-preserving class; no estimator can recover the unspecified ATE from the common observations.

Design and fairness. The matched comparison uses 2,000 datasets on the finite path, cycle, and octahedral complexes defined in Appendix J. There are three geometries, three scenarios (Null, Shape, Density), two sample sizes $(1,000, 10,000)$, two marginal treatment frequencies $(.05, .20)$, and 50 repetitions: 1,800 datasets. A positive mean-effect control adds 200 components datasets. These are controlled mechanism experiments, with simulated responses. Observed $W \sim \mathrm{Bernoulli}(.5)$ affects both assignment and outcome probabilities; $\mathbb{P}(A = 1 \mid W = w)$ is $.5\epsilon$ or 1.5ϵ. All methods see the same records and W.

The null has equal treatment laws given W. Shape removes the center from a density-superlevel complex while leaving symmetric coordinate means unchanged. Density changes center probability while preserving the selected Betti signature. Mean shifts probability between the two outer vertices of the components complex, giving ATE .30 without changing its signature. No thresholds, kernels, scenarios or seeds are selected after evaluation. For 11 fixed statistics, 499 common conditional treatment permutations within (W, fold) strata refit all nuisance estimates. Conservative tie handling is particularly relevant for discrete topological statistics. The nominal level is .05; validity is proved for the full conditional no-distributional-effect null, not a weaker no-topological-effect null.

We implement Janzing's original joint-law KL arrow strength, not an unlabeled pairwise KL substitution (Janzing et al., 2013). Balanced JS and Gaussian contrasts receive equal-treatment weighting as

Method	Null	Shape	Density	Mean
GeoRARE $H_{0:2}$	0.0	96.0	4.0	0.0
Gaussian AIPW	5.0	71.0	26.0	81.0
Janzing KL	4.0	97.0	52.5	85.0
Balanced JS	6.0	98.0	52.5	83.0
Actual packages; native/asymptotic calibration				
DoWhy OLS	4.5	9.0	4.5	92.0
DoubleML ATE	8.0	5.5	4.0	84.0
DoubleML CDF Wald	8.5	98.0	63.0	81.5

Table 10: Rejection percentages on the identical scalar components records (200 datasets per scenario). Shape changes the component signature while keeping the mean zero; Density changes probabilities without changing the signature; Mean changes the mean by 0.30 while preserving the signature. The upper block uses shared conditional permutations; the lower block uses actual package tests with different calibration. DoubleML CDF Wald has an elevated measured null rate, so its high detection rate is not a matched-size superiority claim.

well. The Gaussian AIPW comparator uses cross-fitted categorical scores and a fixed unit-bandwidth kernel. Actual DoWhy OLS/IPW and DoubleML IRM are executed on all 800 scalar components records (PyWhy contributors, 2026; DoubleML contributors, 2026). DoubleML is also fitted to both nontrivial CDF indicators, so it is not restricted to a mean target. Its joint Wald and Bonferroni tests are reported separately from conditional randomization calibration. A finite quantile-grid randomization test is included; continuous-density quantile asymptotics are not imposed on categorical outcomes.

Results. Across all 600 Shape datasets, $H_{0:2}$ and Euler each reject in 91.5%, versus 70.7% for the particular Gaussian AIPW statistic, 95.5% for Janzing KL, and 96.8% for balanced JS. The adjusted mean statistic rejects in 8.2%; its underlying mean contrast is exactly zero. This supports structural usefulness relative to a mean summary, but not superiority to all distributional methods. Table 10 shows the identical scalar subset and confirms that mean methods work on the genuine mean-effect control. DoubleML CDF Wald detects Shape in 98% but has an 8.5% measured null rejection rate; the higher detection rate therefore does not constitute a matched-size comparison.

The prespecified primary setting is Shape with $\epsilon = .05$, $n = 10,000$, pooled equally across three geometries. GeoRARE, Gaussian AIPW and Janzing all reject in $150/150$ datasets. The superiority criterion fails. This ceiling also limits discrimination between their powers. Conservative simultaneous paired Monte Carlo intervals permit differences of roughly ± 11.6 percentage points; zero observed discordance is not proof of identical population powers. Appendix K.2 reports all settings, including weaker-performance cases and the paired uncertainty calculation.

Separately, exact probability-box projection gives 100% observed structural-band coverage over 2,000 datasets and positive lower bounds on 44% of the 600 Shape datasets. No positive lower bounds occur in the 1,400 Null, Density or Mean datasets. These are finite-sample structural certificates; a mere rejection by a topological statistic is not one. The positive-certificate rate is only 5% at $n = 1,000$ and 83% at $n = 10,000$ in Shape, showing the continuing cost of rare-arm information.

K.1 Effect magnitude, identification, and testing power

This appendix distinguishes effect magnitude, identification, and testing power. The first is a property of the declared target; the second is a property of a model class and its observable implications; the third is a property of an estimator, a calibration procedure, and a sampling experiment. None implies the other automatically. All results below use outcome laws, not individual-level cross-world couplings inferred from marginal distributions.

Proposition 12 (A mean-invisible structural effect)**.** *Suppose $\mathbb{E}[Y(a) \mid Z = z] = m_z$ for every treatment a, but $\Psi(P_z^a) \neq \Psi(P_z^b)$ for at least one pair in a context with positive reference weight. Then all conditional coordinate-mean contrasts vanish and $\Theta_\Psi(d, r) > 0$ for every finite d, r with strictly positive reference masses.*

Proof. The first assertion follows by subtraction. Every summand defining Θ_Ψ is nonnegative. The indicated unequal pair contributes a strictly positive squared norm multiplied by strictly positive weights. □

This is a separation of estimands, not a proof that a mean estimator is biased. It is also not a separation from all distributional methods: with a characteristic kernel, unequal laws have unequal mean embeddings. A selected finite set of scalar quantiles can miss a distributional difference, whereas all scalar CDF values characterize a scalar law. Coordinatewise CDFs need not characterize a multivariate joint law.

Proposition 13 (Rarity attenuation and a limit on power claims)**.** *In a randomized two-node system $A \to Y$, let $\mathbb{P}(A = 1) = \epsilon \in (0, 1)$, $P_a = \mathcal{L}(Y(a))$ and $D = \|\Psi(P_1) - \Psi(P_0)\|^2 > 0$. Janzing's original joint-law KL arrow strength equals $I(A; Y)$ and is bounded by binary entropy $H_b(\epsilon)$. It therefore tends to zero as $\epsilon \downarrow 0$. In contrast,*

$$\Theta_\Psi(d) = 2q_d(1 - q_d)D, \qquad q_d = \frac{\epsilon^d}{\epsilon^d + (1 - \epsilon)^d}, \qquad \Theta_\Psi(0) = D/2.$$

For fixed binary treatment counts and reference masses, multiplying a test statistic by $2q_d(1-q_d) > 0$ leaves its permutation p-value unchanged.

Proof. Cutting the only arrow replaces $P_{A,Y}$ by $P_A P_Y$, giving mutual information by the definition of KL divergence. The identity $I(A; Y) = H(A) - H(A \mid Y)$ and nonnegativity of conditional entropy give the bound. Only two off-diagonal ordered pairs occur in Θ_Ψ. Finally, $cT^{(b)} \geq cT^{(0)}$ if and only if $T^{(b)} \geq T^{(0)}$ for $c > 0$. □

The entropy bound is a statement about the randomized two-node example, not an asserted formula for arbitrary confounded graphs. Balanced JS and equally weighted Gaussian contrasts likewise avoid this population-frequency attenuation. Thus reweighting is not specific to topology and cannot compensate for the absence of observations from a rare arm. In a single binary context, different d values alone do not create distinct calibrated permutation tests. Multiple contexts or treatments can change relative summands, which is a different situation requiring a prespecified joint test.

Proposition 14 (A sharp structural-versus-mean identification separation)**.** *Consider the following declared SCM class. Let $U \sim \text{Bernoulli}(.2)$, $A = U$, and use independent uniform exogenous noises to sample each potential outcome from a categorical law $Q_{a,u}$. In vertex order $(-1, +1, 0)$, set*

$$Q_{0,0} = (.35, .35, .30), \qquad Q_{1,1} = (.49, .49, .02),$$
$$Q_{0,1} = (.35 - t_0, .35 + t_0, .30), \qquad Q_{1,0} = (.49 - t_1, .49 + t_1, .02),$$

where $(t_0, t_1) \in [-.2, .2]^2$. The observable law of (A, Y) is constant throughout this class. The ATE identified set is exactly $[-.4, .4]$. On the three-vertex path with the zero vertex in the middle, take the probability superlevel complex at threshold .1 and its component count $\Psi = \beta_0$. Then $\Theta_\Psi(0) = .5$ for every model in the class. More generally $\Theta_\Psi(d) = 2q_d(1 - q_d)$ is point-identified for every fixed d.

Proof. All displayed probabilities are positive and sum to one. Consistency gives $\mathbb{P}(A = 0, Y = y) = .8Q_{0,0}(y)$ and $\mathbb{P}(A = 1, Y = y) = .2Q_{1,1}(y)$, independent of t_0, t_1. Under intervention,

$$P_0 = (.35 - .2t_0, .35 + .2t_0, .30),$$
$$P_1 = (.49 - .8t_1, .49 + .8t_1, .02).$$

Consequently $\text{ATE} = 1.6t_1 - .4t_0$. Its minimum and maximum on the rectangle are $-.4$ and $.4$, attained at opposite corners. Continuity and convexity of the rectangle imply that every intermediate value is attained, proving sharpness. Every vertex of P_0 exceeds .1; its induced path is connected. The center of P_1 lies below .1 and both outer vertices exceed .1; its induced subcomplex has two components. The same signatures hold for the factual laws. The structural distance squared is therefore one, and the asserted escort formulas follow. □

This class satisfies marginal treatment positivity and signature-specific ignorability. It does not satisfy ignorability given observed covariates; conditioning on the latent U would also violate overlap. The class restriction is substantive and is not learned or certified from the shared observational distribution. Without it, the structural conclusion need not hold. This construction extends the target-specific reasoning of Faghihi & Saki (2026); it is not a claim to identify ATE under arbitrary hidden confounding. Linear programming independently checks both sharp endpoints, and all 441 models on a declared parameter grid are checked for observational equality and structural invariance. Those checks supplement, not replace, the proof over the continuous rectangle.

Proposition 15 (Matched conditional randomization calibration)**.** *Assume independent sampling, observed W, consistency, and treatment exchangeability given W. Under $\mathcal{L}(Y(1) \mid W = w) = \mathcal{L}(Y(0) \mid W = w)$ for each w, condition on the outcomes, W, a fixed outcome-independent sample fold, and treatment counts within each (W, fold) stratum. Uniformly permuting treatment labels in these strata gives a valid conditional reference distribution for any fixed measurable statistic, including a discontinuous topological one. For B independent such permutations,*

$$p = \frac{1 + \sum_{b=1}^{B} \mathbf{1}\{T^{(b)} \geq T^{(0)}\}}{B + 1}$$

is super-uniform, with ties treated conservatively.

Proof. Under the null the conditional law of the outcome does not depend on the label within a stratum. Conditional on the number treated, all assignments are equally likely. The original assignment and independently sampled assignments are exchangeable conditional on the pooled data and margins. Their descending rank is uniform when ties are randomly broken; counting all tied values above the observed statistic can only increase the rank and its corresponding p-value. Re-estimating every nuisance function on all permuted assignments makes each statistic the same function of its assigned data, preserving exchangeability. □

The null in this proposition is equality of conditional outcome laws, not merely equality of means or Betti signatures. Rejection by a topological statistic does not by itself certify a nonzero population topological contrast. We separately project simultaneous Clopper–Pearson boxes using Theorem 11. Independent permutations use the multivariate hypergeometric distribution of categorical treatment counts; an exhaustive 81-table example checks size for every implemented statistic.

Distributional estimators and package comparators. Let $g_a(w) = \mathbb{P}(Y = \cdot \mid A = a, W = w)$ and $e(w) = \mathbb{P}(A = 1 \mid W = w)$. The cross-fitted vector score used by the Gaussian comparator is

$$\widehat{\delta} = \frac{1}{n} \sum_i \left[\widehat{g}_1(W_i) - \widehat{g}_0(W_i) + \frac{A_i}{\widehat{e}(W_i)} (e_{Y_i} - \widehat{g}_1(W_i)) - \frac{1 - A_i}{1 - \widehat{e}(W_i)} (e_{Y_i} - \widehat{g}_0(W_i)) \right],$$

where e_y is the category-indicator vector and every fitted nuisance uses the other fold. The statistic is $\widehat{\delta}^T K \widehat{\delta}$ with $K_{uv} = \exp(-\|x_u - x_v\|^2/2)$. This is a cross-fitted AIPW kernel statistic with the exact randomization calibration above; it is not represented as an execution of another paper's entire testing algorithm. Conditional on the training fold, the expectation of the arm-specific score minus its true mean is $\mathbb{E}[(1 - e_a/\widehat{e}_a)(\widehat{g}_a - g_a)]$. It vanishes if either nuisance is correct, proving the familiar double robustness identity. With an arbitrary wrong propensity the identity does not assert finite-sample unbiasedness.

Actual DoWhy OLS and normalized IPW are also executed. Actual DoubleML IRM is fitted to the scalar outcome and to its two nontrivial CDF indicators in the three-category experiment. These indicators span all nonconstant functions on the outcome alphabet; they are a full scalar-distribution comparator, not just a selected quantile grid. The joint Wald test uses the cross-fitted score covariance, with a chi-square reference of its numerical rank, and remains an asymptotic test. The separate Bonferroni combination of package CDF p-values is reported. Continuous-quantile asymptotics are not applied to these discrete outcomes. Package-native mean and CDF tests are tabulated separately from matched conditional randomization tests.

Janzing versus a pairwise KL substitution. For the observed-confounder benchmark the graph is $W \to A$, $W \to Y$, $A \to Y$. Removing the last arrow replaces $g_a(w)$ by $\sum_b p_b g_b(w)$, where

$p_b = \mathbb{P}(A = b)$ is the marginal parent distribution. Accordingly the implemented original arrow statistic is

$$\sum_w h_w \sum_a e_a(w)\,\mathrm{KL}\left(g_a(w)\,\middle\|\,\sum_b p_b g_b(w)\right).$$

The software uses empirical $h_w, e_a(w), p_b$ and a fixed half-count smoothing of conditional outcome probabilities. It is not mislabeled as a pairwise interventional KL. Direct evaluation of KL between the full original and cut joint laws checks this implementation. Balanced JS, total variation, and Gaussian comparisons of standardized interventional laws supply additional non-topological comparators with equal treatment weighting. Thus a weighting advantage is not reserved for GeoRARE.

K.2 Protocol, audit, and complete results

The probability threshold is .1 and the fixed Gaussian bandwidth is one in the declared vertex coordinates. For Null, the conditional center masses are $(.18, .22)$ across W in both arms. Shape changes the treated masses to $(.01, .03)$; Density changes them to $(.12, .16)$. Remaining mass is uniform over outer vertices except in the components Mean control, where .15 is moved from the negative to the positive vertex. All geometries and these qualitative mechanisms are mathematically chosen controls. The experiment tests their specified structural and distributional consequences.

The source and protocol were fixed before full evaluation. Five implementation-check datasets used disjoint seeds and were not used to select settings. Evaluation completed without source changes, omitted conditions, package errors, or reruns; implementation diagnostics are archived. Its 2,000 datasets contain 800 actual DoWhy and DoubleML comparisons. Fixed seeds make runs reproducible; the simulation randomness is not evidence from 2,000 distinct real applications.

The protocol specified paired normal Monte Carlo intervals for two primary comparisons. Both had zero observed discordance, making those intervals degenerate. They remain in the machine-readable audit but are not interpreted as exact equality. We use conservative Clopper–Pearson bounds for each positive and negative discordance probability within each shape, with a union bound over both directions, three shapes and two comparisons, then averaged the corresponding difference bounds equally over shapes. These valid simultaneous bounds are $[-.1162, .1162]$. Because the prespecified paired intervals degenerate under zero observed discordance, we additionally report conservative simultaneous Clopper–Pearson bounds. This additional uncertainty analysis does not alter the observed rates or the prespecified superiority decision. No experimental settings or datasets depend on this uncertainty calculation.

The condition-level tables give Wilson intervals as descriptive Monte Carlo summaries. Primary comparisons and all settings remain distinguished. The package contains every dataset-level statistic, p-value, band, package estimate and source count table, and the exact package versions. Native DoubleML inference is asymptotic; the measured null rate must be considered alongside detection. DoWhy IPW is reported as a point estimator with bias and RMSE; no significance test was computed for this estimator. All such package summaries are supplied as CSV files.

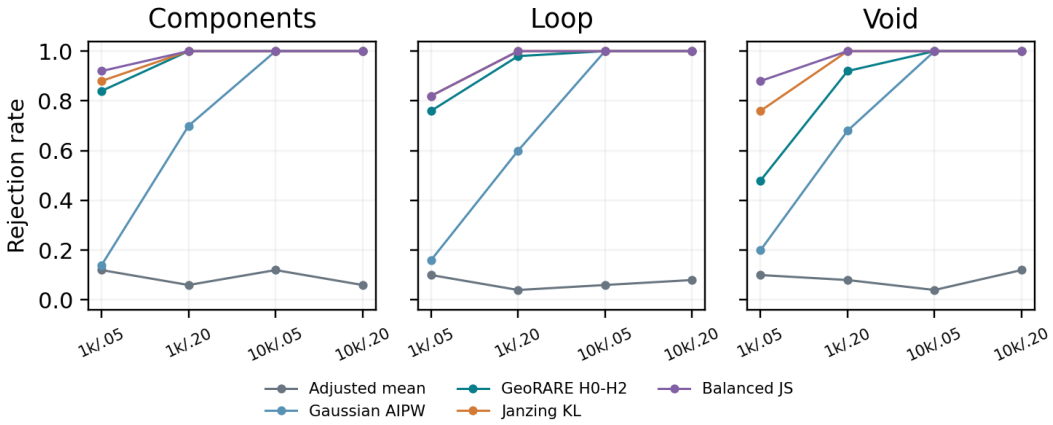


Figure 14: All Shape conditions for selected comparators. Horizontal labels are sample size/treatment frequency. Each point uses 50 paired datasets. The figure displays measured rates, not biological uncertainty. Complete Monte Carlo intervals follow in the tables.

Method	Null	Shape	Density	Mean
GeoRARE $H_{0:2}$	0.2	91.5	3.7	0.0
GeoRARE Euler	0.2	91.5	3.7	0.0
GeoRARE AET	0.8	65.8	3.7	0.0
GeoRARE Hodge	0.8	88.5	3.3	0.0
Adjusted mean	6.2	8.2	6.3	86.5
Quantile grid	0.2	25.3	2.5	65.5
Gaussian DM	6.0	72.7	27.3	86.5
Gaussian AIPW	5.7	70.7	28.2	81.0
Janzing KL	5.8	95.5	48.7	85.0
Balanced JS	4.5	96.8	49.2	83.0
Total variation	4.7	87.7	46.3	85.5

Table 11: All fixed statistics, shared conditional-randomization calibration. Null, Shape and Density each pool 600 datasets equally across three geometries, two sample sizes and two treatment frequencies. Mean contains only 200 components datasets. These pooled rates are secondary descriptive summaries; the primary contrast uses only the declared rare, large-sample setting.

Method	$1k, .05$	$1k, .20$	$10k, .05$	$10k, .20$
GeoRARE $H_{0:2}$	0 [0,7]	0 [0,7]	0 [0,7]	0 [0,7]
GeoRARE Euler	0 [0,7]	0 [0,7]	0 [0,7]	0 [0,7]
GeoRARE AET	0 [0,7]	0 [0,7]	0 [0,7]	0 [0,7]
GeoRARE Hodge	0 [0,7]	0 [0,7]	0 [0,7]	0 [0,7]
Adjusted mean	8 [3,19]	2 [0,10]	6 [2,16]	2 [0,10]
Quantile grid	0 [0,7]	0 [0,7]	0 [0,7]	0 [0,7]
Gaussian DM	8 [3,19]	2 [0,10]	6 [2,16]	2 [0,10]
Gaussian AIPW	8 [3,19]	4 [1,13]	6 [2,16]	2 [0,10]
Janzing KL	6 [2,16]	2 [0,10]	4 [1,13]	4 [1,13]
Balanced JS	8 [3,19]	4 [1,13]	8 [3,19]	4 [1,13]
Total variation	8 [3,19]	2 [0,10]	6 [2,16]	2 [0,10]

Table 12: All 50-replicate conditions for components, null. Entries are rejection percentages and 95% Wilson Monte Carlo intervals; these are not biological intervals.

Method	$1k, .05$	$1k, .20$	$10k, .05$	$10k, .20$
GeoRARE $H_{0:2}$	84 [71,92]	100 [93,100]	100 [93,100]	100 [93,100]
GeoRARE Euler	84 [71,92]	100 [93,100]	100 [93,100]	100 [93,100]
GeoRARE AET	84 [71,92]	100 [93,100]	100 [93,100]	100 [93,100]
GeoRARE Hodge	84 [71,92]	100 [93,100]	100 [93,100]	100 [93,100]
Adjusted mean	12 [6,24]	6 [2,16]	12 [6,24]	6 [2,16]
Quantile grid	0 [0,7]	50 [37,63]	66 [52,78]	36 [24,50]
Gaussian DM	18 [10,31]	72 [58,83]	100 [93,100]	100 [93,100]
Gaussian AIPW	14 [7,26]	70 [56,81]	100 [93,100]	100 [93,100]
Janzing KL	88 [76,94]	100 [93,100]	100 [93,100]	100 [93,100]
Balanced JS	92 [81,97]	100 [93,100]	100 [93,100]	100 [93,100]
Total variation	52 [39,65]	100 [93,100]	100 [93,100]	100 [93,100]

Table 13: All 50-replicate conditions for components, shape. Entries are rejection percentages and 95% Wilson Monte Carlo intervals; these are not biological intervals.

Method	$1k, .05$	$1k, .20$	$10k, .05$	$10k, .20$
GeoRARE $H_{0:2}$	6 [2,16]	10 [4,21]	0 [0,7]	0 [0,7]
GeoRARE Euler	6 [2,16]	10 [4,21]	0 [0,7]	0 [0,7]
GeoRARE AET	6 [2,16]	10 [4,21]	0 [0,7]	0 [0,7]
GeoRARE Hodge	6 [2,16]	10 [4,21]	0 [0,7]	0 [0,7]
Adjusted mean	4 [1,13]	2 [0,10]	10 [4,21]	2 [0,10]
Quantile grid	0 [0,7]	10 [4,21]	0 [0,7]	0 [0,7]
Gaussian DM	4 [1,13]	4 [1,13]	16 [8,29]	76 [63,86]
Gaussian AIPW	6 [2,16]	8 [3,19]	16 [8,29]	74 [60,84]
Janzing KL	8 [3,19]	26 [16,40]	76 [63,86]	100 [93,100]
Balanced JS	2 [0,10]	36 [24,50]	72 [58,83]	100 [93,100]
Total variation	2 [0,10]	22 [13,35]	60 [46,72]	100 [93,100]

Table 14: All 50-replicate conditions for components, density. Entries are rejection percentages and 95% Wilson Monte Carlo intervals; these are not biological intervals.

Method	$1k, .05$	$1k, .20$	$10k, .05$	$10k, .20$
GeoRARE $H_{0:2}$	0 [0,7]	0 [0,7]	0 [0,7]	0 [0,7]
GeoRARE Euler	0 [0,7]	0 [0,7]	0 [0,7]	0 [0,7]
GeoRARE AET	0 [0,7]	0 [0,7]	0 [0,7]	0 [0,7]
GeoRARE Hodge	0 [0,7]	0 [0,7]	0 [0,7]	0 [0,7]
Adjusted mean	50 [37,63]	96 [87,99]	100 [93,100]	100 [93,100]
Quantile grid	0 [0,7]	66 [52,78]	96 [87,99]	100 [93,100]
Gaussian DM	52 [39,65]	94 [84,98]	100 [93,100]	100 [93,100]
Gaussian AIPW	36 [24,50]	88 [76,94]	100 [93,100]	100 [93,100]
Janzing KL	50 [37,63]	90 [79,96]	100 [93,100]	100 [93,100]
Balanced JS	42 [29,56]	90 [79,96]	100 [93,100]	100 [93,100]
Total variation	48 [35,61]	94 [84,98]	100 [93,100]	100 [93,100]

Table 15: All 50-replicate conditions for components, mean. Entries are rejection percentages and 95% Wilson Monte Carlo intervals; these are not biological intervals.

Method	$1k, .05$	$1k, .20$	$10k, .05$	$10k, .20$
GeoRARE $H_{0:2}$	2 [0,10]	0 [0,7]	0 [0,7]	0 [0,7]
GeoRARE Euler	2 [0,10]	0 [0,7]	0 [0,7]	0 [0,7]
GeoRARE AET	2 [0,10]	0 [0,7]	0 [0,7]	0 [0,7]
GeoRARE Hodge	4 [1,13]	0 [0,7]	0 [0,7]	0 [0,7]
Adjusted mean	10 [4,21]	8 [3,19]	4 [1,13]	4 [1,13]
Quantile grid	0 [0,7]	0 [0,7]	2 [0,10]	0 [0,7]
Gaussian DM	10 [4,21]	6 [2,16]	6 [2,16]	4 [1,13]
Gaussian AIPW	12 [6,24]	6 [2,16]	6 [2,16]	4 [1,13]
Janzing KL	10 [4,21]	4 [1,13]	4 [1,13]	6 [2,16]
Balanced JS	4 [1,13]	2 [0,10]	4 [1,13]	2 [0,10]
Total variation	6 [2,16]	2 [0,10]	2 [0,10]	4 [1,13]

Table 16: All 50-replicate conditions for loop, null. Entries are rejection percentages and 95% Wilson Monte Carlo intervals; these are not biological intervals.

Method	$1k, .05$	$1k, .20$	$10k, .05$	$10k, .20$
GeoRARE $H_{0:2}$	76 [63,86]	98 [90,100]	100 [93,100]	100 [93,100]
GeoRARE Euler	76 [63,86]	98 [90,100]	100 [93,100]	100 [93,100]
GeoRARE AET	74 [60,84]	98 [90,100]	100 [93,100]	100 [93,100]
GeoRARE Hodge	72 [58,83]	98 [90,100]	100 [93,100]	100 [93,100]
Adjusted mean	10 [4,21]	4 [1,13]	6 [2,16]	8 [3,19]
Quantile grid	0 [0,7]	0 [0,7]	54 [40,67]	98 [90,100]
Gaussian DM	22 [13,35]	66 [52,78]	100 [93,100]	100 [93,100]
Gaussian AIPW	16 [8,29]	60 [46,72]	100 [93,100]	100 [93,100]
Janzing KL	82 [69,90]	100 [93,100]	100 [93,100]	100 [93,100]
Balanced JS	82 [69,90]	100 [93,100]	100 [93,100]	100 [93,100]
Total variation	48 [35,61]	100 [93,100]	100 [93,100]	100 [93,100]

Table 17: All 50-replicate conditions for loop, shape. Entries are rejection percentages and 95% Wilson Monte Carlo intervals; these are not biological intervals.

Method	$1k, .05$	$1k, .20$	$10k, .05$	$10k, .20$
GeoRARE $H_{0:2}$	12 [6,24]	6 [2,16]	0 [0,7]	0 [0,7]
GeoRARE Euler	12 [6,24]	6 [2,16]	0 [0,7]	0 [0,7]
GeoRARE AET	12 [6,24]	6 [2,16]	0 [0,7]	0 [0,7]
GeoRARE Hodge	12 [6,24]	6 [2,16]	0 [0,7]	0 [0,7]
Adjusted mean	10 [4,21]	10 [4,21]	12 [6,24]	6 [2,16]
Quantile grid	0 [0,7]	0 [0,7]	20 [11,33]	0 [0,7]
Gaussian DM	10 [4,21]	10 [4,21]	14 [7,26]	76 [63,86]
Gaussian AIPW	10 [4,21]	10 [4,21]	18 [10,31]	74 [60,84]
Janzing KL	16 [8,29]	18 [10,31]	62 [48,74]	100 [93,100]
Balanced JS	16 [8,29]	20 [11,33]	56 [42,69]	100 [93,100]
Total variation	18 [10,31]	20 [11,33]	52 [39,65]	100 [93,100]

Table 18: All 50-replicate conditions for loop, density. Entries are rejection percentages and 95% Wilson Monte Carlo intervals; these are not biological intervals.

Method	$1k$, .05	$1k$, .20	$10k$, .05	$10k$, .20
GeoRARE $H_{0:2}$	0 [0,7]	0 [0,7]	0 [0,7]	0 [0,7]
GeoRARE Euler	0 [0,7]	0 [0,7]	0 [0,7]	0 [0,7]
GeoRARE AET	2 [0,10]	4 [1,13]	2 [0,10]	0 [0,7]
GeoRARE Hodge	0 [0,7]	4 [1,13]	2 [0,10]	0 [0,7]
Adjusted mean	10 [4,21]	10 [4,21]	8 [3,19]	2 [0,10]
Quantile grid	0 [0,7]	0 [0,7]	0 [0,7]	0 [0,7]
Gaussian DM	8 [3,19]	8 [3,19]	8 [3,19]	4 [1,13]
Gaussian AIPW	2 [0,10]	6 [2,16]	8 [3,19]	4 [1,13]
Janzing KL	10 [4,21]	4 [1,13]	6 [2,16]	10 [4,21]
Balanced JS	4 [1,13]	6 [2,16]	6 [2,16]	2 [0,10]
Total variation	4 [1,13]	10 [4,21]	6 [2,16]	4 [1,13]

Table 19: All 50-replicate conditions for void, null. Entries are rejection percentages and 95% Wilson Monte Carlo intervals; these are not biological intervals.

Method	$1k$, .05	$1k$, .20	$10k$, .05	$10k$, .20
GeoRARE $H_{0:2}$	48 [35,61]	92 [81,97]	100 [93,100]	100 [93,100]
GeoRARE Euler	48 [35,61]	92 [81,97]	100 [93,100]	100 [93,100]
GeoRARE AET	30 [19,44]	4 [1,13]	0 [0,7]	0 [0,7]
GeoRARE Hodge	12 [6,24]	96 [87,99]	100 [93,100]	100 [93,100]
Adjusted mean	10 [4,21]	8 [3,19]	4 [1,13]	12 [6,24]
Quantile grid	0 [0,7]	0 [0,7]	0 [0,7]	0 [0,7]
Gaussian DM	18 [10,31]	76 [63,86]	100 [93,100]	100 [93,100]
Gaussian AIPW	20 [11,33]	68 [54,79]	100 [93,100]	100 [93,100]
Janzing KL	76 [63,86]	100 [93,100]	100 [93,100]	100 [93,100]
Balanced JS	88 [76,94]	100 [93,100]	100 [93,100]	100 [93,100]
Total variation	52 [39,65]	100 [93,100]	100 [93,100]	100 [93,100]

Table 20: All 50-replicate conditions for void, shape. Entries are rejection percentages and 95% Wilson Monte Carlo intervals; these are not biological intervals.

Method	$1k$, .05	$1k$, .20	$10k$, .05	$10k$, .20
GeoRARE $H_{0:2}$	8 [3,19]	2 [0,10]	0 [0,7]	0 [0,7]
GeoRARE Euler	8 [3,19]	2 [0,10]	0 [0,7]	0 [0,7]
GeoRARE AET	10 [4,21]	0 [0,7]	0 [0,7]	0 [0,7]
GeoRARE Hodge	4 [1,13]	2 [0,10]	0 [0,7]	0 [0,7]
Adjusted mean	4 [1,13]	6 [2,16]	6 [2,16]	4 [1,13]
Quantile grid	0 [0,7]	0 [0,7]	0 [0,7]	0 [0,7]
Gaussian DM	2 [0,10]	8 [3,19]	16 [8,29]	92 [81,97]
Gaussian AIPW	4 [1,13]	6 [2,16]	20 [11,33]	92 [81,97]
Janzing KL	6 [2,16]	10 [4,21]	62 [48,74]	100 [93,100]
Balanced JS	10 [4,21]	10 [4,21]	68 [54,79]	100 [93,100]
Total variation	12 [6,24]	10 [4,21]	60 [46,72]	100 [93,100]

Table 21: All 50-replicate conditions for void, density. Entries are rejection percentages and 95% Wilson Monte Carlo intervals; these are not biological intervals.

Method / scenario	Bias	RMSE	True ATE
DoWhy OLS / null	-0.0016	0.0784	0.00
DoWhy IPW / null	-0.0012	0.0931	0.00
DoubleML ATE / null	-0.0036	0.1860	0.00
DoWhy OLS / shape	-0.0047	0.0907	0.00
DoWhy IPW / shape	-0.0063	0.0977	0.00
DoubleML ATE / shape	-0.0070	0.1574	0.00
DoWhy OLS / density	+0.0050	0.0823	0.00
DoWhy IPW / density	+0.0056	0.0897	0.00
DoubleML ATE / density	+0.0003	0.1569	0.00
DoWhy OLS / mean	-0.0046	0.0615	0.30
DoWhy IPW / mean	-0.0082	0.0645	0.30
DoubleML ATE / mean	-0.0042	0.1086	0.30

Table 22: Actual-package mean-effect estimation on identical scalar records, 200 datasets per scenario, equally pooling the four sample-size/frequency settings. DoWhy IPW has point estimates and no uncomputed significance test.

L MATCHED WALD AND STRUCTURAL-CERTIFICATION COMPARISONS

CDF Wald and discrete Betti procedures can differ in nuisance fitting, studentization, and calibration as well as representation. To separate these factors, we recompute both statistics on the same 800 scalar datasets from Appendix K, and evaluate them on a separate 2,000-dataset scalar simulation with disjoint seeds. Recalculation on the 800 shared datasets is a matched analysis of those observations, not an additional independent sample. "DoubleML CDF Wald" denotes our joint test constructed from actual DoubleML IRM scores, not a built-in package estimator.

A direct Wald test on a hard Betti plug-in estimate is generally invalid: its derivative is zero away from crossings and undefined at crossings. Instead we derive a differentiable integrated-Euler target and a separate improvement to inference for the fixed-threshold target (Appendix L.1). On the path inside the declared center-minimum chamber, the integrated Euler signature is $F_K(p) = 1 - 2p_c$. Its signed contrast has a one-degree Wald test using the same DoubleML CDF scores and covariance. It equals a center-CDF Wald statistic on every executed dataset; no exclusive topological information gain is claimed. A second comparison uses common refitted saturated scores and shared conditional permutations for both CDF and Euler statistics.

The integrated-Euler Wald test attains approximately 98% shape detection, but does not demonstrate superiority over the matched CDF test. It changes the signature: unlike fixed-threshold Betti, integrated Euler responds to some density changes inside the same Betti chamber. We do not report those as detection improvements for the fixed-threshold target.

For the fixed-threshold target, a finite-sample intersection–union procedure instead certifies the required probability inequalities, avoiding an unnecessary multiplicity penalty across conjunctive conditions. It raises positive certification from 54.25% to 68.00% on the 400 Shape datasets: paired gain 13.75 percentage points, stratified Monte Carlo interval $[11.30, 16.20]$. This passes the declared improvement criterion. Both procedures produce zero false positive certificates on 1,600 Null, Density, Mean and Boundary datasets. At Boundary the true fixed-threshold contrast is zero even though the outcome distributions differ. The fixed-threshold Betti permutation statistic rejects in 42% there, correctly illustrating that its distributional-null rejection is not itself a structural certificate. The improvement concerns positive certification, not uniformly tighter confidence sets or biological findings.

L.1 INFERENCE TARGETS, ASSUMPTIONS, AND PROOFS

Software attribution and calibration. The "DoubleML CDF Wald" score is a joint test constructed in this project from actual DoubleML IRM fits of CDF indicators. It is not a native built-in DoubleML CDF-Wald estimator. Its 98% detection rate accompanied an 8.5% no-effect rejection rate on 200 datasets per scenario in the scalar comparison. A comparison with a discrete, conservatively calibrated Betti statistic did not isolate the effects of representation, nuisance fitting, covariance normalization, and null calibration. The software implementation and the calibration must therefore be specified separately from the choice of structural representation.

Why the CDF statistic can detect a zero-mean change. In outcome order $(-1, +1, 0)$ let $p_a = (p_{a,-}, p_{a,+}, p_{a,c})$ be the interventional probability vector for arm a. Its two CDF coordinates are $f_a = (p_{a,-}, p_{a,-} + p_{a,c})$. Thus the signed vector $\delta = f_1 - f_0$ determines every probability difference, although the coordinate-mean contrast can be zero. If $\sqrt{n}(\widehat{\delta} - \delta) \Rightarrow N(0, \Sigma)$ and $\widehat{\Sigma} \to \Sigma$ with fixed rank s, the null statistic

$$W_F = n\widehat{\delta}^T \widehat{\Sigma}^+ \widehat{\delta} \;\Rightarrow\; \chi^2_s$$

uses the covariance to account for differing noise levels and correlated CDF coordinates. Here $+$ denotes the Moore–Penrose inverse. This is an asymptotic statement, not a finite-sample size guarantee for rare arms. For rank-deficient covariance it additionally requires the contrast to lie in the limiting covariance range and a stable rank; our regular benchmark has rank two. Under regular local alternatives $\delta = h/\sqrt{n}$, the limit is noncentral chi-square with noncentrality $h^T \Sigma^+ h$.

Proposition 16 (A differentiable Euler signature)**.** *For a finite simplicial complex K with vertex probabilities p_v, let $K_t(p)$ be the induced subcomplex on vertices satisfying $p_v \geq t$. Its integrated*

Method	Null	Shape	Boundary
DoubleML-score CDF Wald	7.00	98.50	81.75
DoubleML-score Euler Wald	6.50	98.25	84.00
CDF Wald, permutations	3.50	97.75	79.25
Euler Wald, permutations	4.00	98.00	82.00
Full-box CP certificate	0.00	54.25	0.00
Exact structural IUT	0.00	68.00	0.00

Table 23: Results, 400 datasets per column, equally pooling four sample-size/frequency conditions. The first four rows test distributional or Euler effects; Boundary is an alternative for those targets. The last two rows certify the unchanged fixed-threshold structural effect, which is zero at Boundary. Thus rates across these two blocks answer different questions.

Euler signature is

$$F_K(p) = \int_0^1 \chi(K_t(p))\, dt = \sum_{\emptyset \neq \sigma \in K} (-1)^{\dim \sigma} \min_{v \in \sigma} p_v.$$

It is continuous and piecewise linear. In a neighborhood where every simplex minimum is attained at a unique, unchanged vertex v_σ, its derivative is the constant vector $g_K = \sum_{\sigma \in K} (-1)^{\dim \sigma} e_{v_\sigma}$. On the three-vertex path, if the center probability is strictly below both outer probabilities,

$$F_K(p) = 1 - 2p_c, \qquad \Delta_F = F_K(p_1) - F_K(p_0) = \ell^T \delta, \quad \ell = (2, -2)^T.$$

Consequently $W_E = n(\ell^T \widehat{\delta})^2 / (\ell^T \widehat{\Sigma} \ell)$ has a χ^2_1 limit under $\Delta_F = 0$, if its denominator has a strictly positive limit and the stated joint CLT holds.

Proof. A simplex belongs to K_t precisely for $0 \leq t \leq \min_{v \in \sigma} p_v$. Integrate its signed indicator and sum finitely many terms. Unique minimizers remain unchanged under sufficiently small perturbations, proving local linearity. The path has three vertices and two edges, so $F_K(p) = \sum_v p_v - \min(p_-, p_c) - \min(p_+, p_c)$. In the indicated chamber this is $1 - 2p_c$. Since $p_{a,c} = f_{a,2} - f_{a,1}$, the contrast and its variance follow by linear transformation. The normal limit and Slutsky's theorem give the Wald limit. □

All declared scalar populations, including the threshold-boundary control, lie inside this center-minimum chamber. This is a model-class assumption justified from the declared simulation laws, not a claim that an arbitrary real outcome law satisfies it. At simplex-minimum ties outside such a differentiable chamber, an ordinary delta-method Wald test need not be valid. We therefore do not extend this calculation to arbitrary higher-dimensional persistent homology without additional analysis.

What is and is not improved by this signature. The integrated-Euler GeoRARE target is $\Theta_F(d) = 2q_d(1 - q_d)\Delta_F^2$. It is distinct from the fixed-threshold Betti target. Inference is performed on the signed Δ_F, not by applying a first-order Wald approximation to Θ_F at zero, where that derivative vanishes. The hard Betti signature itself is locally constant away from threshold crossings and discontinuous at a crossing; applying an ordinary Wald test to its direct plug-in estimate is not justified.

The one-dimensional center-CDF test has exactly the same statistic as W_E, because its contrast is $-\Delta_F/2$ and its variance is one quarter as large. This identity is verified on every executed dataset. Any detection gain from this construction is therefore not exclusive to topological information. It combines a declared structural summary with a targeted one-dimensional test. It can have greater power than a two-dimensional omnibus test in an aligned direction, while missing alternatives in other directions. Density changes that preserve the fixed-threshold Betti signature can still change the integrated Euler signature. These are reported as changes in the integrated-Euler target, not as successful detection of the fixed-threshold target.

Matched score and calibration experiments. The matched 800-dataset analysis uses identical DoubleML tree/logistic nuisance fits and folds for both Wald statistics. It recomputes the joint score and covariance, rather than merely copying the 98% number. For an additional common-randomization comparison, both CDF and Euler statistics use identical cross-fitted saturated categorical nuisance fits, half-count smoothing, propensity truncation at .005, individual-score covariance, and 499 shared treatment permutations within (W, fold) strata. All nuisances and covariance matrices are refitted for each permutation. This vectorized implementation is labeled separately from the actual DoubleML nuisance fits. Its validity follows from Proposition 15 under the full conditional distributional null. Such calibration does not establish validity under every weaker Euler-only null in a larger model class.

Proposition 17 (Exact positive certification for the fixed-threshold Betti target)**.** *Assume consistency, positivity, ignorability given binary observed W, known reference $\mathbb{P}(W = w) = 1/2$, and the three-vertex path at threshold $c = .1$. The following orientation is sufficient for a nonzero component contrast:*

$$p_{0,c} > c, \quad p_{1,c} < c, \quad p_{1,-} > c, \quad p_{1,+} > c.$$

There is a second orientation with treatment labels exchanged. For each single inequality, construct one-sided Clopper–Pearson bounds in both W strata with tail probability $\alpha/4$, and average their endpoints. Accept an orientation only when all four required strict inequalities are certified. Accept a positive effect if either orientation is accepted. The probability of a false positive structural certificate is at most α in finite samples, including at threshold boundaries.

Proof. For one probability and one direction, two conditional binomial CP bounds cover with probability at least $1 - \alpha/2$ by a union bound. Averaging covered endpoints preserves the bound for the standardized arm probability. If an orientation is false, at least one of its four inequalities is false. Acceptance of all four requires falsely certifying that particular false inequality, whose probability is at most $\alpha/2$. Thus no Bonferroni penalty across the four conjunctive inequalities is necessary: this is an intersection–union test (Berger & Hsu, 1996). A union bound over the two orientations yields α. Conditional bounds remain valid with random stratum sample sizes; an empty stratum receives bounds $[0, 1]$. The indicated probabilities produce one connected arm and one arm with two disconnected outer vertices. Therefore acceptance certifies the same fixed-threshold Betti contrast, not a substituted integrated-Euler target. □

This procedure specializes inference to positive structural effects. It is not a uniformly tighter replacement for the complete confidence set: it does not, for example, certify absence of an effect. Since on this path $\beta_0 \in \{1, 2\}$, a valid interval is $[.5, .5]$ upon positive certification and $[0, .5]$ otherwise for $\Theta_{H_0}(0)$. Both finite methods use the same known W reference. No additional biological data or oracle outcome probabilities are supplied to the intersection–union test. An analogous Wald IUT using estimated one-sided normal bounds is also executed, but remains asymptotic and is assessed separately at the structural boundary.

Regularity, degrees of freedom, and attribution of a gain. For cross-fitted AIPW scores, a sufficient regularity condition is bounded positivity, finite second moments, consistent nuisances, and the product-rate condition $\|\widehat{g} - g\|_2 \|\widehat{e} - e\|_2 = o_p(n^{-1/2})$. The finite observed-confounder model with positive category probabilities provides a regular setting for these asymptotics; it does not remove small-sample normal-approximation errors. For positive definite Σ, Cauchy–Schwarz applied to $\Sigma^{-1/2}\delta$ and $\Sigma^{1/2}\ell$ gives

$$\frac{(\ell^T \delta)^2}{\ell^T \Sigma \ell} \leq \delta^T \Sigma^{-1} \delta.$$

The projected statistic can nevertheless reject when the omnibus does not, because its one-degree critical value is smaller. It can also miss effects orthogonal to its chosen contrast. Covariance normalization and a more focused alternative explain this possibility; topology does not create additional information in this scalar example.

Connection to the hidden-confounding identification result. In Proposition 14, the two interventional center probabilities remain .30 and .02 over the full parameter rectangle, and both centers remain strictly below the outer probabilities. Hence $F_K(P_0) = .40$, $F_K(P_1) = .96$, and

$\Theta_F(0) = \frac{1}{2}(.56)^2 = .1568$ throughout that class, although the ATE identified set remains $[-.4, .4]$. This follows from explicit invariance of this particular Euler signature. Ignorability of an arbitrary hard Betti summary does not imply ignorability of the integrated Euler signature: the latter can change within a fixed-Betti chamber. The observed-W simulations assume full-law ignorability, whereas the hidden-class corollary uses its own stated invariant. This distinction is required when applying the framework of Faghihi & Saki (2026) to the integrated-Euler target.

L.2 Complete condition-level results

Method	$1k, .05$	$1k, .20$	$10k, .05$	$10k, .20$
DoubleML-score CDF Wald	15.0	9.0	1.0	3.0
DoubleML-score Euler Wald	9.0	9.0	6.0	2.0
CDF Wald, permutations	5.0	5.0	1.0	3.0
Euler Wald, permutations	5.0	6.0	3.0	2.0
Betti, permutations	1.0	0.0	0.0	0.0
Full-box CP certificate	0.0	0.0	0.0	0.0
Exact structural IUT	0.0	0.0	0.0	0.0
Structural Wald IUT	0.0	0.0	0.0	0.0

Table 24: Null: all condition-level rejection or positive-certificate percentages, 100 repetitions each. Distributional testing and fixed-threshold certification have different null hypotheses.

Method	$1k, .05$	$1k, .20$	$10k, .05$	$10k, .20$
DoubleML-score CDF Wald	94.0	100.0	100.0	100.0
DoubleML-score Euler Wald	93.0	100.0	100.0	100.0
CDF Wald, permutations	91.0	100.0	100.0	100.0
Euler Wald, permutations	92.0	100.0	100.0	100.0
Betti, permutations	91.0	95.0	100.0	100.0
Full-box CP certificate	0.0	17.0	100.0	100.0
Exact structural IUT	0.0	72.0	100.0	100.0
Structural Wald IUT	50.0	100.0	100.0	100.0

Table 25: Shape: all condition-level rejection or positive-certificate percentages, 100 repetitions each. Distributional testing and fixed-threshold certification have different null hypotheses.

Method	$1k, .05$	$1k, .20$	$10k, .05$	$10k, .20$
DoubleML-score CDF Wald	30.0	33.0	79.0	100.0
DoubleML-score Euler Wald	36.0	44.0	86.0	100.0
CDF Wald, permutations	18.0	32.0	76.0	100.0
Euler Wald, permutations	22.0	40.0	86.0	100.0
Betti, permutations	14.0	8.0	1.0	0.0
Full-box CP certificate	0.0	0.0	0.0	0.0
Exact structural IUT	0.0	0.0	0.0	0.0
Structural Wald IUT	2.0	0.0	0.0	0.0

Table 26: Density: all condition-level rejection or positive-certificate percentages, 100 repetitions each. Distributional testing and fixed-threshold certification have different null hypotheses.

Method	$1k, .05$	$1k, .20$	$10k, .05$	$10k, .20$
DoubleML-score CDF Wald	35.0	92.0	100.0	100.0
DoubleML-score Euler Wald	13.0	9.0	6.0	6.0
CDF Wald, permutations	21.0	91.0	100.0	100.0
Euler Wald, permutations	6.0	8.0	6.0	6.0
Betti, permutations	3.0	0.0	0.0	0.0
Full-box CP certificate	0.0	0.0	0.0	0.0
Exact structural IUT	0.0	0.0	0.0	0.0
Structural Wald IUT	0.0	0.0	0.0	0.0

Table 27: Mean: all condition-level rejection or positive-certificate percentages, 100 repetitions each. Distributional testing and fixed-threshold certification have different null hypotheses.

Method	$1k, .05$	$1k, .20$	$10k, .05$	$10k, .20$
DoubleML-score CDF Wald	47.0	80.0	100.0	100.0
DoubleML-score Euler Wald	50.0	86.0	100.0	100.0
CDF Wald, permutations	40.0	77.0	100.0	100.0
Euler Wald, permutations	45.0	83.0	100.0	100.0
Betti, permutations	29.0	42.0	43.0	54.0
Full-box CP certificate	0.0	0.0	0.0	0.0
Exact structural IUT	0.0	0.0	0.0	0.0
Structural Wald IUT	13.0	4.0	2.0	4.0

Table 28: Boundary: all condition-level rejection or positive-certificate percentages, 100 repetitions each. Distributional testing and fixed-threshold certification have different null hypotheses.

M ADJUSTED-LAW EXPERIMENT PROTOCOL

M.1 KNOWN CAUSAL LAWS WITH OBSERVED CONFOUNDING

The outcome lies on a three-vertex connected path, a five-vertex filled cycle, or a seven-vertex filled octahedral sphere. $Z \sim \mathrm{Bernoulli}(0.2)$ and $W \sim \mathrm{Bernoulli}(0.5)$ independently. Treatment probabilities are $(0.45, 0.35, 0.20)$ when $W = 0$ and $(0.70, 0.28, 0.02)$ when $W = 1$. The center mass is 0.02 or 0.30 according to W, with the remaining mass equally distributed over outer vertices. Thus adjusted center mass is 0.16, while selection can change observed topology. In the rare-effect condition only $(Z, A) = (1, 3)$ changes, with center probabilities $(0.01, 0.03)$. In the boundary condition these are $(0.09, 0.11)$, giving center mass exactly 0.1. The symmetric coordinate mean stays zero; zero mean effect is the correct answer to a different estimand, not a failed mean estimator.

We run 100 datasets for each of 3 shapes, 3 scenarios and 2 sample sizes ($2,000$ and $10,000$): 1,800 datasets, each shared across all nine estimators. Correct nuisance models use saturated (Z, W, A) tables; omitted models remove W, with fixed pseudocount 0.5 and propensity floor 0.005. Full condition-level results are in the appendix. At $n = 10,000$, adjusted point estimators substantially reduce structural error relative to ignoring W. Simple adjusted outcome regression is often better than AIPW at these sample sizes. AIPW provides robustness properties but does not guarantee lower finite-sample error in these settings.

Finite bands start with Bonferroni–Clopper–Pearson bounds for every (Z, W, A) outcome probability, then standardize with the known W reference $(1/2, 1/2)$ before applying Theorem 11. Zero-observation strata receive $[0, 1]$. All 1,800 families covered their targets, but *none* certified a positive effect, and neither full signature nor active-set certification succeeded for all six (Z, A) cells. For 100 successes in one condition, the two-sided exact 95% lower confidence limit for coverage is 96.38%, not 100% certainty. The result exposes conservative uncertainty under a rare (Z, W, A) combination of probability 0.002; it does not establish useful power. Increasing n narrowed intervals without resolving this gap. The analytic strict advantage in Figure 13 should not be confused with an empirical certification advantage in these runs.

M.2 COMPLETE ESTIMATION AND COVERAGE RESULTS

Estimator	Law RMSE	Null	Rare	Boundary
AIPW_both_adjusted	0.0750	0.1049	0.1038	0.1116
AIPW_both_omit_W	0.0485	0.2729	0.2091	0.2505
AIPW_outcome_omit_W	0.0720	0.1000	0.0966	0.1041
AIPW_propensity_omit_W	0.0420	0.1012	0.0972	0.1053
DM_adjusted	0.0370	0.0933	0.0897	0.0863
DM_omit_W	0.0440	0.2726	0.2110	0.2462
IPW_adjusted	0.0686	0.1037	0.1020	0.1092
IPW_omit_W	0.0477	0.2745	0.2104	0.2523
unadjusted	0.0453	0.2726	0.2109	0.2497

Table 29: All nine estimators at $n = 2,000$. Law RMSE averages equally over all nine shape/scenario conditions; the last three columns give structural surface RMSE averaged equally over three shapes. Every constituent condition and replicate is available in the supplied CSV files.

Estimator	Law RMSE	Null	Rare	Boundary
AIPW_both_adjusted	0.0276	0.0381	0.0423	0.0646
AIPW_both_omit_W	0.0354	0.2908	0.2173	0.2688
AIPW_outcome_omit_W	0.0266	0.0373	0.0385	0.0612
AIPW_propensity_omit_W	0.0224	0.0385	0.0255	0.0598
DM_adjusted	0.0219	0.0373	0.0249	0.0560
DM_omit_W	0.0348	0.2912	0.2175	0.2653
IPW_adjusted	0.0260	0.0388	0.0394	0.0633
IPW_omit_W	0.0354	0.2908	0.2173	0.2693
unadjusted	0.0350	0.2912	0.2173	0.2672

Table 30: All nine estimators at $n = 10,000$. Law RMSE averages equally over all nine shape/scenario conditions; the last three columns give structural surface RMSE averaged equally over three shapes. Every constituent condition and replicate is available in the supplied CSV files.

Condition	Coverage	Width	Width, $\delta = .02$	Positive
Components/null/2000	1.00	0.404	0.535	0.00
Components/null/10000	1.00	0.240	0.321	0.00
Components/rare_shape/2000	1.00	0.407	0.548	0.00
Components/rare_shape/10000	1.00	0.238	0.325	0.00
Components/threshold/2000	1.00	0.402	0.542	0.00
Components/threshold/10000	1.00	0.243	0.315	0.00
Loop/null/2000	1.00	0.523	0.815	0.00
Loop/null/10000	1.00	0.265	0.363	0.00
Loop/rare_shape/2000	1.00	0.555	0.823	0.00
Loop/rare_shape/10000	1.00	0.258	0.368	0.00
Loop/threshold/2000	1.00	0.572	0.825	0.00
Loop/threshold/10000	1.00	0.252	0.367	0.00
Void/null/2000	1.00	0.599	0.920	0.00
Void/null/10000	1.00	0.257	0.400	0.00
Void/rare_shape/2000	1.00	0.581	0.899	0.00
Void/rare_shape/10000	1.00	0.265	0.392	0.00
Void/threshold/2000	1.00	0.590	0.916	0.00
Void/threshold/10000	1.00	0.263	0.413	0.00

Table 31: All finite-band conditions (100 datasets each). Coverage is joint over the escort grid. Positive is the rate of any positive lower bound. Full-signature and full-active-set certification rates are zero in all conditions. These results do not demonstrate useful certification power.

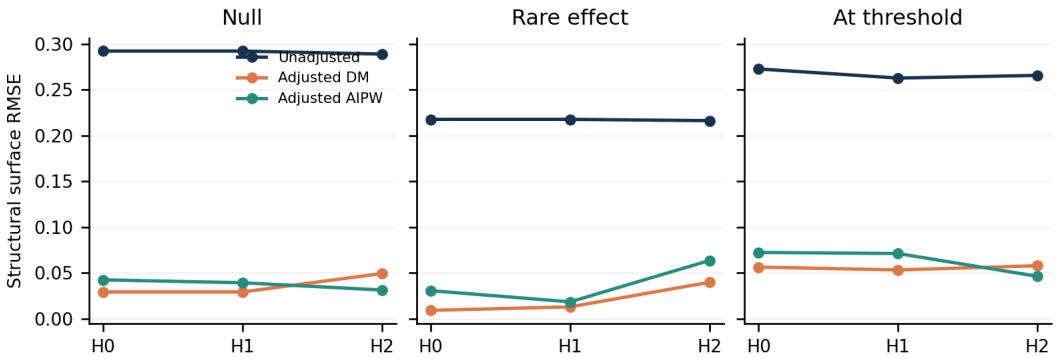


Figure 15: Same-target structural RMSE at $n = 10,000$, over the complete 16-point escort grid. Each point averages 100 datasets. The adjusted estimators use observed confounder W; unadjusted estimation omits it. The boundary scenario remains difficult despite accurate law estimation.

N FINITE-PANEL INTERVENTION PROFILES

N.1 REAL INTERVENTION PROFILES UNDER CONTROLLED MISSING WELLS

We use sci-Plex (Srivatsan et al., 2020), including every compound with all $2 \times 3 \times 4$ replicate–cell-line–dose wells in the supplied surviving-well cache: 157 compounds and 3,768 wells. Each well contributes the same 12 selected cells in 20 PCA coordinates. The PCA and geometric scales are fixed from replicate-1 data that were examined before this analysis. Feature standardizers also use replicate 1 only. The missing-well experiment conditions on these calibrations. Dose prediction is not the task. The reference outcome is a finite-panel profile of squared *within-compound* dose-pair descriptor distances in replicate 2.

The experiment includes all 15 representations: moments; Euclidean H_0, H_1, H_2 and their concatenation; local Gaussian-RKHS KH_0, KH_1, KH_2 and concatenation; exact Euler3; anchored Euler transform; Euclidean and local-RKHS Hodge 0–2; their specified geometric combination; and exact empirical Gaussian mean-embedding distance. Every representation uses the same wells and cell indices. The 252,456 critical Euler events are recorded and exactly reproduce the fitted grids. Persistence is exact up to its declared 95th-percentile training-distance cap, with remaining bars right-censored. Euler files extend through every critical distance. Finite Rips 3-skeleton enumeration includes tetrahedra; higher simplices are unnecessary for H_0, H_1, H_2 but would be necessary for higher-dimensional topology.

Whole wells are removed independently with known retention probabilities. One design retains each well with probability 0.5; the rare design uses probabilities ranging from 0.765 to 0.06. There are 500 masks per design, shared across all representations, estimators and targets. We report both the balanced reference and a *hypothetical* reference with $p = (0.55, 0.25, 0.15, 0.05)$, $h = (0.7, 0.2, 0.1)$. Balanced reference surfaces are exactly flat. Imposed missingness and the hypothetical reference are not evidence of naturally rare treatment allocation.

For compound i and dose pair $j = (a, b)$, let Y_{izj} be its observed replicate-2 descriptor distance, $M_{izj} = R_{iza}R_{izb}$ indicate retaining both wells, $\pi_{zj} = \rho_{za}\rho_{zb}$, and $\tilde{Y}_{izj}$ the corresponding replicate-1 distance. With $w_{zj} = 2\nu_r(z)q_d(a \mid z)q_d(b \mid z)$, the replicate-assisted estimator is

$$\widehat{T}(d, r) = \frac{1}{N} \sum_{i,z,j} w_{zj} \left\{ \tilde{Y}_{izj} + \frac{M_{izj}}{\pi_{zj}} (Y_{izj} - \tilde{Y}_{izj}) \right\}. \tag{45}$$

It is design-unbiased for the fixed-panel target, regardless of replicate-1 prediction quality. This is standard augmentation of an unequal-probability estimator (Horvitz & Thompson, 1952), applied to GeoRARE's pair profile. Baselines are Horvitz–Thompson (HT), stratum-wise Hájek, and pooled observed pairs. Pair indicators sharing a well are dependent. Our exact variance accounts for this dependence and gives a necessary and sufficient condition for replicate assistance to improve design variance (Appendix N.2).

Representation	HT	Hajek	Assisted	MSE reduction
moments	0.09706	0.03274	0.02838	+91.5%
H012	0.35683	0.34423	0.36860	-6.7%
KH012	0.35669	0.34080	0.38478	-16.4%
Hodge012	0.57309	0.56539	0.56849	+1.6%
full_geometry	0.37275	0.34429	0.35120	+11.2%
Gaussian	0.00653	0.00387	0.00211	+89.6%

Table 32: Finite-panel RMSE under rare missingness and the hypothetical reference, averaged over the declared grid. Reduction compares assistance to HT within the same representation; each row has a different target. All 15 representations and both designs/references appear in the appendix.

The geometric combination improves its own finite-panel MSE modestly, while moments and Gaussian geometry benefit much more. Many individual topological representations become worse with replicate assistance; their replicate-1 descriptors are weak predictors of replicate-2 pair distances. The simple Hájek baseline is competitive and sometimes better. These findings do not show a unique real-world advantage of topology. Finite-panel design unbiasedness is exact; population biological

causal identification, uncertainty over independent biological experiments, and generalization to other compounds are not established by this missingness exercise. Cells within a well are never counted as independent treatment assignments.

N.2 Finite-panel unbiasedness and exact variance

Condition on the full measured panel, the replicate-1 predictor and all calibrations. Let independent well indicators have probabilities ρ_{za}. For pair j, $\mathbb{E}(M_{izj}/\pi_{zj}) = 1$, hence expectation of each bracket in (45) equals Y_{izj}, proving design unbiasedness for every escort target. Conditional causal identification of a population law is not part of this argument.

For pairs j, k in the same compound/context, their indicator covariance after inverse weighting is

$$C_{z,jk} = \mathrm{Cov}\left(\frac{M_{izj}}{\pi_{zj}}, \frac{M_{izk}}{\pi_{zk}}\right) = \left(\prod_{a \in j \cap k} \rho_{za}\right)^{-1} - 1, \tag{46}$$

where the empty product is one. Indeed, the joint inclusion probability is the product over $j \cup k$, and division by the two marginal pair probabilities leaves one inverse factor for each shared well. Different compound/context blocks have independent masks. If $b_{iz,j} = w_{zj}(Y_{izj} - \tilde{Y}_{izj})$, then

$$\mathrm{Var}_R(\widehat{T}) = N^{-2} \sum_{i,z} b_{iz}^{\top} C_z b_{iz}. \tag{47}$$

HT has the same formula with $\tilde{Y} = 0$. Therefore assistance improves variance if and only if the quadratic sum in (47) is smaller than the HT sum; poor predictors can increase variance. Both mean and variance identities are independently checked by enumerating all 16 subsets of a four-well block, with maximum error below 2×10^{-15}. The displayed exact variance uses the fully observed benchmark panel; in an actually incomplete future experiment it is not directly available. We do not report it as an implementable confidence interval for an unobserved biological target.

N.3 Representation-specific variance and error comparisons

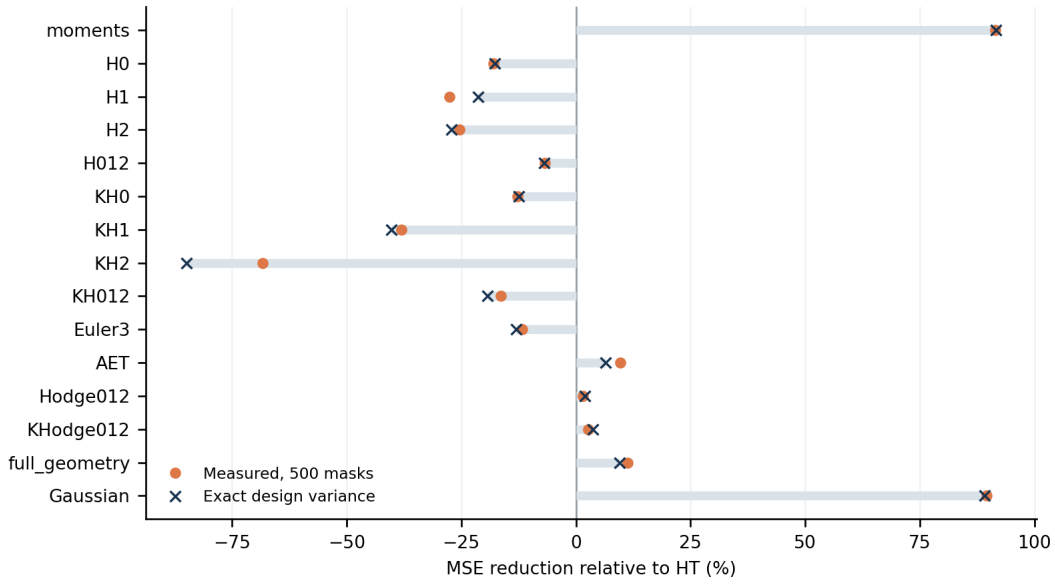


Figure 16: Relative MSE change from replicate assistance versus HT, within each representation's own target, under rare missingness and the hypothetical reference. Positive reduction is beneficial. Empirical reductions use all 500 masks and 16 grid points; exact variance reductions use the known full panel. Raw errors across different signatures do not measure a common target.

Representation	Pooled	HT	Hajek	Assisted
moments	0.02160	0.06771	0.01266	0.01803
H0	0.31221	0.31815	0.31790	0.34230
H1	0.08689	0.10225	0.08743	0.12088
H2	0.31528	0.32242	0.32171	0.36673
H012	0.15128	0.16482	0.15458	0.16744
KH0	0.55370	0.56720	0.55909	0.63516
KH1	0.06641	0.08097	0.06730	0.10224
KH2	0.20680	0.21031	0.20753	0.27625
KH012	0.19655	0.20832	0.19793	0.22795
Euler3	0.38603	0.39200	0.39363	0.40979
AET	0.51948	0.52770	0.51299	0.49402
Hodge012	0.37729	0.38687	0.37909	0.38154
KHodge012	0.06158	0.09158	0.06228	0.08676
full_geometry	0.15965	0.17682	0.16023	0.16129
Gaussian	0.00194	0.00383	0.00185	0.00119

Table 33: All representations: balanced missingness, balanced reference. RMSE is averaged over 500 masks and all 16 escort settings. Pooled and Hajek estimators need not be design-unbiased. An empty Hajek stratum receives zero; its frequency is recorded in the result file. Raw errors across representations address different targets.

Representation	Pooled	HT	Hajek	Assisted
moments	0.01707	0.05530	0.01025	0.01514
H0	0.35192	0.35703	0.35797	0.37788
H1	0.11576	0.12888	0.11649	0.15452
H2	0.45182	0.46123	0.45825	0.48991
H012	0.19514	0.20790	0.19801	0.20596
KH0	0.24344	0.25299	0.24391	0.36514
KH1	0.06409	0.07900	0.06447	0.09497
KH2	0.20878	0.21162	0.21088	0.28235
KH012	0.10848	0.11909	0.10891	0.15159
Euler3	0.46168	0.47234	0.47008	0.49050
AET	0.36177	0.37106	0.36344	0.37074
Hodge012	0.22310	0.22836	0.22211	0.22694
KHodge012	0.05934	0.08568	0.05976	0.08111
full_geometry	0.14188	0.15771	0.14332	0.14513
Gaussian	0.00163	0.00347	0.00158	0.00120

Table 34: All representations: balanced missingness, hypothetical rare reference. RMSE is averaged over 500 masks and all 16 escort settings. Pooled and Hajek estimators need not be design-unbiased. An empty Hajek stratum receives zero; its frequency is recorded in the result file. Raw errors across representations address different targets.

Representation	Pooled	HT	Hajek	Assisted
moments	0.18885	0.18311	0.08282	0.05564
H0	0.58497	0.86268	0.78522	0.96208
H1	0.16236	0.24356	0.22163	0.32511
H2	0.59835	0.83221	0.84190	0.93506
H012	0.25694	0.42871	0.40691	0.44680
KH0	0.85319	2.43101	2.28103	2.47847
KH1	0.11271	0.22383	0.18276	0.24973
KH2	0.26726	0.93359	0.78807	1.00997
KH012	0.32320	0.89499	0.82415	0.90973
Euler3	0.67182	1.03854	0.97195	1.08992
AET	0.72614	3.00522	2.88031	2.75113
Hodge012	0.62074	1.25867	1.26797	1.25100
KHodge012	0.13005	0.25493	0.19168	0.24641
full_geometry	0.30542	0.69665	0.64859	0.63582
Gaussian	0.00956	0.01128	0.00747	0.00356

Table 35: All representations: rare missingness, balanced reference. RMSE is averaged over 500 masks and all 16 escort settings. Pooled and Hajek estimators need not be design-unbiased. An empty Hajek stratum receives zero; its frequency is recorded in the result file. Raw errors across representations address different targets.

Representation	Pooled	HT	Hajek	Assisted
moments	0.11123	0.09706	0.03274	0.02838
H0	0.46777	0.78198	0.76283	0.84924
H1	0.15558	0.21989	0.20361	0.24836
H2	0.61234	0.63249	0.64827	0.70846
H012	0.24999	0.35683	0.34423	0.36860
KH0	0.37113	0.93459	0.91613	0.99240
KH1	0.09055	0.13742	0.11325	0.16145
KH2	0.24817	0.41655	0.39370	0.54052
KH012	0.15933	0.35669	0.34080	0.38478
Euler3	0.59371	1.04551	1.02076	1.10536
AET	0.46859	1.17616	1.12956	1.11812
Hodge012	0.32749	0.57309	0.56539	0.56849
KHodge012	0.09537	0.14780	0.11012	0.14581
full_geometry	0.20695	0.37275	0.34429	0.35120
Gaussian	0.00567	0.00653	0.00387	0.00211

Table 36: All representations: rare missingness, hypothetical rare reference. RMSE is averaged over 500 masks and all 16 escort settings. Pooled and Hajek estimators need not be design-unbiased. An empty Hajek stratum receives zero; its frequency is recorded in the result file. Raw errors across representations address different targets.

O SPECTRAL AND PREDICTABLE-DESCRIPTOR EXTENSIONS

Hard topology and differentiable spectra. Weighted Hodge operators provide an optional spectral refinement. With oriented boundary matrices B_k and positive diagonal weights W_k, set $D_k = W_{k-1}^{-1/2} B_k W_k^{1/2}$ and $L_k = D_k^\top D_k + D_{k+1} D_{k+1}^\top$. Their nullities are Betti numbers over $\mathbb{R}$; persistence in the code is over $\mathbb{F}_2$, a distinction that matters for torsion. On a fixed complex, normalized heat traces are smooth in positive weights and have explicit derivatives (Appendix O.1). Changing these weights does not change Betti numbers. Hard simplex membership is not globally differentiable. Nor are nonzero Hodge spectra homotopy invariants: the weak homotopy-based identification assumption for Betti numbers cannot simply be reused for them. Persistent pair Laplacians are a further established construction (Mémoli et al., 2022); the real study uses multiscale ordinary Hodge operators and does not label them persistent pair operators.

O.1 HODGE OPERATORS: WHAT IS AND IS NOT IDENTIFIED

Since $B_k B_{k+1} = 0$, weighted boundaries satisfy $D_k D_{k+1} = 0$. For any chain x, $x^\top L_k x = \|D_k x\|^2 + \|D_{k+1}^\top x\|^2 \geq 0$. The mutually orthogonal subspaces $\operatorname{im} D_k^\top$, $\ker L_k$, and $\operatorname{im} D_{k+1}$ span the chain space. Invertible positive weight changes preserve boundary ranks, so $\dim \ker L_k = n_k - \operatorname{rank} B_k - \operatorname{rank} B_{k+1} = \beta_k$. This is the usual weighted finite Hodge decomposition.

For a fixed dimension $n_k > 0$, let $H_t(L) = n_k^{-1} \operatorname{tr} e^{-tL}$. Differentiating the matrix exponential by its integral formula and using cyclicity of trace gives

$$\partial_\eta H_t(L) = -\frac{t}{n_k} \operatorname{tr}\{e^{-tL} \partial_\eta L\}. \tag{48}$$

This does not require differentiating eigenvectors, even at repeated eigenvalues. Integrating the same formula along the segment between two same-size positive semidefinite operators gives $|H_t(L) - H_t(L')| \leq t\|L - L'\|_{\mathrm{op}}$. The assumptions of aligned fixed chain coordinates and unchanged complex are essential. There is no comparable conclusion for arbitrary changes of active complex without additional analysis.

Two four-vertex Euclidean metrics set tree-edge squared distances to 1 and other off-diagonal squared distances to 1.2. A path and a star have identical complete H_0, H_1, H_2 Rips bars and Euler3 curves, but graph Laplacian spectra $\{0, 2-\sqrt{2}, 2, 2+\sqrt{2}\}$ and $\{0, 1, 1, 4\}$. Their heat traces differ. This proves that spectra may refine persistence in this example; it also proves they are not invariants of the corresponding homotopy information. Covariance, Gaussian geometry and anchored directional descriptors can distinguish these clouds too. The witness uses an explicit positive-diagonal Cholesky coordinate gauge, stored in `science_checks.json`, to make coordinate-dependent comparisons reproducible across platforms.

For a pair $K \subseteq L$, true persistent homology requires boundaries in L constrained to land in $C_k(K)$. If N is an orthonormal basis for the null space of the rows of B_{k+1}^L outside K, set $C = B_{k+1}^L[K, :]N$. Then $\Delta_k^{K,L} = (B_k^K)^\top B_k^K + CC^\top$ has nullity equal to $\operatorname{rank}(H_k(K) \to H_k(L))$ over $\mathbb{R}$ (Mémoli et al., 2022). This follows by restricting to $\ker B_k^K$ and removing the subspace of cycles that bound in L. Tests include a cycle filled by triangles and two vertices of K connected through a middle vertex present only in L. These checks accompany the implementation; no large-cloud or persistent-pair empirical benefit is claimed here.

O.2 PREDICTABLE HISTORY DESCRIPTORS

The optional history utility emits a signature of $Y_{\max(0,t-w):t-1}$ using a calibration fixed before time t. It is measurable with respect to $\mathcal{F}_{t-1}$ because it is a deterministic function of past data. A test changes all current and future observations and verifies unchanged output. This proves predictability of the construction, not causal sufficiency of the history, and no longitudinal empirical result is claimed.

P COMPUTATIONAL REPRODUCIBILITY

P.1 EXECUTION SCOPE AND AUDIT RECORDS

The CPU workflows, fixed seeds, numerical tolerances, protocol files, and recorded outputs accompany the manuscript. Source hashes distinguish the evaluated implementations from analysis and document-generation scripts. Local protocol records document the computational workflow and the order of analysis decisions. The ACIC regularization sensitivity analysis and the paired-discordance uncertainty analysis are explicitly identified in their experimental protocols.

Recorded source hashes cover the inference implementation, data generator, mathematical check program, protocol, and pre-evaluation theorem note. All 123,300 certification and oscillator dataset seeds are unique and disjoint from the 4,000 matched-comparison and scalar-Wald simulation seeds. Every stored count table sums to its declared n. Six selected transition records reproduce their counts and bands exactly; all nine first-seed oscillator conditions reproduce their permutation p-values exactly. The source hashes agree with the evaluated implementations. The independent radius linear-program comparison has maximum error 1.01×10^{-12}, and the complete joint-law versus rare-cell KL factorization differs by less than 10^{-18}.

The factorial simulation uses aggregated counts because the categorical experiment depends on the sample only through those sufficient statistics. It is exactly equivalent to drawing and storing each iid unit; it does not approximate large datasets by a smaller effective sample. The measured core runtime was 30.2 seconds, excluding dependency installation, checks, summarization, and PDF production. This speed reflects the finite state space and does not demonstrate scalability to arbitrary large complexes or deep representation learning.

P.2 NUMERICAL IMPLEMENTATION AND SOFTWARE CHECKS

The finite-confounder and finite-panel workflow uses the included derived inputs and pins numerical-library threads to two. Optional CUDA reproduction checks matrix-exponential differentiation, eigensolvers, and Gaussian parity using an installed GPU-enabled PyTorch environment, and does not silently fall back to CPU. Combinatorics, confidence sets, and finite-panel sampling remain CPU computations.

Recorded hardware checks report NVIDIA GB10, ARM64, capability 12.1, PyTorch 2.12.1+cu130, and executed CUDA mathematics. They validate the spectral/Gaussian workload on that configuration, not every experimental launcher. The main experimental runs reported here were performed on an x86_64 CPU.

The real-panel computation caches repeated NPZ decompression without changing numerical operations. CSV validation preserves the literal scenario label "null". Radius checks exclude empty strict active-pattern regions at $Mc = 1$ when claiming an exact radius. This has no effect on the declared $c = 0.1$, $M \in \{3, 5, 7\}$ experiments. Radius reporting uses 10^{-8} numerical tolerance rather than equating bisection precision with feasibility precision. Implementation diagnostics, including unsuccessful checks, are archived separately from the scientific result tables.