CERTIFYING INTERVENTIONAL AGREEMENT AMONG OBSERVATIONALLY EQUIVALENT CAUSAL MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Observationally equivalent causal models can still disagree about what happens under intervention, because interventions create inputs that never occur in observational data. We introduce Interventional Separation Selection (ISS), which repeatedly queries the true system with an admissible intervention on which the surviving candidate models disagree, discards the candidates the outcome contradicts, and stops once no intervention within a cost bound separates the survivors. If the true system is among the candidates, this stopping condition certifies that every survivor agrees with it on every admissible intervention within the bound, a guarantee that no observational learner can give, however much data it sees. The stopping condition depends only on the survivors, so it can be checked without knowing the truth. For continuous variables the candidates form an infinite version space, and mixed-integer linear programs decide the stopping condition exactly over all of it, with agreement holding up to a tolerance. On a three-digit colored MNIST causal abstraction task in which ink hue tracks digit size, plain convolutional networks trained on $60{,}000$ examples reach zero held-out error, yet disagree with shape-based labels on 26% of single-digit edits, as often as hue-based labels do. Auditing the causal abstractions of networks observed only on such images, ISS certifies what each network perceives with 13.6 interventions per image on average, and each certificate, checked against every admissible intervention, holds whenever the network's true abstraction is among the candidates. When a network bypasses a unit that every candidate abstraction relies on, certificates covering interventions on that unit can be silently void, and twenty random validation interventions refute 69% of them.

1 INTRODUCTION

Machine learning models are typically evaluated by how well they generalize to unseen data. However, two models can produce the same prediction while arriving at it for different causal reasons. This problem arises naturally under *underspecification* (D'Amour et al., 2022): multiple models can fit the same observational data—and even agree on held-out observations—while relying on different underlying mechanisms and making different predictions under intervention. For instance, one model may reach the correct prediction because it captures the intended causal relationship, while another may reach the same prediction through a shortcut that happens to work on the observed data. Collecting more data from the same observational regime does not resolve this ambiguity when the relevant configurations are never observed (Theorem 1).

We refer to a model together with the context explaining an observed outcome as a *causal perspective*. Different causal perspectives can therefore agree on *what* happened while encoding different explanations for *why* it happened. These differences become visible only under intervention.

This motivates a different question:

> *Can the interventions we run certify a model's behavior under those we do not?*

We operationalize this question through intervention: rather than requiring identification of a unique perspective, we ask whether all perspectives still consistent with the evidence agree on every admissible intervention up to a specified cost. We introduce *Interventional Separation Selection (ISS)*

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

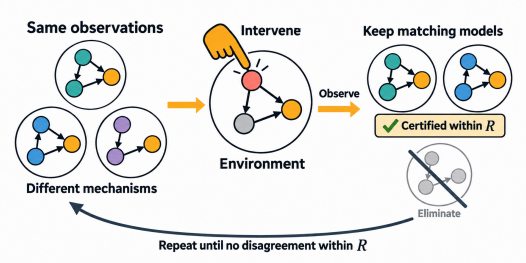

Figure 1: Observationally equivalent models can encode different mechanisms. ISS selects a separating intervention (minimum cost by default), observes the environment's answer, and eliminates inconsistent hypotheses. The loop certifies agreement only when no query within radius R separates any surviving pair in the full version space; with exact answers and the truth in the initial space, every survivor then has zero interventional risk within R.

to provide such a certificate. ISS maintains the set of observationally compatible perspectives and actively queries interventions on which surviving perspectives disagree. After observing the environment's response, inconsistent perspectives are eliminated, as displayed in Figure 1. The procedure stops only when no admissible intervention of cost at most R can separate any surviving pair. If the true system is among the initial perspectives, this stopping condition certifies that every surviving perspective agrees with the true system on all admissible interventions within radius R.

Contributions. Our main contribution is a framework for *certifying causal reasoning through interventional generalization under observational underspecification*. We introduce ISS, which actively separates observationally compatible causal perspectives and provides an exact, cost-bounded certificate of interventional agreement, and we extend the certificate from Boolean to continuous variables (Proposition 4). We apply ISS to the causal abstractions of MNIST classifiers whose ink hue, internal percepts and output are continuous (Section 6). Networks that read a digit's shape and networks that read its ink hue cannot be told apart on the observed images; ISS certifies each network with 13.6 image edits per image on average, and we check every certificate against every admissible intervention. In a Boolean version with no tolerance (Appendix D), ISS never needs more than 6 image edits to certify radius 2, the fewest any learner can guarantee in the worst case (Theorem 3).

2 Related Work

Underspecification and multiplicity. Predictively equivalent models can rely on different mechanisms and therefore behave differently outside the regime on which they were evaluated (D'Amour et al., 2022; Marx et al., 2020). Shortcut learning provides a prominent instance of this phenomenon under distribution shift (Geirhos et al., 2020; Sagawa et al., 2020). Prior work has used interventions or counterfactual augmentation to expose such differences (De Haan et al., 2019; Kaushik et al., 2020). Our focus is not merely to detect multiplicity, but to characterize when the surviving models are guaranteed to agree under every admissible intervention within a specified budget.

Experimental design and active causal learning. Model-discrimination design chooses experiments that separate competing explanations (Box & Hill, 1967; Chaloner & Verdinelli, 1995), while active causal discovery chooses interventions to resolve uncertainty about causal structure (Eberhardt & Scheines, 2007; Hauser & Bühlmann, 2014; Squires et al., 2020); the state of the art amortizes the design policy (Tigas et al., 2022; Annadani et al., 2024), prices the measuring oracle rather than the intervention (Zhang et al., 2023), or plans under an explicit budget (Guo et al., 2026), and cost-aware and verification-oriented variants ask how interventions should be allocated under experimental constraints (Ghassami et al., 2018; Choo et al., 2022). ISS differs in its target: given a version space of fully specified causal perspectives at a fixed context, it seeks a cost-bounded certificate of interventional equivalence rather than recovery of a unique graph, and its budget indexes that guarantee rather than only constraining the design. Those methods stop when a horizon or budget is spent rather than when nothing remains to separate; an information criterion can express that test, but only if its expectation is exact over the whole version space, which an amortized or sampled

posterior does not give. We therefore port their acquisition principle, expected information per unit cost, into our loop (Appendix D) and discuss the differences in Appendix E.5.

Disagreement-based learning and testing. Version-space methods and disagreement-based active learning select queries that distinguish surviving hypotheses (Mitchell, 1982; Seung et al., 1992; Tong & Koller, 2001); differential testing similarly searches for inputs on which implementations disagree (McKeeman, 1998; Pei et al., 2017). ISS gives this idea a causal and cost-sensitive form: queries are interventions, disagreement is defined by counterfactual responses, and exhaustion of all separating queries within a radius yields a certificate. This objective also differs from algorithmic recourse, where the goal is to find a low-cost action that changes a particular prediction (Ustun et al., 2019; Karimi et al., 2021); here, cost measures how difficult it is to expose a disagreement between hypotheses.

Causal abstraction and counterfactual reasoning. Causal abstraction evaluates whether interventions on a lower-level system realize the interventions of a higher-level causal model (Beckers & Halpern, 2019; Geiger et al., 2021; 2025). We use candidate abstractions and alignments as the hypothesis class, so that observationally plausible abstractions can be actively separated and, under realizability, certified over a specified intervention family. Counterfactual inference instead typically assumes a causal model and reasons about outcomes under hypothetical interventions (Balke & Pearl, 1994; Oberst & Sontag, 2019); our problem is deciding among multiple such models by choosing interventions. Identification theory asks whether causal quantities are determined from available distributions (Pearl, 1995; Shpitser & Pearl, 2008); we ask how much intervention cost is required before the remaining candidate models become indistinguishable for the interventions of interest. Further connections, including query learning, grounding, the computational complexity of causal reasoning, and additional causal-discovery work, are discussed in Appendix E.

3 Interventional Risk and the Observational Floor

Causal perspectives. Let $[n] := \{1, \dots, n\}$ and $X = (X_1, \dots, X_n)$, $X_j \in \{0, 1\}$. An SCM $M = (U, X, F)$ has exogenous variables $U = (U_1, \dots, U_n)$ with finite domains $\mathcal{U}_j$ and assignments $X_j = f_j(X_{\mathrm{Pa}_M(j)}, U_j)$, where $\mathrm{Pa}_M(j) \subseteq [n] \setminus \{j\}$. By acyclicity every context $u \in \mathcal{U} := \prod_j \mathcal{U}_j$ determines a unique solution $x(M, u)$. A *causal perspective* is a pair $P = (M, u)$ with observational state $x(P) := x(M, u)$; two perspectives are *history-equivalent at* x^{obs} when $x(P) = x(P') = x^{\mathrm{obs}}$.

Queries and cost. An intervention $I = \mathrm{do}(X_S = a_S)$ replaces the assignments of X_S by constants, giving M^I; a query $q = (I, T)$ has target $T \subseteq [n]$ and answer $A_P(q) := x(M^I, u)_T$. The admissible family $\mathcal{Q}$ encodes which variables may be manipulated, how many at once and which targets are observable ($\mathcal{Q}_{\mathrm{full}}$: all). Given x^{obs},

$$c_{\mathrm{int}}(I; x^{\mathrm{obs}}) = \sum_{s \in S} w_s \, \mathbb{1}\left[a_s \neq x_s^{\mathrm{obs}}\right], \qquad w_s > 0, \tag{1}$$

prices departure from the observed world; $c(q) := c_{\mathrm{int}}(I_q; x^{\mathrm{obs}})$, $w_{\min} := \min_s w_s$, and $\mathcal{Q}_r := \{q \in \mathcal{Q} : 0 < c(q) \leq r\}$. r is the maximum intervention cost allowed when deciding which queries count as admissible.

Definition 1 (Separation cost and local equivalence)**.** The *separating query set* is $\mathcal{W}_{PP'} := \{q \in \mathcal{Q} : A_P(q) \neq A_{P'}(q)\}$ and the *interventional separation cost* is $k_{\mathcal{Q},c}(P, P') := \min_{q \in \mathcal{W}_{PP'}} c(q)$, with $\min \emptyset := \infty$; $P \equiv^r_{\mathcal{Q},c} P'$ (*r-local equivalence*) when $A_P(q) = A_{P'}(q)$ for every $q \in \mathcal{Q}_r$.

The learning problem. An environment is a true perspective $P^\star = (M^\star, u^\star)$. A learner receives N observational samples S_N from a regime over contexts and the observed state $x^{\mathrm{obs}} = x(P^\star)$, may run queries in $\mathcal{Q}$ and observe $A_{P^\star}$, and outputs $\hat{P}$ from a class $\mathcal{H}$ of perspectives history-equivalent at x^{obs}.

Definition 2 (Interventional risk)**.** For r with $\mathcal{Q}_r \neq \emptyset$ and π_r uniform on $\mathcal{Q}_r$, $R_r(P, P') := \Pr_{q \sim \pi_r}[A_P(q) \neq A_{P'}(q)]$ and $R_r(P) := R_r(P, P^\star)$.

R_r is a normalized Hamming distance between answer vectors, hence a pseudo-metric.

Proposition 1 (Separation cost is the margin)**.** *For all P, P' history-equivalent at x^{obs} and all r with $\mathcal{Q}_r \neq \emptyset$: $R_r(P, P') = 0$ if $r < k_{\mathcal{Q},c}(P, P')$ and $R_r(P, P') > 0$ if $r \geq k_{\mathcal{Q},c}(P, P')$; hence $P \equiv^r_{\mathcal{Q},c} P'$ iff $R_r(P, P') = 0$ iff $k_{\mathcal{Q},c}(P, P') > r$.*

Proofs are in Appendix A. Separation cost is symmetric, monotone in $\mathcal{Q}$ and c, at least $w_{\min}$ for history-equivalent perspectives (no free query is informative), and satisfies the ultrametric inequality $k_{\mathcal{Q},c}(P, P'') \geq \min\{k_{\mathcal{Q},c}(P, P'), k_{\mathcal{Q},c}(P', P'')\}$, so $\equiv^r_{\mathcal{Q},c}$ is an equivalence relation whose classes refine as r grows (Proposition 6, Appendix B). This makes identification well posed: *identifying the truth's class is the most any learner can achieve at budget r—it cannot distinguish within a class—and by Proposition 1 it is enough.*

The observational version space. Write each mechanism as $f_j(x_{\mathrm{Pa}(j)}, u_j) = g_j(x_{\mathrm{Pa}(j)}) \oplus u_j$, $u_j \in \{0, 1\}$, and call (j, z), $z \in \{0, 1\}^{|\mathrm{Pa}(j)|}$, a *mechanism cell*. A learner that knows the parent sets and the noise form must estimate $g_j(z)$ for every cell.

Assumption 1 (Observational regime)**.** Contexts are independent and identically distributed (i.i.d.); roots satisfy $X_j = U_j$ and non-root noise bits are 1 with probability $p < 1/2$.

Proposition 2 (Zero-training-error version space)**.** *Under Assumption 1 with $p = 0$, let $\mathrm{sup}(S_N)$ be the set of cells that appeared in the dataset S_N. The set $V_0(S_N)$ of hypotheses with zero training error is $\{g : g_j(z) = g^\star_j(z)\ \forall (j, z) \in \mathrm{sup}(S_N)\}$, a product over the unobserved cells, and every member has the same likelihood. The m_N unobserved cells split into $m_N - m_\infty$ cells the regime reaches with positive probability but have not yet occurred in the dataset, and m_∞ cells it reaches with probability zero; $|V_0(S_N)| = 2^{m_N}$, $m_N \downarrow m_\infty$ almost surely, and only the completions of the m_∞ permanent cells are indistinguishable on held-out observational data. For $p > 0$ the maximum-likelihood fit is the majority value on each observed cell and unobserved cells remain free.*

By Proposition 5 every member of V_0 abduces a context at x^{obs}: identical observational, different interventional behavior.

Theorem 1 (Observational floor)**.** *Let L be any (possibly randomized) learner with $\hat{P} = L(S_N)$. For every r with $\mathcal{Q}_r \neq \emptyset$,*

$$\sup_{P^\star \in V_0(S_N)} \mathbb{E}\, R_r(L(S_N); P^\star) \;\geq\; \tfrac{1}{2} \max_{P, P' \in V_0(S_N)} R_r(P, P'). \tag{2}$$

The right side depends on S_N only through $\mathrm{sup}(S_N)$, so it does not vanish as $N \to \infty$, and it is positive whenever two completions of the m_∞ permanently unobserved cells are separated by some $q \in \mathcal{Q}_r$.

The bound is about the problem, not an estimator: it holds for every observational learner at every N. Held-out data can still realize a missing finite-sample cell; for the completions of the m_∞ permanent cells no held-out evaluation detects the residual risk, since they behave identically on observations.

4 The Certificate, and Making It Exact

Definition 3 (Version-space separation cost)**.** For a set V of perspectives, $\kappa(V) := \min_{P \neq P' \in V} k_{\mathcal{Q},c}(P, P')$ (∞ when all members are interventionally equivalent, in particular when $|V| \leq 1$).

Theorem 2 (Soundness and certified zero risk)**.** *Assume $P^\star \in V_0$ and let Algorithm 1 run with any rule that queries a separating query. Then (a) $P^\star \in V$ at every iteration; (b) $|V|$ strictly decreases at every iteration, so it halts after at most $|V_0| - 1$ interventions; (c) if it halts with $\kappa(V) > R$ then every $\hat{P} \in V$ satisfies $\hat{P} \equiv^R_{\mathcal{Q},c} P^\star$ and therefore $R_r(\hat{P}) = 0$ for every $r \leq R$, and if it halts on the budget the same holds with R replaced by $\kappa(V) - w_{\min}$; (d) the halting condition is a function of V alone, so (c) is a certificate verifiable without knowing $P^\star$.*

The certificate belongs to the stopping rule, not to the query rule: any rule that never wastes a query on a non-separating intervention inherits (a)–(d), the loop guard leaves one affordable (it costs $\kappa(V) \leq B - s$), and by (a) an empty V refutes $P^\star \in V_0$. In our experiments (Section 6 and Appendix D), every rule certifies correctly wherever the truth lies in the class, including the information-per-cost rule of Bayesian design and a rule that probes at random; the rules differ only

Algorithm 1 Interventional separation selection (ISS)

Require: data S_N, state x^{obs}, class $\mathcal{H}$, family $\mathcal{Q}$, cost c, radius R, budget B, evidence $(E, u_E^\star)$, level δ

1: $V \leftarrow$ hypotheses in $\mathcal{H}$ within likelihood level δ of the fit on S_N, history-equivalent at x^{obs}, with $u_{P,E} = u_E^\star$; spent cost $s \leftarrow 0$ ▷ δ: Theorem 4; δ=0: zero training error
2: **while** $\kappa(V) \le R$ and $s + \kappa(V) \le B$ **do**
3: $q \leftarrow$ a query in $\mathcal{Q}_R$ with $c(q) \le B - s$ separating two members of V ▷ default: $c(q) = \kappa(V)$
4: run q, $s \leftarrow s + c(q)$, observe $y \leftarrow A_{P^\star}(q)$, set $V \leftarrow \{P \in V : A_P(q) = y\}$
5: **end while**
6: **if** $V = \emptyset$ **return** s and "refuted" (no hypothesis fits, so $P^\star \notin V_0$)
7: **return** $\hat{P} \in V$, s, and the certified radius: R if $\kappa(V) > R$, else $\kappa(V) - w_{\min}$

in cost. Part (c) is deterministic, not a high-probability bound, whereas an observational learner cannot certify any positive radius (Theorem 1). What stays uncertified is the terminal class $\{P : A_P = A_{P^\star} \text{ on } \mathcal{Q}_R\}$: membership follows if the induced mechanisms $\varphi_j^P(x) := f_j(x_{\mathrm{Pa}_M(j)}, u_j)$ agree with $P^\star$'s on every state some query in $\mathcal{Q}_R$ realizes (Corollary 1, Appendix B), but not only then, since a difference on a realized cell survives when every query of cost at most R masks it at the target; both kinds of survivor carry risk beyond the radius.

How much does certification cost? The minimum-cost rule pays $\kappa(V_t)$ at step t, non-decreasing along the run since $V_{t+1} \subseteq V_t$, which is why it is the default. The lower bounds hold for *every* learner, adaptive or not; note their quantifiers.

Theorem 3 (Cost of certification)**.** *Assume $P^\star \in V_0$, let T be the number of* ISS *interventions and $C := |V_0/\equiv^R|$ the number of R-local classes. Then*

(a) (Upper bound, minimum-cost rule) the total cost is $\sum_{t<T} \kappa(V_t) \le T \cdot \kappa(V_{T-1}) \le (|V_0|-1)R$.
(b) (Lower bounds) Any learner that halts and correctly certifies radius R in environment $P^\star$ spends at least $k^\star(P^\star) := \max\{k_{\mathcal{Q},c}(P^\star, P') : P' \in V_0,\ k_{\mathcal{Q},c}(P^\star, P') \le R\}$, hence in the worst case over $P^\star \in V_0$ at least $\max\{k_{\mathcal{Q},c}(P, P') : P, P' \in V_0, k_{\mathcal{Q},c}(P, P') \le R\}$; with a binary target its answer strings over the C classes form a prefix-free code, so in the worst case over $P^\star$ it runs at least $\lceil \log_2 C \rceil$ interventions.
(c) (Greedy approximation, non-adaptive) Let OPT *be the minimum total cost of a set of queries in $\mathcal{Q}_R$ separating every pair in V_0 of separation cost at most R. Greedily adding the query maximizing newly separated pairs per unit cost yields such a set of cost at most $(1 + \ln \binom{|V_0|}{2})$OPT, and running it certifies radius R.*

Part (b) bounds the cost of *any* learner in a given environment by the hardest hypothesis it must refute; part (c) bounds *non-adaptive* certification, which adaptive learning can undercut (Appendix D).

The acquisition step is computationally hard. Deciding whether a separating query of cost at most r exists is NP-complete, even for Boolean SCMs with in-degree 2, unit intervention costs and a single target. Bounded intervention size makes it tractable, since all candidates can then be enumerated in polynomial time; so when every variable is manipulable, deciding *whether* two hypotheses can be separated is easy, while finding the *cheapest* separating intervention can still be hard. Algorithm 1 handles this exactly by cost-ordered enumeration.

The full version space, implicitly. The version space has $|V_0| = 2^{m_N}$ members, exponentially many in the number of unobserved cells, and a certificate over a *sample* of it is not a certificate. Under Proposition 2 the answer of P to q is a Boolean function f_q of the $m := m_N$ unobserved cells, and propagating q while branching on every unobserved cell met on the way writes f_q as a DNF whose cubes are the branches. An answer y to q thus adds the constraint $f_q = y$, which can be written exactly as linear inequalities over binary variables, with one indicator variable per cube that answers y. After t answers the version space is the set of completions consistent with all t constraints, and each new answer appends its inequalities to the existing system.

Proposition 3 (Exact acquisition over the implicit version space)**.** *(i) f_q has at most $2^{n'}$ cubes, n' the number of non-root variables, one per trajectory of the propagation. (ii) A query separates two members of V_t iff two of its cubes with different answers are each consistent with the constraints—one feasibility check per cube, a 0-1 integer program with the cube's literals fixed. (iii) Scanning*

$\mathcal{Q}_R$ in cost order therefore yields $\kappa(V_t)$, a minimum-cost separating query and the certificate of Theorem 2(c), exact over the full V_0, in $O(|\mathcal{Q}_R| \max_q |f_q|)$ feasibility checks per step. (iv) A uniform sample of V_0 conditioned on the answers of fixed *queries is a uniform sample of V_t, but in general not when the queries were chosen as a function of the sample; so the sample serves only acquisition (tie-breaking, split scores), and every risk estimate is computed on a fresh, independent uniform sample of V_t drawn by enumeration or rejection.*

Noisy mechanisms: a soft certificate. When $p > 0$ the maximum-likelihood fit can be wrong on an observed cell, $P^\star \notin V_0$, and Theorem 2 is void. The remedy is to keep every hypothesis whose likelihood is within a calibrated margin of the fit. For a fitted cell z of variable j with n_z samples of which k_z are ones, let $d_z := |2k_z - n_z|$; flipping the fitted value of z costs $d_z\,\ell$ in log-likelihood, $\ell := \log \frac{1-p}{p}$, so the truth's deficit on variable j is $D_j := \ell \sum_{z \in j:\ \mathrm{fit}_z \neq g_j^\star(z)} d_z$.

Theorem 4 (Soft certificate)**.** *Under Assumption 1 with known $p \in (0, \frac{1}{2})$, conditional on the cell counts $(n_z)_{z\in j}$ of variable j, D_j/ℓ is distributed as $\sum_{z\in j} \max(n_z - 2K_z, 0)$ with independent $K_z \sim \mathrm{Bin}(n_z, 1-p)$, whatever the noise at other variables. Let T_j be the smallest integer with $\Pr[D_j/\ell > T_j \mid (n_z)_{z\in j}] \leq \delta/n'$, with n' the number of non-root variables, and let V_0^δ be the set of hypotheses that agree with the fit except on unobserved cells and on fitted cells of each variable j of total margin at most T_j. Then $\Pr[P^\star \in V_0^\delta] \geq 1 - \delta$, and* ISS *run on V_0^δ certifies correctly—every survivor has zero risk within the certified radius—with probability at least $1 - \delta$.*

The quantile is a convolution of capped binomials computed from the observed counts, the soft version space is again a product structure amenable to Proposition 3 (flip variables with an allowed-set constraint), and the price of the guarantee is the budget needed to eliminate the flipped hypotheses.

Continuous variables. Nothing in Algorithm 1 needs finite domains. Let variables take real values, let c be any departure from x^{obs} (for example $\sum_s w_s |a_s - x_s^{\mathrm{obs}}|$), and let answers be real vectors. For a tolerance $\varepsilon > 0$, a query q *ε-separates* P and P' when $\|A_P(q) - A_{P'}(q)\|_\infty > \varepsilon$, and $\kappa^\varepsilon(V)$ is the least cost of such a query over pairs in V. Run Algorithm 1 with ε-separation, and after an answer y keep the hypotheses whose answer lies within η of y, where η bounds the error of the observed answers.

Proposition 4 (Continuous certificate)**.** *If $P^\star \in V_0$ and every observed answer lies within η of $P^\star$'s, then $P^\star$ is never eliminated, and if the loop halts with $\kappa^\varepsilon(V) > R$, every survivor is within ε of $P^\star$ on every $q \in \mathcal{Q}_R$. If $2\eta < \varepsilon$ and $\mathcal{Q}_R$ is finite, the loop halts after at most $|\mathcal{Q}_R|$ interventions.*

Theorem 1 carries over with the risk $\mathbb{E}_{q\sim\pi_r} \|A_P(q) - A_{P'}(q)\|_\infty$, again a pseudo-metric. When answers are linear in a hypothesis's parameters, up to decisions on the sign of a linear function, the version space is a finite union of polytopes and ε-separation is decided exactly by mixed-integer linear programs, which extend the 0-1 programs of Proposition 3 to real-valued parameters (Section 6).

Misspecification and validation. Every certificate above is conditional on the truth lying in the maintained class. When it does not—the fit is wrong beyond the soft margin, or no candidate abstraction of a network is faithful—the loop still halts once the survivors agree, and its certificate is void *silently*. A *validation phase* compares the survivors' shared predictions on unexecuted queries with the environment on a separate budget of random admissible queries and refutes the class at the first contradiction; it certifies nothing, but it alone can falsify the class.

5 Exogenous Grounding: Supervision at Zero Intervention Cost

A hypothesis can fit the observed state by *choosing* its context—attributing an observed fact to an exogenous surprise rather than to its mechanism.

Proposition 5 (Abductive freedom)**.** *If $f_j(x_{\mathrm{Pa}(j)}, u_j) = g_j(x_{\mathrm{Pa}(j)}) \oplus u_j$ for every j, then for every structure g and every x^{obs} exactly one context, $u_j = x_j^{\mathrm{obs}} \oplus g_j(x_{\mathrm{Pa}(j)}^{\mathrm{obs}})$, makes (M_g, u) history-equivalent at x^{obs}; history-equivalence constrains structure not at all.*

An *exogenous evidence set* $E \subseteq [n]$ with verified values $u_E^\star$ (an audit confirming that a component operated nominally) defines the grounded version space $V_0^E := \{P \in V_0 : u_{P,E} = u_E^\star\}$, whose complement is eliminated at cost 0; $E \subseteq E'$ implies $V_0^{E'} \subseteq V_0^E$ and $\kappa(V_0^{E'}) \geq \kappa(V_0^E)$ (Proposition 7). The two are complementary: evidence is blind to structural disagreement, interventions

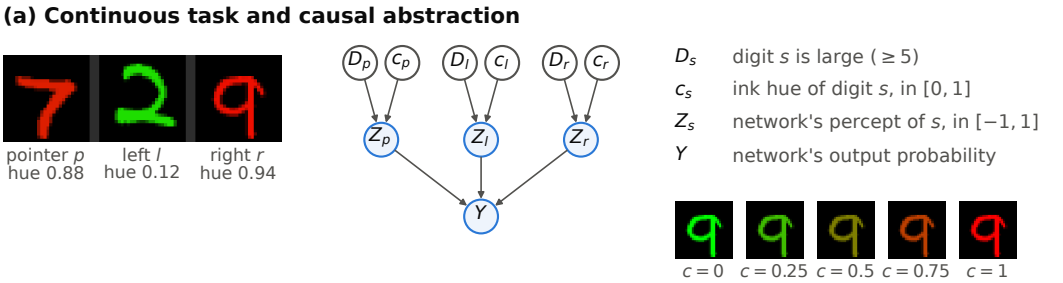

Figure 2: ISS with continuous variables. (a) An observed image, the high-level causal model and a digit drawn at five hues. (b) ISS on this image for two networks, both perfect on observed images. Each step shows the intervention ISS chose (edited digits framed; a hue edit sets the digit's hue to the value shown) and the network's output probability y. Both certificates hold within ε on every admissible intervention of cost at most 0.5.

see context disagreement only where it is unmasked, and for every r there are structurally identical hypotheses with separation cost r under the full family and ∞ below intervention size r that one verified fact tells apart (Theorems 7 and 8, Appendix B): *data constrains mechanisms on the visited configurations, evidence constrains contexts, interventions constrain mechanisms off them.*

6 Experiments: Continuous Causal Abstraction on MNIST

We test our theory on vision networks with continuous causal variables: ink hue, internal percepts, and output probability. Observational data cannot reveal what the network has learned, but the certificates can still be checked exactly. A Boolean version with red/green ink appears in Appendix D, and the full continuous protocol in Appendix C.

Task. Each image contains three MNIST digits (LeCun et al., 1998): a pointer p, a left digit l, and a right digit r (Figure 2a). Each digit has a hue $c \in [0, 1]$, from green ($c = 0$) to red ($c = 1$). A digit is *large* if it is at least 5 and *small* otherwise; we call this its *size*. A small pointer selects the left digit, while a large pointer selects the right. The label is 1 if the selected digit is large. In Figure 2a, the pointer 7 selects the 9, so the label is 1. This is a version of MNIST pointer-value retrieval (Zhang et al., 2021), a benchmark for causal abstraction (Geiger et al., 2022).

Causal abstraction. We describe the network with a small causal model aligned with its internal components (Geiger et al., 2021). For each digit $s \in \{p, l, r\}$, the model has its size D_s, hue c_s, and a percept $Z_s = g_s(D_s, c_s) \in [-1, 1]$. The output is $Y = \sigma(\gamma Z_{\text{sel}} + \beta)$, where $Z_{\text{sel}} = Z_r$ if $Z_p > 0$ and Z_l otherwise.

This alignment is exact in our networks: each digit is processed by its own CNN ending in a single $\tanh$ unit representing Z_s, and a router passes the selected percept to the output (Koh et al., 2020). The unknown part is the percept function g_s. It may depend on digit shape, hue, or a mixture of both.

Why observation cannot decide. The auditor only sees images where hue follows digit size: large digits are close to red and small digits are close to green, as in a continuous Colored MNIST shortcut

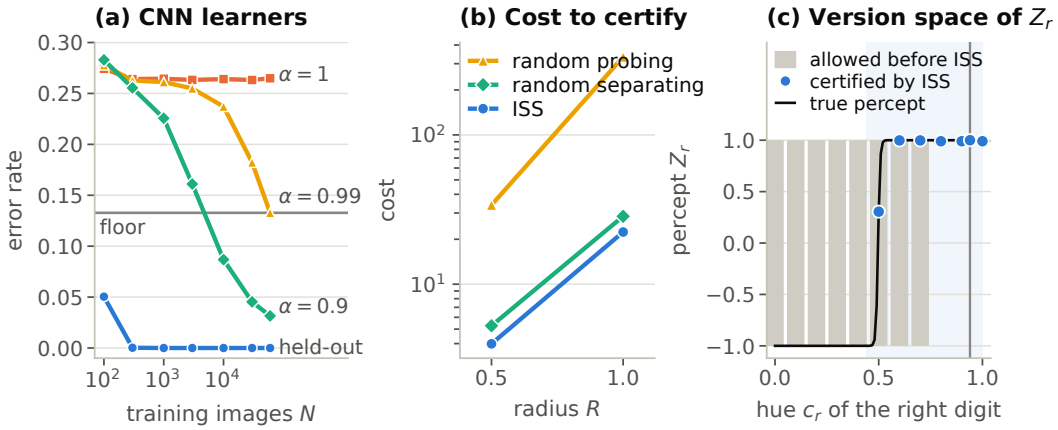

Figure 3: (a) Plain CNNs trained on N observed images: held-out error, and error on single-digit edits against the shape world; gray: the floor of Theorem 1 for the $\alpha = 1$ regime, whose missing cells the $\alpha < 1$ learners eventually observe. (b) Cost of certifying radius R with visual interventions, networks in the class. (c) The right digit's percept in the hue reader of Figure 2: what the log allows at each admissible hue, and what ISS certified within $R = 0.5$ of the observed hue (shaded).

(Arjovsky et al., 2019). From 3,000 such images, it can estimate γ, β, and constrain the percept curves on observed hues. Outside these hues, the curves remain unconstrained, so the version space V_0 is infinite. Each hypothesis specifies 72 percept values, of which 69 remain free or only bounded by the log.

Interventions and certificates. A *visual* intervention recolors a digit, with cost equal to the hue change, or swaps it with a digit of the other size, at cost 1. A *white-box* intervention also allows setting the pointer percept to one of $-1, -0.5, 0, 0.5, 1$ (Geiger et al., 2021). The network's output probability is the answer.

With tolerance $\varepsilon = 0.05$, ISS stops when no intervention of cost at most R can make two surviving hypotheses disagree by more than ε. If the truth lies in V_0, Proposition 4 guarantees that every survivor is within 0.05 of the network on all such interventions. We compute separation exactly with small mixed-integer linear programs and verify each certificate by testing every admissible intervention.

Networks. We audit 18 networks on 10 observed images each:

- *trained* (9): hue matches size with probability $\alpha \in \{1, 0.9, 0.5\}$, with three seeds per setting;
- *planted* (6): each percept uses a known random mix of shape and hue;
- *bypass* (3): the router uses a separate pointer encoder, so none of our candidate abstractions matches the network.

Result 1: observation cannot reveal the abstraction. Held-out accuracy on observed images cannot tell what a network reads. Every network in the class scores at least 98.6%, and the hue readers ($\alpha = 1$) score highest, 100.0%. Yet the intended abstraction mispredicts 32% of their interventions of cost at most 1 by more than ε (Table 1). By Theorem 1, in its continuous form, whatever an observational auditor reports is off by at least 0.16 in expected output probability for some network consistent with its log, averaged over observed images, however large the log.

The same limit binds ordinary learners. Call the intended labeling rule the *shape world*, and the same rule applied to hues, with $c \geq 0.5$ counting as large, the *hue world*; the two label every observed image identically. Plain CNNs trained on up to 60,000 observed images reach zero held-out error, yet disagree with the shape world on 26% of single-digit edits (Figure 3a): they learned the hue world. When one digit in ten has a random hue ($\alpha = 0.9$), mismatched hues occur in the data, and the error falls with N to 3%.

Result 2: certificates are correct and cheap. Figure 2b shows ISS on one image for two networks, both perfect on observed images. At each step ISS runs the cheapest intervention on which survivors

Table 1: Auditing 18 networks at $R = 0.5$ (visual interventions unless marked; means over 10 observed images each). *In class*: images at which the true percepts lie in V_0. *Intended risk*: admissible interventions on which the intended abstraction is off by more than ε. ISS: interventions and total cost; *Random*: total cost. *Correct*: certificates within ε of the network on every admissible intervention, as all are where the truth lies in V_0.

	Held-out	In	Intended	ISS		Random	Random	
Networks (n)	accuracy	class	risk	int.	cost	sep.	probing	Correct
Trained, $\alpha = 1$ (3)	100.0%	100%	6%	7.6	2.5	3.3	34.4	100%
Trained, $\alpha = 0.9$ (3)	99.5%	93%	2%	13.4	4.0	5.3	43.1	100%
Trained, $\alpha = 0.5$ (3)	98.7%	97%	0%	21.1	6.1	8.3	38.0	100%
Planted (6)	99.2%	95%	3%	13.0	3.7	4.7	27.1	98%
Bypass (3)	100.0%	–	5%	7.9	2.7	3.3	35.4	100%
white-box			4%	7.9	2.7	3.4	52.6	63%

disagree by more than ε and keeps those that predict the answer. For the hue reader it moves the pointer's and the right digit's hues down toward 0.5, where this network's percepts switch. The shape reader's percepts vary from one handwriting to another, so the log bounds them only loosely even at observed hues, and ISS pins this digit's percepts at each hue within reach. Figure 3c shows the outcome for one percept: every admissible hue within the radius gets a certified value on the true curve, and beyond the radius the version space still allows anything.

The truth lies in V_0 at 96% of observed images. On all of them every run ends with a certificate, and every certificate holds within ε on every admissible intervention: 1,728 runs, without exception. At the other images the log's bounds miss some digit's percept, so the class is misspecified, and 12% of the certificates issued there are void (Appendix C). At $R = 0.5$ with visual interventions, ISS uses 13.6 interventions of total cost 4.0. Random separating queries cost 5.3, and random probing, with the same stopping test, 33.9 (Figure 3b).

Result 3: a wrong class is caught only by validation. The bypass networks separate behavior from abstraction. Their pointer's percept and their router both read the pointer's hue and agree on every observed image, so the log fits the class exactly. With visual interventions, ISS certifies their behavior correctly: the class can express a router that reads the pointer's hue. With interchange interventions no candidate matches, because setting the pointer's percept changes nothing. Yet at $R = 1$, 97% of white-box runs end with a void certificate: ISS only asks where the candidates disagree, and they all agree about what the pointer's percept does. Twenty random validation interventions refute 69% of these certificates, and random probing, which ignores disagreement, refutes the class in 97% of runs.

7 Discussion and Conclusion

Choosing among observationally equivalent causal hypotheses is an interventional learning problem: interventional risk is the loss, minimum separation cost the margin, and the budget fixes what can be certified. Intervening buys a *certificate*, not a better estimate: once no admissible intervention within radius R separates the survivors, every survivor agrees with the true system on all such interventions, provided the truth lies in the version space. The guarantee belongs to the stopping condition, not the acquisition heuristic, and it needs the full version space. Under noise it becomes probabilistic; under misspecification only a separate validation phase can falsify the class.

Limitations exact separation needs finitely many admissible interventions (our hues lie on a grid) and answers linear in the parameters up to sign decisions; the worst-case cost is exponential; parent sets and noise form are given; validation needs $\ln(1/\beta)/\rho$ queries to catch a void class of risk ρ with probability $1 - \beta$; and our networks expose their abstraction through one unit per digit, whereas distributed representations would first need a learned alignment (Geiger et al., 2024).

AI USE STATEMENT

Generative AI tools were used to draft text, to write and debug the experimental code. The authors take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI. Further we used Generative AI as an aid for proofs and mathematical theory. Finally we used AI to polish some of our sentences for readability.

REFERENCES

Eldar David Abraham, Karel D'Oosterlinck, Amir Feder, Yair Ori Gat, Atticus Geiger, Christopher Potts, Roi Reichart, and Zhengxuan Wu. CEBaB: Estimating the causal effects of real-world concepts on NLP model behavior. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Raj Agrawal, Chandler Squires, Karren Yang, Karthikeyan Shanmugam, and Caroline Uhler. Abcd-strategy: Budgeted experimental design for targeted causal structure discovery. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 3400–3409. PMLR, 2019.

Tatsuya Akutsu, Satoru Kuhara, Osamu Maruyama, and Satoru Miyano. Identification of genetic networks by strategic gene disruptions and gene overexpressions under a boolean model. *Theoretical Computer Science*, 298(1):235–251, 2003.

Gadi Aleksandrowicz, Hana Chockler, Joseph Y Halpern, and Alexander Ivrii. The computational complexity of structure-based causality. *Journal of Artificial Intelligence Research*, 58:431–451, 2017.

Steen A Andersson, David Madigan, and Michael D Perlman. A characterization of Markov equivalence classes for acyclic digraphs. *The Annals of Statistics*, 25(2):505–541, 1997.

Dana Angluin. Queries and concept learning. *Machine learning*, 2(4):319–342, 1988.

Yashas Annadani, Panagiotis Tigas, Stefan Bauer, and Adam Foster. Amortized active causal induction with deep reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Joshua F Apgar, Jared E Toettcher, Drew Endy, Forest M White, and Bruce Tidor. Stimulus design for model selection and validation in cell signaling. *PLoS computational biology*, 4(2):e30, 2008.

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization, 2019.

Anthony C. Atkinson and Valerii V. Fedorov. The design of experiments for discriminating between two rival models. *Biometrika*, 62(1):57–70, 1975. doi: 10.1093/biomet/62.1.57.

Alexander Balke and Judea Pearl. Counterfactual probabilities: Computational methods, bounds and applications. In *Uncertainty in artificial intelligence*, pp. 46–54. Elsevier, 1994.

Elias Bareinboim and Judea Pearl. Causal inference and the data-fusion problem. *Proceedings of the National Academy of Sciences*, 113(27):7345–7352, 2016.

Elias Bareinboim, Juan D Correa, Duligur Ibeling, and Thomas Icard. On Pearl's hierarchy and the foundations of causal inference. In *Probabilistic and Causal Inference: The Works of Judea Pearl*, pp. 507–556. Association for Computing Machinery, 2022.

Sander Beckers and Joseph Y Halpern. Abstracting causal models. In *Proceedings of the aaai conference on artificial intelligence*, volume 33, pp. 2678–2685, 2019.

Sander Beckers, Frederick Eberhardt, and Joseph Y Halpern. Approximate causal abstractions. In *Uncertainty in artificial intelligence*, pp. 606–615. PMLR, 2020.

Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in terra incognita. In *European conference on computer vision*, pp. 472–489. Springer, 2018.

Yoshua Bengio, Nicholas Léonard, and Aaron Courville. Estimating or propagating gradients through stochastic neurons for conditional computation. arXiv preprint arXiv:1308.3432, 2013.

Emily Black, Manish Raghavan, and Solon Barocas. Model multiplicity: Opportunities, concerns, and solutions. In *Proceedings of the 2022 ACM conference on fairness, accountability, and transparency*, pp. 850–863, 2022.

George EP Box and WILLIAM J Hill. Discrimination among mechanistic models. *Technometrics*, 9(1):57–71, 1967.

Leo Breiman. Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science*, 16(3):199–231, 2001.

Kathryn Chaloner and Isabella Verdinelli. Bayesian experimental design: A review. *Statistical Science*, 10(3):273–304, 1995. doi: 10.1214/SS/1177009939.

Davin Choo and Kirankumar Shiragur. Subset verification and search algorithms for causal DAGs. In *International Conference on Artificial Intelligence and Statistics*, pp. 4409–4442. PMLR, 2023.

Davin Choo, Kirankumar Shiragur, and Arnab Bhattacharyya. Verification and search algorithms for causal DAGs. In *Advances in Neural Information Processing Systems*, volume 35, pp. 12787–12799, 2022.

Arthur Conmy, Augustine N Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. Towards automated circuit discovery for mechanistic interpretability. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Alexander D'Amour, Katherine Heller, Dan Moldovan, Ben Adlam, Babak Alipanahi, Alex Beutel, Christina Chen, Jonathan Deaton, Jacob Eisenstein, Matthew D Hoffman, et al. Underspecification presents challenges for credibility in modern machine learning. *Journal of Machine Learning Research*, 23(226):1–61, 2022.

Pim De Haan, Dinesh Jayaraman, and Sergey Levine. Causal confusion in imitation learning. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Frederick Eberhardt. Almost optimal intervention sets for causal discovery. In *Proceedings of the 24th Conference on Uncertainty in Artificial Intelligence*, pp. 161–168, 2008.

Frederick Eberhardt and Richard Scheines. Interventions and causal inference. *Philosophy of science*, 74(5):981–995, 2007.

Frederick Eberhardt, Clark Glymour, and Richard Scheines. On the number of experiments sufficient and in the worst case necessary to identify all causal relations among N variables. In *Proceedings of the 21st Conference on Uncertainty in Artificial Intelligence*, pp. 178–184, 2005.

Thomas Eiter and Thomas Lukasiewicz. Complexity results for structure-based causality. *Artificial Intelligence*, 142(1):53–89, 2002.

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. Amnesic probing: Behavioral explanation with amnesic counterfactuals. *Transactions of the Association for Computational Linguistics*, 9:160–175, 2021.

Aaron Fisher, Cynthia Rudin, and Francesca Dominici. All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *Journal of machine learning research*, 20(177):1–81, 2019.

Yoav Freund, H Sebastian Seung, Eli Shamir, and Naftali Tishby. Selective sampling using the query by committee algorithm. *Machine learning*, 28(2):133–168, 1997.

Matt Gardner, Yoav Artzi, Victoria Basmov, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hannaneh Hajishirzi, Gabriel Ilharco, Daniel Khashabi, Kevin Lin, Jiangming Liu, Nelson F Liu, Phoebe Mulcaire, Qiang Ning, Sameer Singh, Noah A Smith, Sanjay Subramanian, Reut Tsarfaty, Eric Wallace, Ally Zhang, and Ben Zhou. Evaluating models' local decision boundaries via contrast sets. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 1307–1323, 2020.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. Causal abstractions of neural networks. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 9574–9586. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/4f5c422f4d49a5a807eda27434231040-Paper.pdf.

Atticus Geiger, Zhengxuan Wu, Hanson Lu, Josh Rozner, Elisa Kreiss, Thomas Icard, Noah Goodman, and Christopher Potts. Inducing causal structure for interpretable neural networks. In *Proceedings of the 39th International Conference on Machine Learning*, pp. 7324–7338, 2022.

Atticus Geiger, Zhengxuan Wu, Christopher Potts, Thomas Icard, and Noah Goodman. Finding alignments between interpretable causal variables and distributed neural representations. In *Causal Learning and Reasoning*, pp. 160–187. PMLR, 2024.

Atticus Geiger, Duligur Ibeling, Amir Zur, Maheep Chaudhary, Sonakshi Chauhan, Jing Huang, Aryaman Arora, Zhengxuan Wu, Noah Goodman, Christopher Potts, et al. Causal abstraction: A theoretical foundation for mechanistic interpretability. *Journal of Machine Learning Research*, 26(83):1–64, 2025.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

AmirEmad Ghassami, Saber Salehkaleybar, Negar Kiyavash, and Elias Bareinboim. Budgeted experiment design for causal structure learning. In *International Conference on Machine Learning*, pp. 1724–1733. PMLR, 2018.

Yujia Guo, Daolang Huang, Xinyu Zhang, Sammie Katt, Samuel Kaski, and Ayush Bharti. Constrained Bayesian experimental design via online planning. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026.

Joseph Y Halpern. *Actual causality*, volume 3. MiT Press Cambridge, 2016.

Joseph Y Halpern and Judea Pearl. Causes and explanations: A structural-model approach. part i: Causes. *The British journal for the philosophy of science*, 2005a.

Joseph Y Halpern and Judea Pearl. Causes and explanations: A structural-model approach. part ii: Explanations. *The British journal for the philosophy of science*, 2005b.

Alain Hauser and Peter Bühlmann. Characterization and greedy learning of interventional Markov equivalence classes of directed acyclic graphs. *The Journal of Machine Learning Research*, 13 (1):2409–2464, 2012.

Alain Hauser and Peter Bühlmann. Two optimal strategies for active learning of causal models from interventional data. *International Journal of Approximate Reasoning*, 55(4):926–939, 2014.

Miguel A. Hernán and James M. Robins. *Causal Inference: What If*. Chapman & Hall/CRC, 2020. URL https://miguelhernan.org/whatifbook.

Neil Houlsby, Ferenc Huszár, Zoubin Ghahramani, and Máté Lengyel. Bayesian active learning for classification and preference learning, 2011.

William G Hunter and Albey M Reiner. Designs for discriminating between two rival models. *Technometrics*, 7(3):307–323, 1965.

Duligur Ibeling and Thomas Icard. Probabilistic reasoning across the causal hierarchy. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 10170–10177, 2020.

Duligur Ibeling and Thomas Icard. Comparing causal frameworks: Potential outcomes, structural models, graphs, and abstractions. In *Advances in Neural Information Processing Systems*, volume 36, pp. 80130–80141, 2023.

Amjad Ibrahim and Alexander Pretschner. From checking to inference: Actual causality computations as optimization problems. In *International Symposium on Automated Technology for Verification and Analysis*, pp. 343–359. Springer, 2020.

Trey E. Ideker, Vesteinn Thorsson, and Richard M. Karp. Discovery of regulatory interactions through perturbation: Inference and experimental design. In *Pacific Symposium on Biocomputing 2000*, pp. 305–316. World Scientific, 2000.

Yiding Jiang, Vaishnavh Nagarajan, Christina Baek, and J Zico Kolter. Assessing generalization of SGD via disagreement. In *International Conference on Learning Representations*, 2022.

Zhijing Jin, Yuen Chen, Felix Leeb, Luigi Gresele, Ojasv Kamal, Zhiheng Lyu, Kevin Blin, Fernando Gonzalez Adauto, Max Kleiman-Weiner, Mrinmaya Sachan, et al. CLadder: Assessing causal reasoning in language models. In *Advances in Neural Information Processing Systems*, volume 36, pp. 31038–31065, 2023.

Amir-Hossein Karimi, Bernhard Schölkopf, and Isabel Valera. Algorithmic recourse: from counterfactual explanations to interventions. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pp. 353–362, 2021.

Amir-Hossein Karimi, Gilles Barthe, Bernhard Schölkopf, and Isabel Valera. A survey of algorithmic recourse: contrastive explanations and consequential recommendations. *ACM Computing Surveys*, 55(5):1–29, 2022.

Stuart A Kauffman. Metabolic stability and epigenesis in randomly constructed genetic nets. *Journal of theoretical biology*, 22(3):437–467, 1969.

Divyansh Kaushik, Eduard Hovy, and Zachary C Lipton. Learning the difference that makes a difference with counterfactually-augmented data. In *International Conference on Learning Representations*, 2020.

Emre Kıcıman, Robert Ness, Amit Sharma, and Chenhao Tan. Causal reasoning and large language models: Opening a new frontier for causality. *Transactions on Machine Learning Research*, 2024.

Murat Kocaoglu, Alex Dimakis, and Sriram Vishwanath. Cost-optimal learning of causal graphs. In *International Conference on Machine Learning*, pp. 1875–1884. PMLR, 2017.

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 5338–5348, 2020.

Kevin B Korb, Lucas R Hope, Ann E Nicholson, and Karl Axnick. Varieties of causal intervention. In *Pacific Rim international conference on artificial intelligence*, pp. 322–331. Springer, 2004.

Clemens Kreutz and Jens Timmer. Systems biology: experimental design. *The FEBS journal*, 276 (4):923–942, 2009.

Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

David Lee and Mihalis Yannakakis. Principles and methods of testing finite state machines-a survey. *Proceedings of the IEEE*, 84(8):1090–1123, 1996.

Erik Lindgren, Murat Kocaoglu, Alexandros G Dimakis, and Sriram Vishwanath. Experimental design for cost-aware learning of causal graphs. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Dennis V Lindley. On a measure of the information provided by an experiment. *The Annals of Mathematical Statistics*, 27(4):986–1005, 1956.

Lars Lorch, Scott Sussex, Jonas Rothfuss, Andreas Krause, and Bernhard Schölkopf. Amortized inference for causal structure learning. In *Advances in Neural Information Processing Systems*, volume 35, pp. 13104–13118, 2022.

Chaochao Lu, Biwei Huang, Ke Wang, José Miguel Hernández-Lobato, Kun Zhang, and Bernhard Schölkopf. Sample-efficient reinforcement learning via counterfactual-based data augmentation, 2020.

Charles Marx, Flavio Calmon, and Berk Ustun. Predictive multiplicity in classification. In *International conference on machine learning*, pp. 6765–6774. PMLR, 2020.

William M McKeeman. Differential testing for software. *Digital Technical Journal*, 10(1):100–107, 1998.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*, volume 35, pp. 17359–17372, 2022.

Tom M Mitchell. Generalization as search. *Artificial intelligence*, 18(2):203–226, 1982.

Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: a simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2574–2582, 2016.

Milan Mossé, Duligur Ibeling, and Thomas Icard. Is causal reasoning harder than probabilistic reasoning? *The Review of Symbolic Logic*, 17(1):106–131, 2024.

Ramaravind K Mothilal, Amit Sharma, and Chenhao Tan. Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pp. 607–617, 2020.

Kevin P Murphy. Active learning of causal Bayes net structure. Technical report, University of California, Berkeley, 2001.

Michael Oberst and David Sontag. Counterfactual off-policy evaluation with gumbel-max structural causal models. In *International Conference on Machine Learning*, pp. 4881–4890. PMLR, 2019.

Giambattista Parascandolo, Niki Kilbertus, Mateo Rojas-Carulla, and Bernhard Schölkopf. Learning independent causal mechanisms. In *International Conference on Machine Learning*, pp. 4036–4044. PMLR, 2018.

Judea Pearl. Causal diagrams for empirical research. *Biometrika*, 82(4):669–688, 1995.

Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, Cambridge, UK, 2 edition, 2009.

Kexin Pei, Yinzhi Cao, Junfeng Yang, and Suman Jana. Deepxplore: Automated whitebox testing of deep learning systems. In *proceedings of the 26th Symposium on Operating Systems Principles*, pp. 1–18, 2017.

Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 78(5):947–1012, 2016.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. The MIT press, 2017.

Karl Popper. *The Logic of Scientific Discovery*. Routledge, London, 2002. Routledge Classics edition; first English edition Hutchinson, London, 1959.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4902–4912, 2020.

Jonathan Richens and Tom Everitt. Robust agents learn causal world models. In *International Conference on Learning Representations*, 2024.

Paul K Rubenstein, Sebastian Weichwald, Stephan Bongers, Joris M Mooij, Dominik Janzing, Moritz Grosse-Wentrup, and Bernhard Schölkopf. Causal consistency of structural equation models. In *Proceedings of the 33rd Conference on Uncertainty in Artificial Intelligence*, 2017.

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 *International Conference on Learning Representations*, 2020.

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

Lesia Semenova, Cynthia Rudin, and Ronald Parr. On the existence of simpler machine learning models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1827–1858, 2022.

H Sebastian Seung, Manfred Opper, and Haim Sompolinsky. Query by committee. In *Proceedings of the fifth annual workshop on Computational learning theory*, pp. 287–294, 1992.

Karthikeyan Shanmugam, Murat Kocaoglu, Alexandros G Dimakis, and Sriram Vishwanath. Learning causal graphs with small interventions. In *Advances in Neural Information Processing Systems*, volume 28, 2015.

Galit Shmueli. To explain or to predict? *Statistical science*, pp. 289–310, 2010.

Ilya Shpitser and Judea Pearl. Complete identification methods for the causal hierarchy. *Journal of Machine Learning Research*, 9:1941–1979, 2008.

Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, Prediction, and Search*. MIT Press, Cambridge, MA, 2 edition, 2000.

Chandler Squires, Sara Magliacane, Kristjan Greenewald, Dmitriy Katz, Murat Kocaoglu, and Karthikeyan Shanmugam. Active structure learning of causal DAGs via directed clique trees. In *Advances in Neural Information Processing Systems*, volume 33, pp. 21500–21511, 2020.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations*, 2014.

Jin Tian and Judea Pearl. Probabilities of causation: Bounds and identification. *Annals of Mathematics and Artificial Intelligence*, 28(1):287–313, 2000.

Panagiotis Tigas, Yashas Annadani, Andrew Jesson, Bernhard Schölkopf, Yarin Gal, and Stefan Bauer. Interventions, where and how? Experimental design for causal models at scale. In *Advances in Neural Information Processing Systems*, volume 35, pp. 24130–24143, 2022.

Simon Tong and Daphne Koller. Active learning for structure in Bayesian networks. In *Proceedings of the 17th International Joint Conference on Artificial Intelligence*. Morgan Kaufmann, 2001.

Berk Ustun, Alexander Spangher, and Yang Liu. Actionable recourse in linear classification. In *Proceedings of the conference on fairness, accountability, and transparency*, pp. 10–19, 2019.

Victor Veitch, Alexander D'Amour, Steve Yadlowsky, and Jacob Eisenstein. Counterfactual invariance to spurious correlations in text classification. In *Advances in Neural Information Processing Systems*, volume 34, pp. 16196–16208, 2021.

Thomas S Verma and Judea Pearl. Equivalence and synthesis of causal models. In *Proceedings of the 6th Conference on Uncertainty in Artificial Intelligence*, 1990.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. Investigating gender bias in language models using causal mediation analysis. In *Advances in Neural Information Processing Systems*, volume 33, pp. 12388–12401, 2020.

Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law & Technology*, 31 (2):841–887, 2018.

Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: a circuit for indirect object identification in GPT-2 small. In *International Conference on Learning Representations*, 2023.

James Woodward. *Making Things Happen: A Theory of Causal Explanation*. Oxford University Press, Oxford, 2003.

Kevin Xia, Kai-Zhan Lee, Yoshua Bengio, and Elias Bareinboim. The causal-neural connection: Expressiveness, learnability, and inference. In *Advances in Neural Information Processing Systems*, volume 34, pp. 10823–10836, 2021.

Xiaofei Xie, Lei Ma, Haijun Wang, Yuekang Li, Yang Liu, and Xiaohong Li. DiffChaser: Detecting disagreements for deep neural networks. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pp. 5772–5778, 2019.

Karren Yang, Abigail Katcoff, and Caroline Uhler. Characterizing and learning equivalence classes of causal dags under interventions. In *International Conference on Machine Learning*, pp. 5541–5550. PMLR, 2018.

Matej Zečević, Moritz Willig, Devendra Singh Dhami, and Kristian Kersting. Causal parrots: Large language models may talk causality but are not causal. *Transactions on Machine Learning Research*, 2023.

Chiyuan Zhang, Maithra Raghu, Jon Kleinberg, and Samy Bengio. Pointer value retrieval: A new benchmark for understanding the limits of neural network generalization. arXiv preprint arXiv:2107.12580, 2021.

Junzhe Zhang and Elias Bareinboim. Fairness in decision-making—the causal explanation formula. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Zeyu Zhang, Chaozhuo Li, Xu Chen, and Xing Xie. Bayesian active causal discovery with multi-fidelity experiments. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

A Proofs

A.1 Proof of Proposition 1

By definition $R_r(P, P') = |\mathcal{W}_{PP'} \cap \mathcal{Q}_r|/|\mathcal{Q}_r|$. If $r < k_{\mathcal{Q},c}(P, P')$ then every $q \in \mathcal{W}_{PP'}$ has $c(q) \geq k_{\mathcal{Q},c}(P, P') > r$, so $\mathcal{W}_{PP'} \cap \mathcal{Q}_r = \emptyset$ and $R_r = 0$. If $r \geq k_{\mathcal{Q},c}(P, P')$ then a minimizing $q^\star \in \mathcal{W}_{PP'}$ has $0 < w_{\min} \leq c(q^\star) = k_{\mathcal{Q},c}(P, P') \leq r$ by Proposition 6(c), which applies because P and P' are history-equivalent, so $q^\star \in \mathcal{W}_{PP'} \cap \mathcal{Q}_r$ and $R_r > 0$. The three stated equivalences follow, the last one because $P \equiv^r_{\mathcal{Q},c} P'$ was defined as agreement on all of $\mathcal{Q}_r$. History-equivalence is needed: otherwise a cost-zero query can separate P and P', giving $k_{\mathcal{Q},c}(P, P') = 0 \leq r$, while $\mathcal{Q}_r$, which excludes cost-zero queries, may contain no separating query. □

A.2 Proof of Proposition 2

With $p = 0$ every non-root noise bit is 0, so each sample x satisfies $x_j = g^\star_j(x_{\mathrm{Pa}(j)})$ for every non-root j; a hypothesis g has zero training error exactly when $g_j(x_{\mathrm{Pa}(j)}) = x_j$ for every sample and every j, that is, when $g_j(z) = g^\star_j(z)$ for every cell (j, z) realized by some sample. Cells outside $\mathrm{supp}(S_N)$ appear in no constraint, so $V_0(S_N)$ is the product of $\{g^\star_j(z)\}$ over observed cells with $\{0, 1\}$ over the m_N unobserved ones, of size 2^{m_N}. Every member has zero empirical error and therefore the same likelihood. Since $\mathrm{supp}(S_N) \subseteq \mathrm{supp}(S_{N+1})$ and both are contained in the set of cells of positive probability under the regime, m_N is non-increasing and converges to the number m_∞ of cells of probability zero.

For $p > 0$ the log-likelihood of a hypothesis factorizes over cells: a cell (j, z) observed $n_z \geq 1$ times with k_z ones contributes $k_z \log \rho + (n_z - k_z) \log(1 - \rho)$ with $\rho = p$ if $g_j(z) = 0$ and $\rho = 1 - p$ if $g_j(z) = 1$. Since $p < 1/2$ this is maximized by the majority value, independently across cells, and cells with $n_z = 0$ contribute nothing. □

A.3 Proof of Theorem 1

Fix S_N and abbreviate $V_0 := V_0(S_N)$. For each r, $R_r(\cdot, \cdot)$ is a normalized Hamming distance between the answer vectors $(A_P(q))_{q \in \mathcal{Q}_r}$, hence a pseudo-metric; in particular it satisfies the triangle inequality. Let $P, P' \in V_0$ be arbitrary and put $\delta := R_r(P, P')$.

A learner using only observational data sees the same input S_N whether the environment is P or P', because both have zero training error on S_N and are therefore consistent with it; its output distribution is the same in the two cases. Writing $\hat{P} = L(S_N)$ for a draw from that distribution,

$$\mathbb{E}\, R_r(\hat{P}, P) + \mathbb{E}\, R_r(\hat{P}, P') \;=\; \mathbb{E}\big[R_r(\hat{P}, P) + R_r(\hat{P}, P')\big] \;\geq\; R_r(P, P') = \delta,$$

so at least one of the two expectations is at least $\delta/2$. Since both P and P' lie in V_0, the supremum over $P^\star \in V_0$ of $\mathbb{E}\, R_r(L(S_N); P^\star)$ is at least $\delta/2$. Taking the maximum over $P, P' \in V_0$ gives equation 2, and the second inequality is the definition of R_r for the particular pair.

By Proposition 2, V_0 is determined by $\mathrm{supp}(S_N)$ alone, so the right-hand side is a function of the observational support. Any two completions P, P' that agree on all observed cells and are separated by some $q \in \mathcal{Q}_r$ give $R_r(P, P') \geq 1/|\mathcal{Q}_r| > 0$. If such a pair exists among the completions of the m_∞ cells that the regime reaches with probability zero, then it remains in $V_0(S_N)$ almost surely for every N, and the bound does not vanish. □

A.4 Proof of Theorem 2

(a) At each iteration the algorithm keeps exactly those P with $A_P(q) = y = A_{P^\star}(q)$, a condition $P^\star$ satisfies; and $P^\star \in V_0$ by assumption, the initial V being V_0 intersected with the evidence constraint, which $P^\star$ also satisfies because $u_E^\star$ are its own exogenous values. In particular $V \neq \emptyset$ throughout, so the refutation branch of Algorithm 1 is reached only if $P^\star \notin V_0$.

(b) The chosen q separates two members $P, P' \in V$, so $A_P(q) \neq A_{P'}(q)$ and at least one of them differs from y and is removed; $P^\star$ is never removed by (a), so $1 \leq |V_{t+1}| < |V_t|$ and the loop runs at most $|V_0| - 1$ times.

(c) Halting with $\kappa(V) > R$ means no pair in V has separation cost at most R, i.e. no query in $\mathcal{Q}_R$ separates any two members. By (a) $P^\star \in V$, so for every $\hat{P} \in V$ and every $q \in \mathcal{Q}_R$ we have $A_{\hat{P}}(q) = A_{P^\star}(q)$, which is the definition of $\hat{P} \equiv^R_{\mathcal{Q},c} P^\star$. Proposition 1 then gives $R_r(\hat{P}) = 0$ for all $r \leq R$, since $\mathcal{Q}_r \subseteq \mathcal{Q}_R$. If instead it halts on the budget, let $r := \kappa(V) - w_{\min} < \kappa(V)$. Every $\hat{P} \in V$ equals $P^\star$ or satisfies $k_{\mathcal{Q},c}(\hat{P}, P^\star) \geq \kappa(V) > r$, since $P^\star \in V$, so $\hat{P} \equiv^r_{\mathcal{Q},c} P^\star$ by Proposition 1, and the same argument applies with R replaced by r.

(d) $\kappa(V)$ is computed from the answers of the members of V to admissible queries, which the learner can evaluate; no property of $P^\star$ beyond the observed answers y is used. □

A.5 Proof of Theorem 3

(a) At step t the algorithm selects a minimum-cost query separating two members of V_t, whose cost is $\min_{P \neq P' \in V_t} \min_{q \in \mathcal{W}_{PP'}} c(q) = \kappa(V_t)$. Since $V_{t+1} \subseteq V_t$, the minimum over the smaller pair set is no smaller, so $\kappa(V_t)$ is non-decreasing in t and $\sum_{t<T} \kappa(V_t) \leq T\kappa(V_{T-1})$. The loop condition enforces $\kappa(V_t) \leq R$, and $T \leq |V_0| - 1$ by Theorem 2(b).

(b) *Truth-specific cost bound.* Fix the environment $P^\star$ and any $P' \in V_0$ with $k_{\mathcal{Q},c}(P^\star, P') \leq R$. Run the learner in environment $P^\star$; it executes queries $q_1, \ldots, q_T$ (chosen adaptively) and halts with a certificate. Suppose no q_i separates $P^\star$ from P'. Then in environment P' the learner receives the same answers $A_{P'}(q_i) = A_{P^\star}(q_i)$, hence executes the same queries and halts with the same output and the same certificate. Both $P^\star$ and P' are consistent with the observational data and with all observed answers, but they are not R-locally equivalent, so a certificate asserting that

every consistent hypothesis is R-locally equivalent to the output is false in at least one of the two environments, contradicting correctness. Hence some q_i separates $P^\star$ from P' and, by the definition of $k_{\mathcal{Q},c}$, costs at least $k_{\mathcal{Q},c}(P^\star, P')$. Taking the maximum over P' gives the bound $k^\star(P^\star)$. For a randomized learner, fix its random seed: the learner is then deterministic, and the argument applies to every run that halts with a correct certificate. The bound therefore holds run by run, and so does the query count below, seed by seed. For the worst case over $P^\star \in V_0$, let (P, P') attain $\max\{k_{\mathcal{Q},c}(P, P') : k_{\mathcal{Q},c}(P, P') \le R\}$ and take $P^\star = P$.

Query count. With a binary target, the run in environment $P^\star$ produces an answer string $a(P^\star) \in \{0,1\}^{T(P^\star)}$. Let $P^\star, P^{\star\prime}$ lie in different R-local classes. If $a(P^\star)$ were a prefix of $a(P^{\star\prime})$ (or equal), the learner in environment $P^{\star\prime}$ would, after $T(P^\star)$ answers, have seen exactly the string $a(P^\star)$, executed the same queries, and halted with the same certificate as in environment $P^\star$—false for one of the two, as above. Hence the strings $a(\cdot)$ of representatives of the C classes form a prefix-free code with C codewords. By Kraft's inequality $\sum_{i \le C} 2^{-T_i} \le 1$, so $\max_i T_i \ge \lceil \log_2 C \rceil$, and by the source-coding bound the expected length under the uniform distribution over classes is at least $\log_2 C$.

(c) Let the ground set be $\Pi := \{\{P, P'\} \subseteq V_0 : k_{\mathcal{Q},c}(P, P') \le R\}$, with $M := |\Pi| \le \binom{|V_0|}{2}$, and for $q \in \mathcal{Q}_R$ let $\mathrm{cov}(q) := \{\{P, P'\} \in \Pi : q \in \mathcal{W}_{PP'}\}$ with weight $c(q) > 0$. Every pair in Π is covered by at least one $q \in \mathcal{Q}_R$, by the definition of $k_{\mathcal{Q},c}$. Minimizing $\sum_{q \in \mathcal{Q}'} c(q)$ over covers $\mathcal{Q}'$ of Π is weighted set cover, and the stated rule is its greedy algorithm, whose output has cost at most $H_M \,\mathrm{OPT} \le (1 + \ln M)\mathrm{OPT}$ with $H_M = \sum_{i \le M} 1/i$. Running all queries of a cover leaves a version space in which no two members are separated by any query of cost at most R: if P, P' both survive they agreed with $P^\star$, hence with each other, on every executed query, and any pair with $k_{\mathcal{Q},c}(P, P') \le R$ is covered and so was separated. Therefore $\kappa > R$ and the radius is certified by Theorem 2(c). Adaptivity is not used by the cover, and an adaptive learner may spend less than OPT: it skips the queries of the cover that only separate pairs already refuted. □

A.6 Proof of Corollary 1

Let $q = (\mathrm{do}(X_S = a_S), T) \in \mathcal{Q}_R$ and let $x^\star$ be the state of $P^\star$ under q; by assumption $x^\star$ is realized by a query in $\mathcal{Q}_R$. We show $x(M^I, u_P) = x^\star$ by induction along a topological order of M. For $s \in S$ both equal a_s. For $l \notin S$, assume the claim for all predecessors; then φ_l^P and $\varphi_l^{P^\star}$ are evaluated at the same state $x^\star$, on which they agree by hypothesis, so the l-th coordinates agree. Hence $A_P(q) = A_{P^\star}(q)$ for every $q \in \mathcal{Q}_R$, so P is never eliminated by any answer, whichever queries the rule chooses. That the condition is not necessary: take P differing from $P^\star$ at a single cell (l, z) with $l \notin T$ and l not an ancestor of T in either model; the state under a query realizing z differs at l but x_T is unchanged, so P answers every query like $P^\star$ and survives although a reachable difference exists. The description of the terminal class is Theorem 2(c) read as a set, and $R_r(\hat{P})$ for $r > R$ counts the queries on which a member of that class disagrees with $P^\star$, whichever of the two kinds of difference causes it. □

A.7 Proof of Proposition 7

If $E \subseteq E'$ and $u_{P,E'} = u^\star_{E'}$ then $u_{P,E} = u^\star_E$, so $V_0^{E'} \subseteq V_0^E$; a minimum over a smaller set of pairs is no smaller, giving $\kappa(V_0^{E'}) \ge \kappa(V_0^E)$. The bound of Theorem 3(a) is increasing in $|V_0|$, which shrinks, and the halting condition $\kappa > R$ is implied for E' whenever it holds for E. □

A.8 Proof of Proposition 3

(i) Under Proposition 2 all members of V_0 share the observed cells and the abduced context, so the state of P under $q = (\mathrm{do}(X_S = a_S), T)$ is obtained by propagating in topological order: intervened variables take their clamped values, roots their context values, and a non-root j the table entry of the cell (j, z) its parents currently realize. If that cell is observed the entry is common to all of V_0; if it is unobserved the entry is the cell's free bit. Branching on the free bit whenever an unobserved cell is met, and continuing each branch, produces a tree whose leaves are the possible trajectories; a leaf fixes the target's value and is reached by all completions agreeing with the partial assignment read along the way, which is its cube. Each non-root contributes at most one branching per trajectory,

so there are at most $2^{n'}$ leaves, and the cubes partition V_0 (each completion follows exactly one trajectory), so f_q equals the DNF whose cubes with answer y are the leaves labeled y.

(ii) $V_t = \{b \in V_0 : f_{q_s}(b) = y_s,\ s \le t\}$. Two members b, b' answer q differently iff they lie in cubes of f_q with different labels, i.e. iff two cubes with different labels each contain a member of V_t. A cube κ contains a member of V_t iff the constraints have a 0-1 solution that agrees with the literals of κ, which is one feasibility check with those literals fixed as variable bounds. Each constraint $\{f_{q_s} = y_s\}$ is encoded with one binary indicator e_λ per cube λ of f_{q_s} labeled y_s: the indicators sum to at least one, and each forces the literals of its cube, with $b_i \ge e_\lambda$ if λ sets cell i to 1 and $b_i \le 1 - e_\lambda$ if it sets it to 0. Constraints are added as answers arrive.

(iii) $\kappa(V_t)$ is the cost of the first separating query in cost order, and the absence of a separating query in $\mathcal{Q}_R$ is exactly the halting condition $\kappa(V_t) > R$ of Theorem 2(c); both are decided by the tests of (ii), at most $|f_q|$ per query, hence $O(|\mathcal{Q}_R| \max_q |f_q|)$ feasibility checks per step, each on a program whose size is linear in the number of cubes seen so far.

(iv) If b is uniform on V_0 and $q_1, \dots, q_t$ are fixed, then conditional on $f_{q_s}(b) = y_s$ for $s \le t$, b is uniform on V_t, so filtering a uniform sample by the answers to fixed queries yields a uniform sample of V_t. This fails when q_s is chosen as a function of the sample: conditional on the event that a particular query maximized the sample's split, the sample's composition is no longer exchangeable with the rest of V_t (elements that make that query the maximizer are over-represented), so statistics averaged over the surviving sample are biased in general. We therefore use the acquisition sample only to choose queries, and evaluate risk on a fresh uniform sample of V_t, drawn independently: by enumerating the assignments of the variables the constraints mention (the others are free and uniform) when there are at most 22, otherwise by rejection from uniform draws of V_0, each draw accepted iff it satisfies every constraint. Since acceptance depends only on the draw, the accepted draws are i.i.d. uniform on V_t. □

A.9 Proof of Proposition 4

Each elimination keeps the hypotheses whose answer lies within η of the observed one, and $P^\star$'s answer does, so $P^\star$ survives every step. Halting with $\kappa^\varepsilon(V) > R$ means that no $q \in \mathcal{Q}_R$ ε-separates two members of V; since $P^\star \in V$, every survivor is within ε of $P^\star$ on every $q \in \mathcal{Q}_R$. For termination, the loop runs only ε-separating queries. Once it has run q and observed y, every survivor is within η of y at q, so any two survivors are within $2\eta < \varepsilon$ there and q never ε-separates again; each query of the finite family $\mathcal{Q}_R$ is therefore run at most once. The proof of Theorem 1 uses only the triangle inequality of the risk and the fact that members of V_0 cannot be told apart by observation; both hold for $\mathbb{E}_{q\sim\pi_r}\|A_P(q) - A_{P'}(q)\|_\infty$. □

A.10 Proof of Theorem 4

Order the variables topologically. The cell counts $(n_z)_{z\in j}$ of variable j are determined by the roots and by the noise bits of the ancestors of j; the noise bits $u_j^{(i)}$ of the N samples at variable j are i.i.d. $\mathrm{Bern}(p)$ and independent of everything at other variables (Assumption 1). Conditional on the counts of j—indeed on all roots and all noise at other variables—the samples in cell z have $x_j^{(i)} = g_j^\star(z) \oplus u_j^{(i)}$ with i.i.d. noise, so the number K_z of samples agreeing with the true value $g_j^\star(z)$ is $\mathrm{Bin}(n_z, 1-p)$, independently across the (disjoint) cells of j. The majority fit is wrong at z iff $K_z < n_z/2$, and then its margin is $n_z - 2K_z$; flipping it back to the truth changes the log-likelihood by $(n_z - 2K_z)\log\frac{1-p}{p}$, so $D_j/\ell = \sum_{z\in j}\max(n_z - 2K_z, 0)$ with the stated distribution. (Conditioning on the counts of *other* variables would break this independence, because downstream counts depend on x_j; this is why the budgets are per variable.)

Here j ranges over the n' non-root variables: a root satisfies $X_j = U_j$, has no fitted cells and so needs no budget. For each such j, T_j is a function of $(n_z)_{z\in j}$ and $\Pr[D_j/\ell > T_j \mid (n_z)_{z\in j}] \le \delta/n'$ by construction; by the tower property $\Pr[D_j/\ell > T_j] \le \delta/n'$ unconditionally, and by the union bound over these n' variables $\Pr[\exists j : D_j/\ell > T_j] \le n' \cdot \delta/n' = \delta$. On the complementary event the truth differs from the fit, on each variable j, only on fitted cells of total margin $\sum_z d_z \le T_j$ (and arbitrarily on unobserved cells), i.e. $P^\star \in V_0^\delta$. On that event Theorem 2 applies verbatim to

the initial version space V_0^δ: the truth is never eliminated, and at halting every survivor is R-locally equivalent to it. □

A.11 Proof of Proposition 6

(a) Symmetry holds because $\mathcal{W}_{ij} = \mathcal{W}_{ji}$, and $\mathcal{W}_{ii} = \emptyset$ gives $k_{\mathcal{Q},c}(P_i, P_i) = \min \emptyset = \infty$.

(b) If $\mathcal{Q} \subseteq \mathcal{Q}'$ then the separating set with respect to $\mathcal{Q}$ is contained in the one with respect to $\mathcal{Q}'$, so the minimum over the larger set is no larger. If $c \leq c'$ pointwise, the minimum of c over the (unchanged) separating set is at most the minimum of c'.

(c) Let $q = (I, T)$ with $I = \mathrm{do}(X_S = a_S)$ and $c(q) = 0$; since all $w_s > 0$, this means $a_s = x_s^{\mathrm{obs}}$ for every $s \in S$. By history-equivalence, x^{obs} satisfies every structural equation of M_i under u_i. It also satisfies the replaced equations $X_s = a_s = x_s^{\mathrm{obs}}$. Hence x^{obs} is a solution of M_i^I under u_i, and since M_i^I is acyclic (removing incoming edges preserves acyclicity) its solution is unique, so $x(M_i^I, u_i) = x^{\mathrm{obs}}$ and $A_{P_i}(q) = x_T^{\mathrm{obs}}$. The same argument applies to P_j. Therefore no zero-cost query belongs to $\mathcal{W}_{ij}$, and every $q \in \mathcal{W}_{ij}$ has $c(q) \geq w_{\min}$.

(d) Let $q \in \mathcal{W}_{il}$, so $A_{P_i}(q) \neq A_{P_l}(q)$. Then $A_{P_j}(q)$ differs from at least one of $A_{P_i}(q)$ and $A_{P_l}(q)$, i.e. $q \in \mathcal{W}_{ij} \cup \mathcal{W}_{jl}$. Hence $\mathcal{W}_{il} \subseteq \mathcal{W}_{ij} \cup \mathcal{W}_{jl}$ and

$$k_{\mathcal{Q},c}(P_i, P_l) = \min_{q \in \mathcal{W}_{il}} c(q) \;\geq\; \min_{q \in \mathcal{W}_{ij} \cup \mathcal{W}_{jl}} c(q) = \min\{k_{\mathcal{Q},c}(P_i, P_j), k_{\mathcal{Q},c}(P_j, P_l)\}.$$

Reflexivity and symmetry of $\equiv^r_{\mathcal{Q},c}$ are immediate; transitivity follows because $k_{\mathcal{Q},c}(P_i, P_j) > r$ and $k_{\mathcal{Q},c}(P_j, P_l) > r$ imply $k_{\mathcal{Q},c}(P_i, P_l) > r$. If $r \leq r'$ then $\equiv^{r'}$ implies $\equiv^r$, so the classes of $\equiv^{r'}$ refine those of $\equiv^r$. Finally $d = 1/k_{\mathcal{Q},c}$ is symmetric, non-negative, satisfies $d(P, P) = 0$, and $d(P_i, P_l) \leq \max\{d(P_i, P_j), d(P_j, P_l)\}$ is the inequality just proved. □

A.12 Proof of Lemma 1

Let $D := \{l : x_l^i \neq x_l^j\} \neq \emptyset$. For $s \in S$ both states equal a_s, so $D \cap S = \emptyset$. Choose $l \in D$ that is minimal with respect to a topological order of M_i. Every parent $p \in \mathrm{Pa}_{M_i}(l)$ precedes l in that order, hence $p \notin D$ and $x_p^i = x_p^j$. Since $l \notin S$,

$$x_l^i = \varphi_l^{P_i}(x^i) = f_l^i\big(x^i_{\mathrm{Pa}_{M_i}(l)}, u_{i,l}\big) = f_l^i\big(x^j_{\mathrm{Pa}_{M_i}(l)}, u_{i,l}\big) = \varphi_l^{P_i}(x^j),$$

while $x_l^j = \varphi_l^{P_j}(x^j)$ because $l \notin S$. As $x_l^i \neq x_l^j$, we conclude $\varphi_l^{P_i}(x^j) \neq \varphi_l^{P_j}(x^j)$. □

A.13 Proof of Theorem 5

Infinite separation when induced mechanisms agree. If $\varphi^{P_i} = \varphi^{P_j}$, Lemma 1 shows that $x(M_i^I, u_i) = x(M_j^I, u_j)$ for every intervention I, so $A_{P_i}(q) = A_{P_j}(q)$ for every query and $\mathcal{W}_{ij} = \emptyset$ for every $\mathcal{Q}$. Hence $k_{\mathcal{Q},c}(P_i, P_j) = \infty$ for all $\mathcal{Q}, c$, and in particular $\Delta = \emptyset$ implies $k_{\mathcal{Q}_{\mathrm{full}},c} = \infty$.

Finite separation when they differ. Let $l \in \Delta$ and $z \in \{0,1\}^n$ with $\varphi_l^{P_i}(z) \neq \varphi_l^{P_j}(z)$. Consider $q = (\mathrm{do}(X_{R_l} = z_{R_l}), \{l\}) \in \mathcal{Q}_{\mathrm{full}}$, where $R_l = \mathrm{Pa}_{M_i}(l) \cup \mathrm{Pa}_{M_j}(l)$; acyclicity gives $l \notin R_l$. In M_i^I the variables in R_l take the values z_{R_l}, and $l \notin R_l$ takes the value $f_l^i(z_{\mathrm{Pa}_{M_i}(l)}, u_{i,l}) = \varphi_l^{P_i}(z)$; likewise $A_{P_j}(q) = \varphi_l^{P_j}(z)$. So $q \in \mathcal{W}_{ij}$ with $c(q) = \sum_{s \in R_l} w_s \mathbb{1}\big[z_s \neq x_s^{\mathrm{obs}}\big]$. Minimizing over $l \in \Delta$ and admissible z gives the middle bound of equation 3; the right-hand bound follows from $\mathbb{1}[\cdot] \leq 1$ and $w_s \leq w_{\max}$, and the left-hand bound is Proposition 6(c).

Complexity of the witness. $\varphi_l^{P_i}$ depends only on the coordinates in $\mathrm{Pa}_{M_i}(l)$, so $\varphi_l^{P_i}$ and $\varphi_l^{P_j}$ can be compared on the $2^{|R_l|} \leq 2^{2d}$ configurations of R_l by two table look-ups each; doing so for every l costs $O(n 2^{2d})$ and yields both Δ and a minimizing (l, z). □

A.14 Proof of Theorem 6

(i) A certificate is a query $q = (\mathrm{do}(X_S = a_S), T)$ with $S \subseteq A$; its cost and both answers are computable in polynomial time by evaluating the two acyclic Boolean networks.

(ii) *A the roots, unit weights.* We reduce from SAT. Let $\phi = C_1 \wedge \cdots \wedge C_m$ be a CNF over $z_1, \ldots, z_p$. If $\phi(0, \ldots, 0) = 1$ output a fixed yes-instance. Otherwise build an SCM with root variables $X_1, \ldots, X_p$ ($X_l = U_l$), a NOT gate for each negated literal, a chain of binary OR gates computing each clause, and a chain of binary AND gates computing ϕ, with output variable X_{out}; all gate mechanisms ignore their exogenous variable. The size is polynomial and every in-degree is at most 2. Perspective P_i uses these mechanisms; perspective P_j is identical except that $f_{\text{out}} \equiv 0$. Take $u_i = u_j = 0$, so x^{obs} has all roots 0 and $X_{\text{out}} = \phi(0) = 0$ under both perspectives: they are history-equivalent, and they agree on every gate under every intervention, so only queries with $\text{out} \in T$ can separate them. Let A be the set of roots and $w \equiv 1$. Under $\text{do}(X_S = a_S)$ with $S \subseteq A$ the roots take a value z with $z_s = a_s$ on S and $z_s = 0$ elsewhere, $A_{P_i} = \phi(z)$ at X_{out}, $A_{P_j} = 0$, and the cost is the number of $s \in S$ with $a_s = 1$, i.e. the Hamming weight of z. Hence $k_{\mathcal{Q}_A, c} \leq p$ if and only if ϕ is satisfiable.

$A = [n]$, general weights. When gates are manipulable, a query may clamp a gate at its observed value at zero cost, which can shortcut the circuit (for instance a satisfied clause gate can be frozen at 1 while the root that satisfied it is flipped), so a different reduction is needed. We reduce from VERTEX COVER. Given a graph $G = (V, E)$ with $|V| = p$ and a budget $r \leq p$, let $\phi = \bigwedge_{\{u,v\} \in E} (z_u \vee z_v)$, a *monotone* CNF, and build the circuit above without NOT gates. At $z = 0$ every clause gate and every AND gate equals 0, so x^{obs} is the all-zero state and the perspectives (P_j with $f_{\text{out}} \equiv 0$) are history-equivalent. Set $w_s = 1$ for roots and $w_s = p+1$ for gates. If $C \subseteq V$ is a vertex cover of size at most r, the query $\text{do}(X_C = 1)$ with target out costs $|C| \leq r$ and separates, since $\phi(\mathbb{1}_C) = 1$. Conversely, let q separate at cost at most $r \leq p$. No gate can be set to 1 (cost $p + 1$), so every intervened gate is clamped at its observed value 0. The circuit is monotone, and clamping a gate at 0 can only decrease the values of downstream gates relative to the unclamped circuit evaluated at the same root values; hence $A_{P_i}(q) = 1$ at X_{out} requires $\phi(z) = 1$ for the root values z realized by q, i.e. $\{v : z_v = 1\}$ is a vertex cover, and its size equals the cost of q, which is at most r. Thus $k_{\mathcal{Q}_{[n]}, c} \leq r$ if and only if G has a vertex cover of size at most r.

(iii) In the SAT construction with A the roots, $k < \infty$ iff ϕ is satisfiable. With $A = [n]$, $\mathcal{Q}_A = \mathcal{Q}_{\text{full}}$ and Theorem 5 decides finiteness by comparing induced mechanisms in polynomial time.

(iv) Enumerate all $q \in \mathcal{Q}$, evaluate both answers, and take the minimum cost over the separating ones. □

A.15 PROOF OF PROPOSITION 5

Under $f_j(x_{\text{Pa}(j)}, u_j) = g_j(x_{\text{Pa}(j)}) \oplus u_j$, the equation $x_j^{\text{obs}} = g_j(x_{\text{Pa}(j)}^{\text{obs}}) \oplus u_j$ has the unique solution $u_j = x_j^{\text{obs}} \oplus g_j(x_{\text{Pa}(j)}^{\text{obs}})$ for each j. With this context, x^{obs} satisfies all equations, and by uniqueness of solutions $x(M_g, u(g, x^{\text{obs}})) = x^{\text{obs}}$. Since the argument applies to every g, every structure is history-equivalent at x^{obs} for its abduced context. □

A.16 PROOF OF THEOREM 7

(a) If $u_i = u_j$ then $u_{i,E} = u_E^\star$ iff $u_{j,E} = u_E^\star$ for every E. The characterization of finiteness is Theorem 5.

(b) If $f_l^i = f_l^j$ and $u_{i,l} = u_{j,l}$ then $\varphi_l^{P_i} = \varphi_l^{P_j}$. Hence when $\Delta_F = \emptyset$, $\varphi_l^{P_i} \neq \varphi_l^{P_j}$ forces $l \in \Delta_u$, and the variable produced by Lemma 1 lies in Δ_u. If $u_i = u^\star$ and $E \cap \Delta_u \ni l$, then $u_{j,l} \neq u_{i,l} = u_l^\star$, so $P_j \notin \Pi_E$.

(c) For $P_i, P_j \in \Pi_{[n]}$ we have $u_i = u_j = u^\star$, so $\varphi_l^{P_i} = f_l^i(\cdot, u_l^\star)$ and $\varphi_l^{P_j} = f_l^j(\cdot, u_l^\star)$; Theorem 5 gives the equivalence. Every answer $A_P(q)$ of a perspective with context $u^\star$ depends on f_l only through $f_l(\cdot, u_l^\star)$, so differences at other values of u_l affect no observational state and no answer to any query. □

A.17 PROOF OF THEOREM 8

Let $X_l = U_l$ for $l \leq r$ and $X_{r+1} = (X_1 \vee \cdots \vee X_r) \vee U_{r+1}$, all variables Boolean, and $x^{\text{obs}} = (1, \ldots, 1)$. Let $u_i = (1, \ldots, 1, 0)$ and $u_j = (1, \ldots, 1, 1)$. Both contexts produce x^{obs}, so the

perspectives are history-equivalent, share all structural functions, and differ exactly in coordinate $l_0 = r + 1$.

(i) Since $\Delta_F = \emptyset$, by Theorem 7(b) any divergence occurs at X_{r+1}, so a separating query must have $r+1 \in T$ and $r+1 \notin S$. Under any intervention, $X_{r+1} = 1$ in both perspectives unless $X_1 = \cdots = X_r = 0$. The roots have no parents, so $X_l = 0$ for $l \le r$ requires $l \in S$ with $a_l = 0 \neq x_l^{\mathrm{obs}}$, at cost 1 each. Hence every separating query costs at least r, and $q = (\mathrm{do}(X_1 = \cdots = X_r = 0), \{r+1\})$ separates ($A_{P_i} = 0$, $A_{P_j} = 1$) at cost exactly r. Thus $k_{\mathcal{Q},c}(P_i, P_j) = r$.

(ii) A separating query needs all r roots in its domain, which a family with domain size at most $r-1$ does not admit; so $\mathcal{W}_{ij} = \emptyset$ and $k_{\mathcal{Q},c} = \infty$.

(iii) With $E = \{r+1\}$ and $u^\star = u_i$, $u_{j,r+1} = 1 \neq 0 = u^\star_{r+1}$, so $P_j \notin \Pi_E$. □

B Structural Results Used in the Main Text

Proposition 6 (Basic properties). *(a) $k_{\mathcal{Q},c}$ is symmetric and $k_{\mathcal{Q},c}(P,P) = \infty$; (b) $\mathcal{Q} \subseteq \mathcal{Q}'$ implies $k_{\mathcal{Q}',c} \le k_{\mathcal{Q},c}$, and $c \le c'$ pointwise implies $k_{\mathcal{Q},c} \le k_{\mathcal{Q},c'}$; (c) for history-equivalent P, P' and $c(q) = 0$, $A_P(q) = A_{P'}(q) = x_T^{\mathrm{obs}}$, hence $k_{\mathcal{Q},c}(P,P') \ge w_{\min}$; (d) $k_{\mathcal{Q},c}(P,P'') \ge \min\{k_{\mathcal{Q},c}(P,P'), k_{\mathcal{Q},c}(P',P'')\}$, so $\equiv^r_{\mathcal{Q},c}$ is an equivalence relation whose classes refine as r grows.*

Induced mechanisms. The interventional behavior of $P = (M,u)$ is determined by its induced mechanisms $\varphi_j^P(x) := f_j(x_{\mathrm{Pa}_M(j)}, u_j)$: under $I = \mathrm{do}(X_S = a_S)$, $x(M^I, u)$ is the unique solution of $x_s = a_s$ ($s \in S$) and $x_j = \varphi_j^P(x)$ ($j \notin S$).

Lemma 1 (First divergence). *Let $x^P, x^{P'}$ be the states of P, P' under an intervention with domain S. If $x^P \neq x^{P'}$, there is $l \notin S$ with $x_l^P \neq x_l^{P'}$ and $\varphi_l^P(x^{P'}) \neq \varphi_l^{P'}(x^{P'})$.*

Corollary 1 (A sufficient condition for survival). *If φ^P and $\varphi^{P^\star}$ agree on every state that some query in $\mathcal{Q}_R$ realizes, then P survives every run of Algorithm 1 to radius R, whatever the query rule. The condition is sufficient, not necessary: P may differ from $P^\star$ on a cell that a query in $\mathcal{Q}_R$ does realize and still survive, because the difference can be* masked*—it changes a variable that does not, for that query, propagate to the target set T (Lemma 1 locates a divergence at some variable, not necessarily in T). The terminal class is therefore $\{P : A_P(q) = A_{P^\star}(q)\ \forall q \in \mathcal{Q}_R\}$, which contains the hypotheses differing only on cells that no query of cost at most R realizes together with those whose reachable differences are masked at T; the risk of $\hat{P}$ beyond radius R is carried by both.*

Theorem 5 (Finite separation under the full family). *Let P, P' be history-equivalent at x^{obs}, $\Delta := \{l : \varphi_l^P \neq \varphi_l^{P'}\}$ and $R_l := \mathrm{Pa}_M(l) \cup \mathrm{Pa}_{M'}(l)$. Then $k_{\mathcal{Q}_{\mathrm{full}},c}(P,P') < \infty$ iff $\Delta \neq \emptyset$, and then*

$$w_{\min} \;\le\; k_{\mathcal{Q}_{\mathrm{full}},c}(P,P') \;\le\; \min_{l \in \Delta} \min_{z:\, \varphi_l^P(z) \neq \varphi_l^{P'}(z)} \sum_{s \in R_l} w_s \mathbb{1}\big[z_s \neq x_s^{\mathrm{obs}}\big] \;\le\; w_{\max} \min_{l \in \Delta} |R_l|. \quad (3)$$

With truth tables of in-degree at most d, Δ and a witness attaining the middle bound are computable in $O(n 2^{2d})$ time. If $\varphi^P = \varphi^{P'}$ then $k_{\mathcal{Q},c}(P,P') = \infty$ for every *$\mathcal{Q}$ and c.*

The middle bound clamps the parents of a differing mechanism to a configuration on which the mechanisms differ; it can be loose, because an upstream intervention may steer several parents at once more cheaply. For restricted families $\mathcal{Q} \subsetneq \mathcal{Q}_{\mathrm{full}}$ separation can be infinite even when $\Delta \neq \emptyset$; that is when the terminal class of Corollary 1 is non-trivial, which is why the experiments compare a visual and a white-box intervention family.

Theorem 6 (Complexity). *Let $\mathcal{Q}_A$ allow any intervention on a manipulable set $A \subseteq [n]$ with any target. Deciding $k_{\mathcal{Q}_A,c}(P,P') \le r$ for Boolean SCMs of in-degree at most 2 is NP-complete: it is in NP; it is NP-hard even with unit weights, a single target and A the roots, and remains NP-hard for $A = [n]$ under general weights. Deciding $k_{\mathcal{Q}_A,c} < \infty$ is NP-hard for A the roots but polynomial for $A = [n]$ by Theorem 5. If $|\mathcal{Q}|$ is polynomial in n (interventions of size at most a constant s) the problem is in P by enumeration.*

The contrast is informative: when everything is manipulable, *whether* two hypotheses can be separated is easy but *how cheaply* is hard, because a cheap witness may have to steer several mechanisms

through few upstream facts. Cost-ordered enumeration with a size bound s costs $O((2n)^s n)$ mechanism evaluations, vectorizable over a whole sample of hypotheses in one pass.

Grounding. For history-equivalent P, P' let $\Delta_F := \{l : f_l^P \neq f_l^{P'}\}$ and $\Delta_u := \{l : u_{P,l} \neq u_{P',l}\}$.

Proposition 7 (Monotonicity)**.** *If $E \subseteq E'$ then $V_0^{E'} \subseteq V_0^E$ and $\kappa(V_0^{E'}) \geq \kappa(V_0^E)$; the cost bound of Theorem 3(a) and the certified-radius condition can only improve.*

Theorem 7 (Complementarity)**.** *(a) If $\Delta_u = \emptyset$ then $P \in V_0^E \iff P' \in V_0^E$ for every E. (b) If $\Delta_F = \emptyset$ then every divergence in Lemma 1 occurs at some $l \in \Delta_u$, and if P is fully grounded any E meeting Δ_u eliminates P' at zero cost. (c) For fully grounded P, P', $k_{\mathcal{Q}_{\text{full}},c} < \infty$ iff $f_l^P(\cdot, u_l^\star) \neq f_l^{P'}(\cdot, u_l^\star)$ for some l: differences at other context values are neither observable nor testable.*

Theorem 8 (Evidence can be arbitrarily cheaper than intervention)**.** *For every $r \geq 1$ there are an SCM on $r+1$ variables, a history x^{obs}, and history-equivalent P, P' with identical structural functions such that (i) $k_{\mathcal{Q},c}(P, P') = r$ under unit weights and $\mathcal{Q}_{\text{full}}$; (ii) $k_{\mathcal{Q},c}(P, P') = \infty$ for any $\mathcal{Q}$ whose interventions have domain size at most $r-1$; (iii) $u_P, u_{P'}$ differ in one coordinate l_0, so the single verified fact $E = \{l_0\}$ eliminates P' at zero cost.*

Roots $X_1, \ldots, X_r$ equal their exogenous variables, $X_{r+1} = (X_1 \vee \cdots \vee X_r) \vee U_{r+1}$, the history is all ones, and the two hypotheses differ only in U_{r+1}, which the observed causes mask until all r of them are clamped to 0.

C Details of the Continuous MNIST Study

This appendix completes Section 6. The code is `code/mnist/cont.py` (networks), `audit_cont.py` (the auditor) and `floor_cnn.py cont` (the plain CNNs). Every number is generated from the result files into `macros/` by `analyze_cont.py`.

C.1 Networks

Rendering. A digit of hue $c \in [0, 1]$ puts its gray levels, scaled by c, in the red channel and, scaled by $1 - c$, in the green channel of a black RGB image; $c = 1$ is pure red, $c = 0$ pure green, and $c = 0.5$ an olive mix (Figure 2a). In observed images a large digit has a hue drawn uniformly from $[0.75, 1]$ and a small digit from $[0, 0.25]$.

Architecture. Each slot has the encoder of Appendix D.2, whose output unit u_s gives the percept $Z_s = \tanh(u_s)$. The head is a top-1 router: it outputs $\gamma Z_r + \beta$ if $Z_p > 0$ and $\gamma Z_l + \beta$ otherwise, with learned γ, β. The routing is hard in the forward pass and uses the gradient of $\sigma(4Z_p)$ in the backward pass. In a bypass network the router reads the sign of a second encoder of the pointer image instead of Z_p.

Training. Adam with learning rate 10^{-3} and a cosine schedule, 4,000 steps of 256 fresh images; the label follows the shape world.

- *Trained* (α): each digit's hue is drawn from its size's observed range with probability α and uniformly otherwise. The loss is the task loss plus a concept loss training each u_s to report its digit's size.
- *Planted*: each slot s gets an angle ϕ_s drawn uniformly from $[0, \pi/2]$, and its percept is trained by squared error towards $\tanh\big(3(\cos\phi_s(2D_s - 1) + \sin\phi_s(2c_s - 1))\big)$, which reads shape at $\phi_s = 0$, hue at $\phi_s = \pi/2$, and a continuous mix in between. Hues are uniform, and the label is routed by the planted percepts.
- *Bypass*: value percepts as in the $\alpha = 1$ networks. The pointer's percept is trained on uniform hues to report whether the pointer's hue is at least 0.5, and the router's encoder is trained on the task with observed hues. Both read the pointer's hue and agree on every observed image, so the log fits the class exactly, yet the router never reads the percept.

C.2 The auditor

Log and version space. For each network the auditor records 3,000 observed images with their percepts and outputs. It fits γ and β by least squares on the logit, which reproduces every logged

output of every network, the bypass networks included, to within 10^{-4} in the logit. For each slot, size and hue within the observed range, it bounds the percept by the 0.5% and 99.5% quantiles of the logged percepts in the hue bin of width 0.05 that contains that hue, widened by 0.02. The version space V_0 is the box that these bounds cut out of $[-1, 1]^{72}$, with the observed digit's percept at its observed hue fixed to its observed value. A hypothesis is a point of this box. An observed image is used only if its own percepts lie within their bounds.

When the truth is outside the class. The bounds are estimated from other digits, so a particular digit's percept can fall outside them, for instance that of the held-out digit a swap puts in, which the log never saw. Up to a floating-point error of 10^{-5}, the true percepts lie in the box at 96% of the 150 observed images of the in-class networks. At the others the class is misspecified and Proposition 4 does not apply; 12% of the certificates issued there are void.

Queries and answers. A visual query sets each slot's hue to one of the eleven grid hues or keeps it, and may swap the digit for one fixed held-out digit of the other size; its cost is the total change in hue plus one per swap. The white-box family may also set the pointer's percept to -1, -0.5, 0, 0.5 or 1, at a cost of half the change. The answer is the network's output probability, computed exactly.

Exact separation. Write θ for the hypothesis. A query's selected percept is a coordinate of θ or a patched constant, chosen by the sign of the pointer's percept. An answer y constrains the selected percept to the interval that y allows, with $\eta = 10^{-4}$ in probability. When the pointer's sign is already determined by the current bounds, this is a bound on one coordinate. Otherwise it is a disjunction encoded with one binary variable and big-M constraints. Since the network routes right exactly when $Z_p > 0$, routing right requires $Z_p \geq \mu$ with $\mu = 10^{-3}$, far above the solver's feasibility tolerance; the true pointer percepts of the in-class networks never fall in $(0, \mu)$. The largest and smallest output a query can have over the version space are two mixed-integer linear programs, solved exactly with HiGHS. A cheap interval bound, computed from exact per-coordinate projections, skips queries it already proves narrower than $\varepsilon = 0.05$, and a query once proven narrow is never tested again, since the version space only shrinks.

Rules and checks. The rules are those of Appendix D.3, restricted to minimum cost, random separating and random probing. A certificate is *correct* when every survivor is within ε of the network on every admissible query within the radius; all of these queries are run for the check. Validation draws up to 20 unexecuted admissible queries within the radius and refutes at the first answer farther than ε from some survivor.

C.3 Results

Table 2 reports every network family, intervention family and radius. Wherever the truth lies in the version space, no certificate fails: 0 errors in 1,728 runs. Table 3 gives the plain CNNs.

Compute. On one CPU core a minimum-cost run takes 0.3 s on average (at most 7 s) and solves 223 MILPs. Random probing executes far more queries and piles up disjunctions: it averages 5.9 s and takes up to 263 s. Networks are audited in parallel, one process each; the whole continuous study, including training, takes under an hour on one RTX 4080 Super GPU and 15 CPU cores.

Table 2: The continuous study, all networks, families and radii (means over observed images). $|\mathcal{Q}_R|$ is the number of admissible interventions; ISS gives the number of interventions and their total cost; the next two columns give the cost of random separating queries and of random probing; then the fraction of minimum-cost certificates that are correct and void, and the fraction of random-probing runs that refute the class.

				ISS		random	random			probing
Networks	Family	R	$\lvert\mathcal{Q}_R\rvert$	int.	cost	sep.	probing	correct	void	refutes
Trained, $\alpha = 1$	visual	0.5	210	7.6	2.47	3.27	34.4	100%	0%	0%
		1.0	752	24.1	14.87	18.52	332.8	100%	0%	0%
	white-box	0.5	486	7.6	2.47	3.14	80.3	100%	0%	0%
		1.0	2224	24.2	14.92	17.82	839.8	100%	0%	0%
Trained, $\alpha = 0.9$	visual	0.5	221	13.4	4.00	5.34	43.1	100%	0%	0%
		1.0	785	38.9	23.63	31.14	375.4	100%	0%	0%
	white-box	0.5	510	13.4	4.00	5.50	82.4	100%	0%	0%
		1.0	2318	36.9	21.70	25.57	1060.3	100%	0%	0%
Trained, $\alpha = 0.5$	visual	0.5	206	21.1	6.10	8.33	38.0	100%	0%	0%
		1.0	742	58.8	35.16	46.95	337.3	100%	0%	0%
	white-box	0.5	478	21.1	6.12	8.05	78.6	100%	0%	0%
		1.0	2185	54.3	31.18	33.98	1058.0	100%	0%	0%
Planted	visual	0.5	210	13.0	3.68	4.70	27.1	98%	2%	3%
		1.0	753	33.2	19.06	22.94	310.1	98%	2%	3%
	white-box	0.5	487	13.2	3.76	5.13	60.8	98%	2%	3%
		1.0	2280	29.8	16.16	21.46	942.9	98%	2%	3%
Bypass	visual	0.5	210	7.9	2.72	3.33	35.4	100%	0%	0%
		1.0	753	25.2	15.83	18.74	287.1	100%	0%	0%
	white-box	0.5	483	7.9	2.72	3.42	52.6	63%	37%	23%
		1.0	2208	25.3	15.88	18.33	47.4	3%	97%	97%
All in class	visual	0.5	212	13.6	3.99	5.27	33.9	99%	1%	1%
		1.0	757	37.6	22.35	28.50	333.1	99%	1%	1%
	white-box	0.5	490	13.7	4.02	5.39	72.6	99%	1%	1%
		1.0	2258	35.0	20.02	24.05	968.8	99%	1%	1%

Table 3: Plain CNNs trained on N images with continuous hue (mean of three seeds): held-out error on observed images, and error against the shape world on every single-slot edit of cost at most 1 (a recolor to any hue of the grid, or a swap). The shape and color worlds disagree on 26.6% of these edits.

	$\alpha = 1$		$\alpha = 0.99$		$\alpha = 0.9$	
N	held-out	edits	held-out	edits	held-out	edits
100	5.0%	27.4%	6.3%	27.8%	14.3%	28.3%
300	0.0%	26.4%	0.3%	26.3%	3.9%	25.5%
1,000	0.0%	26.4%	0.1%	26.1%	1.4%	22.6%
3,000	0.0%	26.3%	0.0%	25.5%	1.1%	16.1%
10,000	0.0%	26.4%	0.1%	23.7%	0.9%	8.7%
30,000	0.0%	26.3%	0.1%	18.2%	0.6%	4.5%
60,000	0.0%	26.5%	0.1%	13.3%	0.4%	3.2%

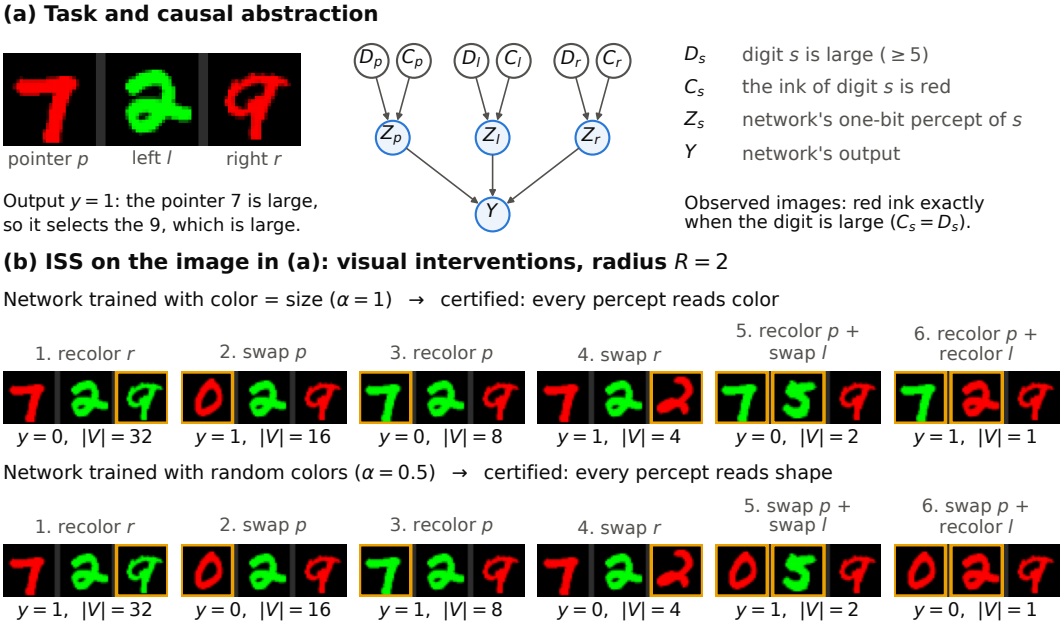

Figure 4: ISS on MNIST. (a) An observed image and the high-level causal model; image edits set D_s and C_s, and the percepts Z_s are the network's one-bit units. (b) Two networks whose logs fit the same cells, audited on this image: the intervention ISS chose at each step (edited digits framed), the network's answer y and the number $|V|$ of surviving candidates. Both certificates hold on all 21 interventions of cost at most 2, checked on the networks.

D THE BOOLEAN MNIST STUDY

This appendix reports the Boolean special case of Section 6. The ink is pure red or pure green, so each digit's color is a Boolean variable C_s (the ink is red), and each percept is a single binary unit. The version space is then finite, every certificate is exact with no tolerance, and every claim of Theorems 2 and 3 can be checked on every run. The code is in `code/mnist`, and every number is generated from its result files into `macros/` by `make mnist-tables`.

D.1 SETTING AND RESULTS

Task. Each image shows three handwritten MNIST digits (LeCun et al., 1998) side by side, each drawn in red or green ink (Figure 4a): a pointer p, a left digit l and a right digit r. Call a digit *large* if it is at least 5. A small pointer selects the left digit and a large pointer the right one, and the label says whether the selected digit is large. In Figure 4a the pointer 7 is large, so it selects the 9, which is large: the label is 1. The task is a Boolean version of MNIST pointer-value retrieval (Zhang et al., 2021), a benchmark for causal abstraction (Geiger et al., 2022).

Causal abstraction. A causal abstraction explains a network by a small causal model whose variables are aligned with parts of the network (Geiger et al., 2021). Ours is drawn in Figure 4a. For each digit $s \in \{p, l, r\}$ it has two input variables, D_s (the digit is large) and C_s (its ink is red), and a percept $Z_s = g_s(D_s, C_s)$, which is what the network believes about the digit. The output is $Y = h(Z_p, Z_l, Z_r)$. The audited networks make the alignment exact. Each digit is read by its own small CNN that ends in a single binary unit, which plays the role of Z_s, and an MLP computes the output from the three bits: a concept-bottleneck network (Koh et al., 2020). What remains open is the mechanisms. The intended abstraction reads shape, $g_s(D, C) = D$, and routes with the pointer (h is a multiplexer). But a percept could as well read the ink, $g_s(D, C) = C$, or need both (*and*) or either (*or*).

Why observation cannot decide. The auditor observes the network only on images in which the ink is red exactly when the digit is large, the shortcut of Colored MNIST (Arjovsky et al., 2019).

On such images shape and ink always agree, so a percept that reads shape and one that reads ink produce the same bits and the same outputs. A log of $5{,}000$ observed images, with the network's bits and outputs, fixes h and the two cells of each g_s that occur: small green digits and large red ones. The other two cells, a small red digit and a large green one, never occur, however large the log. They are the permanently unobserved cells of Proposition 2, $m_\infty = 6$ in all, so the version space V_0 holds $2^6 = 64$ candidate abstractions: one of the four tables for each digit (Table 5).

Interventions and certificates. A *visual* intervention edits the image: it swaps a digit for a held-out digit of the other size, $\mathrm{do}(D_s)$, or recolors it, $\mathrm{do}(C_s)$. An *interchange* intervention (Geiger et al., 2021) acts inside the network: it overwrites a percept bit with the value it takes on another input, $\mathrm{do}(Z_s)$. The visual family allows image edits only; the white-box family allows both kinds. Each changed variable costs 1, so the radius R is the number of variables an intervention may change. For example, $R = 2$ covers recoloring the pointer while swapping the left digit. The answer is the network's output, by majority over five renderings of each swap with different held-out digits. ISS stops at radius R once all surviving candidates predict the same output for every intervention of cost at most R. By Theorem 2, the survivors then predict the network's output on all of these interventions, provided its abstraction is among the 64 candidates. The binary bottleneck lets us check this claim directly, by running every admissible intervention on the network.

Networks. We audit 50 networks at 50 observed images each:

- *trained* (20): trained on the task with ink agreeing with size with probability $\alpha \in \{1, 0.99, 0.9, 0.5\}$, five seeds each; at $\alpha = 1$ the training data contain the shortcut, and at $\alpha = 0.5$ the ink is random;
- *planted* (20): each percept is trained to one of the four tables at random, so the true abstraction is known in advance;
- *bypass* (10): the pointer's bit is trained to report the pointer's size, but the output ignores that bit and reads the pointer through a second, separate encoder, so that no candidate describes the network.

Result 1: observation cannot reveal the abstraction. Held-out accuracy on the observed kind of image cannot tell what a network reads. Every network in the class scores at least 97.6%, and the color readers ($\alpha = 1$) score highest, 100.0%, because the ink never misleads them. Yet the intended abstraction mispredicts 52% of their interventions of cost at most 2 (Table 4). Theorem 1 makes this a limit of observation itself: whatever abstraction an observational auditor reports, it mispredicts at least 36% of these interventions for some network consistent with its log, averaged over observed images, however large the log.

The same limit binds ordinary learners. Call the intended labeling rule, which reads digit sizes, the *shape world*, and the same rule applied to ink colors the *color world*. The two worlds label every observed image identically. Plain CNNs trained on up to $60{,}000$ observed images reach zero held-out error, yet their outputs on single-variable interventions disagree with the shape world 50% of the time (Figure 5a): they learned the color world. When one digit in ten carries the other ink ($\alpha = 0.9$), the mismatched cells do occur in the data, and the error falls with N to 3%. These are the finite-sample cells of Proposition 2, which more data does reveal.

Result 2: certificates are correct and cheap. Figure 4b shows ISS on one image for two networks whose logs fit the same cells. At each step ISS runs the cheapest intervention on which the surviving candidates disagree, and keeps the candidates that predicted the network's answer. The two networks answer each of the first four interventions oppositely. From then on ISS steers them differently: to reach the hidden left digit, it recolors the pointer of the color reader and swaps the pointer of the shape reader, because by then it knows how each network reads the pointer. After six interventions a single candidate survives for each network, and it is the right one.

The network's abstraction lies in the class at every observed image, except at 8% of those of the $\alpha = 0.99$ networks, whose percepts of mismatched digits vary from one handwriting to another. Wherever it lies in the class, every run of every acquisition rule ends with a certificate, and every certificate agrees with the network on every intervention within its radius: 53,496 runs, without exception (Table 8).

Certification is also cheap. At $R = 2$ with visual interventions, ISS never needs more than 6 interventions, which is $\lceil \log_2 C \rceil$, the fewest any learner can guarantee (Theorem 3(b)). Together they

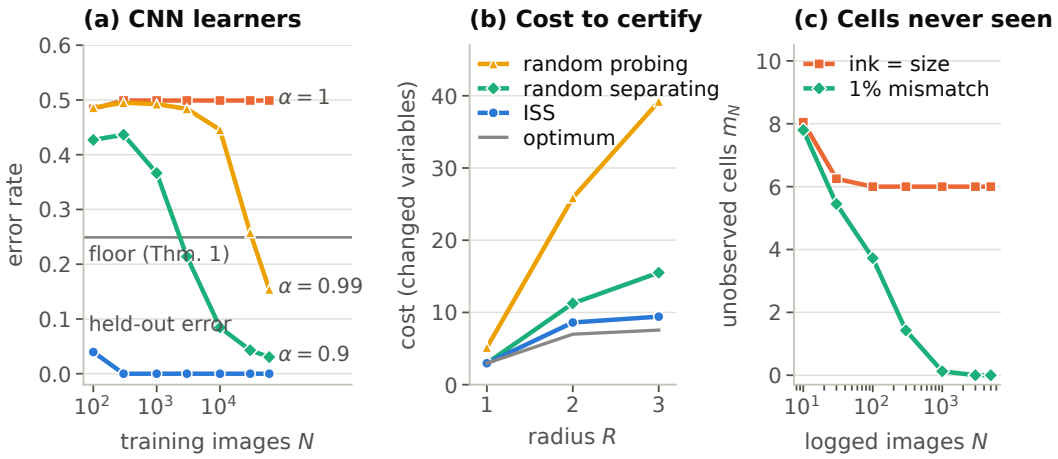

Figure 5: (a) Plain CNNs trained on N observed images: held-out error, and error on single-variable interventions against the shape world, for three values of α; the gray line is the floor of Theorem 1 for the $\alpha = 1$ regime; under $\alpha < 1$ the missing cells occur in the data and the floor vanishes with N. (b) Variables changed to certify radius R with visual interventions, over the 40 networks in the class. (c) Mechanism cells never seen in the auditor's log after N images.

Table 4: Auditing 50 networks at $R = 2$ (visual interventions unless marked; means over 50 observed images). *In class*: images at which one of the 64 candidates matches the network. *Intended risk*: fraction of interventions of cost at most 2 that the intended abstraction mispredicts. *Cost to certify*: variables changed by ISS, by a learner told the truth (optimum), and by random probing. *Correct*: certificates that hold on every intervention within the radius.

	Held-out		In	Intended	Cost to certify			
Networks (n)	accuracy	Reads	class	risk	ISS	optimum	random	Correct
Trained, $\alpha = 1$ (5)	100.0%	color	100%	0.52	9.2	7.0	26.2	100%
Trained, $\alpha = 0.99$ (5)	99.7%	shape (84%)	92%	0.06	9.0	7.1	26.5	99%
Trained, $\alpha \leq 0.9$ (10)	98.6%	shape (98%)	100%	0.01	9.0	7.0	26.4	100%
Planted (20)	99.0%	as planted (98%)	100%	0.33	8.2	6.9	25.3	100%
Bypass (10)	100.0%	color	100%	0.52	9.0	7.0	26.4	100%
white-box			0%	0.36	9.0	–	19.1	0%

change 8.6 variables on average. A learner told the true abstraction still needs 7.0, the non-adaptive greedy cover of Theorem 3(c) needs 15.1, and random probing with the same stopping test needs 25.9 (Figure 5b).

Result 3: a wrong class is caught only by validation. The bypass networks separate behavior from abstraction. With visual interventions, ISS certifies that the pointer is read by color. As a claim about behavior this is correct: the network answers every image edit within the radius as predicted. As an abstraction it is wrong. The pointer's bit reads shape (it reports the size of 99.0% of mismatched digits correctly), so a probe trained on that bit would endorse the intended abstraction (Elazar et al., 2021); but the output never uses the bit.

Interchange interventions expose the difference: flipping the pointer's bit changes nothing, while every candidate predicts that it reroutes the output. No candidate matches these networks, yet ISS still halts with agreement in every white-box run at $R = 2$, and every such certificate is void. The reason is structural. ISS only asks about interventions on which the candidates disagree, and they all agree about the pointer's bit, so it never asks the question that would expose the bypass (Figure 6). A validation phase of twenty random interventions refutes 93% of these certificates, and random probing, which does not look for disagreement, refutes the class during learning in 97% of runs (Table 10). The $\alpha = 0.99$ networks are a milder case: 1% of their certificates at $R = 2$ are void, and validation refutes 100% of those.

Table 5: The four completions of a percept's table. The log fixes the two diagonal cells; the two off-diagonal cells are never observed.

Percept reads	small, green $(D, C) = (0, 0)$	small, red $(0, 1)$	large, green $(1, 0)$	large, red $(1, 1)$
shape	0	0	1	1
color	0	1	0	1
and (both)	0	0	0	1
or (either)	0	1	1	1

Table 6: The audited networks. Held-out accuracy is on the observed regime (ink matches size) against the shape world's label, as a mean with the minimum in parentheses; percept accuracy is the agreement of the bits with the digits' sizes there; $\hat{p}$ is the fraction of logged percepts that disagree with the fitted cells. *Reads* summarises the truth's tables over all observed images, and the last two columns give the fraction of observed images at which some member of V_0 matches the network on every query of the visual and of the white-box family.

Networks	n	Held-out accuracy	Percept accuracy	$\hat{p}$	Reads	Truth in V_0 visual	Truth in V_0 white-box
Trained, $\alpha = 1$	5	100.0% (100.0%)	100.0%	0.00%	color	100%	100%
Trained, $\alpha = 0.99$	5	99.7% (99.7%)	99.8%	0.18%	shape (84%)	92%	92%
Trained, $\alpha \leq 0.9$	10	98.6% (97.6%)	99.1%	0.82%	shape (98%)	100%	100%
Planted	20	99.0% (98.0%)	99.3%	0.65%	as planted (98%)	100%	100%
Bypass	10	100.0% (100.0%)	99.6%	0.38%	color	100%	0%

D.2 Task, data and networks

Data. Networks are trained on the $60{,}000$ MNIST training digits. The $10{,}000$ test digits are split at random into two disjoint pools of $5{,}000$: the *observational pool* supplies the auditor's log and the observed images, and the *exemplar pool* supplies the digits that realize interventions. A digit is drawn in red or green by placing its gray levels in the red or the green channel of a black RGB image. An image has three slots, pointer p, left l and right r.

Architecture. Each slot has its own encoder: two 5×5 convolutions with 16 and 32 channels, each followed by ReLU and 2×2 max pooling, a 64-unit ReLU layer and a linear unit whose sign is the percept bit Z_s. The forward pass uses the hard bit, and the backward pass uses the gradient of the sigmoid, as in the straight-through estimator (Bengio et al., 2013). The head is an MLP with two hidden layers of 64 ReLU units that reads the three bits as ± 1. In a bypass network the pointer has a second encoder of the same shape, whose 64 features pass through a 4-unit $\tanh$ layer. The head reads these four numbers and the two value bits, and never the pointer bit.

Training. All networks use Adam with learning rate 10^{-3} and a cosine schedule, for $4{,}000$ steps of 256 freshly drawn images. The label follows the shape world, $Y = \mathrm{MUX}(D_p; D_l, D_r)$, except for planted networks.

- *Trained* networks (α): ink agrees with size independently in each slot with probability α. The loss is the task loss plus a concept loss that trains each bit to report its digit's size, as in joint concept-bottleneck training (Koh et al., 2020). With $\alpha = 1$ size and ink coincide on every training image, so this loss cannot favor shape over ink.
- *Planted* networks: each slot is assigned one of the four tables of Table 5 at random (20 distinct assignments), the images have random ink, each bit is trained to its planted table and the label is the multiplexer of the planted percepts.
- *Bypass* networks: the value bits are trained as in the $\alpha = 1$ networks. The pointer bit is trained to report the pointer's size on a separate batch with random ink, so it reads shape. The head is trained on the task with ink matching size.

D.3 THE AUDIT

Log and version space. For each network the auditor draws $5{,}000$ images from the observational pool with ink matching size, records the concepts, the three bits and the output, and fits every cell that occurs by majority (Proposition 2 with $p > 0$). The fitted head is the multiplexer for every network. The fraction of logged percepts that disagree with the fit, the perception noise $\hat{p}$, averages 0.56%. The soft certificate of Theorem 4 costs nothing here. At $\delta = 0.01$ and $\hat{p}$, the probability that any fitted cell of a variable is wrong is below δ/n' already at a budget of $T_j = 0$, so no fitted cell may be flipped and $V_0^\delta = V_0$. The six off-diagonal cells are never realized, and V_0 has 64 members for every network. Figure 5(c) shows the unobserved cells as N grows. When 1% of the ink is mismatched instead, every cell is realized within $3{,}000$ images in every log.

Observed images. For each network we draw observed images from the observational pool, with ink matching size, and keep the first 50 whose percepts agree with the fitted cells, so that the context of the observed state has no perception noise. About 1.0% of the drawn images are skipped.

Queries and answers. A query sets a subset of the roots to values other than the observed ones and, in the white-box family, sets bits: a bit may be flipped (cost 1) or clamped at its observed value (cost 0, allowed only in a slot whose digit is edited, since elsewhere it changes nothing). The visual family has 6, 15 and 20 queries of cost 1, 2 and 3; the white-box family has 15, 87 and 248. For each observed image we draw a bank of five exemplar digits of each size for each slot. Rendering i of a query replaces every swapped digit by the i-th exemplar of the required size and ink, sets the patched bits, and runs the network. The answer is the majority of the five outputs. The answer to every query is thus a fixed function of the query, and the truth is the member of V_0 that matches the network on every query of the family, when there is one.

Acquisition rules. All rules share Algorithm 1's elimination and stopping test and differ only in the query they pick among the separating queries within the radius:

- *minimum cost* (the default): a cheapest separating query; ties go to the most balanced split of the survivors, then at random;
- *information per cost*: the query maximizing the entropy of the survivors' split per unit cost, the acquisition principle of Bayesian experimental design (Chaloner & Verdinelli, 1995);
- *random separating*: a separating query uniformly at random;
- *random probing*: an admissible query within the radius uniformly at random, separating or not, without replacement, until the stopping test holds.

Checks. A certificate is *correct* when every survivor answers every admissible query within the radius as the network does; all of these queries are run for the check, not only those the learner executed. The optimum of a learner told the truth is the cheapest set of queries that separates the truth from every hypothesis within separation cost R of it, a weighted set cover solved exactly as a 0-1 integer program with HiGHS; any correct learner pays at least this much (proof of Theorem 3(b)). The greedy cover of Theorem 3(c) and its optimum are computed in the same way. Validation draws up to 20 admissible queries within the radius that the run did not execute and refutes the class at the first answer that contradicts the survivors.

D.4 ADDITIONAL RESULTS

Rules and radii. Table 7 reports every rule at every radius. The two rules that look for disagreement, minimum cost and information per cost, pay the same. Random separating queries cost more, and random probing far more, most of all in the white-box family, where most admissible queries separate nothing. Correctness does not depend on the rule: wherever the truth lies in the class, every rule's certificate is correct (Table 8).

Planted abstractions. The certificates recover the planted tables at 98% of observed images, and at 99.5% of slots. At the other images the network departs from its planted table on some exemplar digit of the bank, and the certificate follows the network. This is the intended behavior: ISS certifies the network, not its training target.

Theorem checks. No claim of Theorems 2 and 3 fails on any run whose truth lies in the class (Table 8). Adaptivity usually pays: the non-adaptive greedy cover costs more than the adaptive minimum-cost run in 4,887 of 15,000 instances, and less in 15.

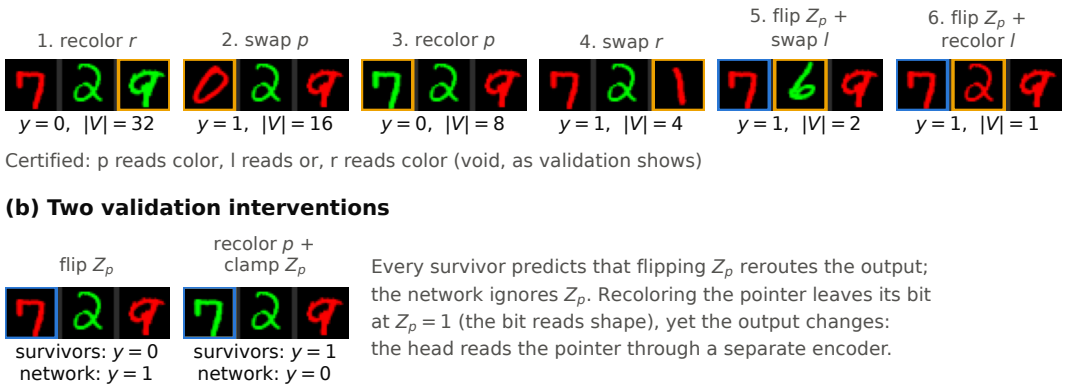

Figure 6: A bypass network audited with visual and interchange interventions on the image of Figure 4. Yellow frames mark edited digits and blue frames patched bits. (a) Steps 1–4 are those of the color reader in Figure 4; at steps 5 and 6 the flipped pointer bit changes nothing, and the answers are misread as facts about the left digit. (b) Two admissible interventions within the radius on which every survivor is wrong.

Bypass networks. Table 10 gives the bypass results at every radius. At $R = 1$ about half of the white-box certificates are still correct. These are the observed images whose two value digits have the same size, where rerouting the pointer changes nothing within one edit. Figure 6 shows a white-box run on the image of Figure 4. To reach the left digit, ISS flips the pointer's bit instead of editing the pointer. The network ignores the flip, so ISS reads its unchanged answers as facts about the left digit and certifies a wrong table for it. Both validation interventions contradict every survivor.

A larger version space. In a four-way pointer, two pointer digits select one of four value digits. The version space then has $2^{12} = 4{,}096$ members, and a value digit is hidden unless both pointer percepts route to it, so separating queries need up to three edits. We audit 20 networks (five trained with $\alpha = 1$, five with $\alpha = 0.5$ and ten planted) at 25 observed images each (Table 11). The truth lies in the class at 100% of observed images, and no certificate fails in 12,000 runs. The gap to random probing widens with the masking: at $R = 3$ it pays more than an order of magnitude more than ISS.

Compute. Training one network takes under a minute on one GPU. The audit takes about 5 s per network for the three-digit task and about 80 s for the four-way pointer, most of it in the exhaustive checks and the exact set-cover optima. The whole experiment, including the floor study, runs in about one and a half hours on a single consumer GPU.

Table 7: Certification at radius R, averaged over observed images. *ISS* is the minimum-cost rule (number of interventions and total cost); the next three columns give the total cost of the other rules; *opt.* is the optimum of a learner told the truth, $k^\star$ the truth-specific bound of Theorem 3(b) and $\lceil \log_2 C \rceil$ the query bound (both averaged; blank where the truth is not in V_0). The last two columns give the fraction of correct certificates for the minimum-cost rule and for the worst of the four rules.

		ISS		info	random	random				correct
Networks	R	int.	cost	/cost	sep.	probing	opt.	$k^\star$	$\lceil \log_2 C \rceil$	ISS / worst
Visual interventions										
Trained, $\alpha = 1$	1	2.8	2.8	2.8	2.8	5.0	2.8	1.0	2.8	100% / 100%
	2	6.0	9.2	9.2	11.4	26.2	7.0	2.0	6.0	100% / 100%
	3	6.0	9.2	9.2	15.5	35.5	7.0	2.0	6.0	100% / 100%
Trained, $\alpha = 0.99$	1	2.9	2.9	2.9	2.9	5.0	2.9	1.0	2.9	100% / 100%
	2	6.0	9.0	9.0	11.3	26.5	7.1	2.0	6.0	99% / 99%
	3	6.0	9.1	9.1	15.3	36.2	7.1	2.0	6.0	92% / 92%
Trained, $\alpha \leq 0.9$	1	3.0	3.0	3.0	3.0	5.2	3.0	1.0	3.0	100% / 100%
	2	6.0	9.0	9.0	11.4	26.4	7.0	2.0	6.0	100% / 100%
	3	6.0	9.0	9.0	14.8	35.3	7.0	2.0	6.0	100% / 100%
Planted	1	3.0	3.0	3.0	3.0	5.1	3.0	1.0	3.0	100% / 100%
	2	5.6	8.2	8.2	11.1	25.3	6.9	1.9	6.0	100% / 100%
	3	6.1	9.8	9.7	15.9	42.9	8.1	2.2	6.0	100% / 100%
Bypass	1	3.0	3.0	3.0	3.0	5.2	3.0	1.0	3.0	100% / 100%
	2	6.0	9.0	9.0	11.4	26.4	7.0	2.0	6.0	100% / 100%
	3	6.0	9.0	9.0	15.0	35.1	7.0	2.0	6.0	100% / 100%
All in class	1	3.0	3.0	3.0	3.0	5.1	3.0	1.0	3.0	100% / 100%
	2	5.8	8.6	8.6	11.3	25.9	7.0	2.0	6.0	100% / 100%
	3	6.1	9.4	9.4	15.5	39.2	7.6	2.1	6.0	99% / 99%
Visual and interchange interventions (white-box)										
Trained, $\alpha = 1$	1	2.8	2.8	2.8	2.8	11.3	2.8	1.0	2.8	100% / 100%
	2	6.0	9.2	9.2	12.1	97.7	7.0	2.0	6.0	100% / 100%
	3	6.0	9.2	9.2	16.9	107.7	7.0	2.0	6.0	100% / 100%
Trained, $\alpha = 0.99$	1	2.9	2.9	2.9	2.9	11.8	2.9	1.0	2.9	100% / 100%
	2	6.0	9.1	9.1	11.9	91.7	7.2	2.0	6.0	95% / 95%
	3	6.0	9.1	9.1	16.9	107.0	7.1	2.0	6.0	92% / 92%
Trained, $\alpha \leq 0.9$	1	3.0	3.0	3.0	3.0	11.7	3.0	1.0	3.0	100% / 100%
	2	6.0	9.0	9.0	12.0	94.0	7.0	2.0	6.0	100% / 100%
	3	6.0	9.0	9.0	17.0	103.8	7.0	2.0	6.0	100% / 100%
Planted	1	3.0	3.0	3.0	3.0	11.8	3.0	1.0	3.0	100% / 100%
	2	6.0	9.0	9.0	12.1	96.4	7.7	2.0	6.0	100% / 100%
	3	6.0	9.0	9.0	17.1	114.4	7.5	2.0	6.0	100% / 100%
Bypass	1	3.0	3.0	3.0	3.0	8.0	–	–	3.0	49% / 49%
	2	6.0	9.0	9.0	12.0	19.1	–	–	6.0	0% / 0%
	3	6.0	9.0	9.0	17.0	20.2	–	–	6.0	0% / 0%
All in class	1	3.0	3.0	3.0	3.0	11.7	3.0	1.0	3.0	100% / 100%
	2	6.0	9.0	9.0	12.0	95.3	7.4	2.0	6.0	99% / 99%
	3	6.0	9.0	9.0	17.0	110.0	7.3	2.0	6.0	99% / 99%

Table 8: The claims of Theorems 2 and 3 checked on every run whose truth lies in V_0 (all rules, radii, families and networks).

Runs (all rules, radii, families, networks)	60,000
Runs whose truth lies in V_0	53,496
truth eliminated (Theorem 2(a))	0
halted without a certificate	0
certificate contradicted within the radius (Theorem 2(c))	0
cost above $(\lvert V_0 \rvert - 1)R$ (Theorem 3(a))	0
cost below the optimum of a learner told the truth	0
Greedy covers within $(1 + \ln M)$ of the optimal cover (Theorem 3(c))	15,000/15,000
Greedy covers costlier than the adaptive minimum-cost run	4,887/15,000
Greedy covers cheaper than the adaptive minimum-cost run	15/15,000

Table 9: Plain CNNs as observational learners (mean of three seeds). *Held-out* is the error on the observed regime; $r = 1, 2$ give the interventional risk against the shape world over all edits of cost at most r of 1,000 observed images. The floor of Theorem 1 is 0.25 at $r = 1$ and 0.36 at $r = 2$; the shape and color worlds disagree on 0.50 and 0.52 of these edits.

	$\alpha = 1$			$\alpha = 0.99$			$\alpha = 0.9$		
N	held-out	$r = 1$	$r = 2$	held-out	$r = 1$	$r = 2$	held-out	$r = 1$	$r = 2$
100	4.0%	0.48	0.52	2.1%	0.49	0.52	15.5%	0.43	0.47
300	0.0%	0.50	0.52	0.3%	0.50	0.52	7.6%	0.44	0.47
1,000	0.0%	0.50	0.52	0.3%	0.49	0.52	4.3%	0.37	0.38
3,000	0.0%	0.50	0.52	0.2%	0.48	0.51	2.6%	0.21	0.25
10,000	0.0%	0.50	0.52	0.2%	0.45	0.46	1.6%	0.08	0.11
30,000	0.0%	0.50	0.52	0.2%	0.26	0.28	1.1%	0.04	0.06
60,000	0.0%	0.50	0.52	0.2%	0.15	0.19	0.8%	0.03	0.04

Table 10: Bypass networks. *Truth in* V_0: some member of the class matches the network on every query of the family. A certificate is *void* when it halts with agreement but a survivor contradicts the network within the radius; *refuted* runs eliminate every hypothesis during learning; *validation* is the fraction of void certificates refuted by 20 random admissible queries within the radius.

		Truth	minimum-cost rule					random separating	
Family	R	in V_0	certified	correct	void	refuted	validation	void	validation
visual	1	100%	100%	100%	0%	0%	–	0%	–
	2	100%	100%	100%	0%	0%	–	0%	–
	3	100%	100%	100%	0%	0%	–	0%	–
white-box	1	0%	100%	49%	51%	0%	100%	51%	100%
	2	0%	100%	0%	100%	0%	93%	100%	93%
	3	0%	100%	0%	100%	0%	96%	100%	99%

Table 11: The four-way pointer: two pointer digits select one of four value digits, so V_0 has $2^{12} = 4{,}096$ members and the visual and white-box families have 298 and 3,535 queries of cost at most 3. 20 networks, 500 observed images; columns as in Table 7.

			ISS			random	random		
Networks	Family	R	int.	cost	opt.	sep.	probing	$\lceil \log_2 C \rceil$	correct
Trained, $\alpha = 1$	visual	1	3.9	3.9	3.9	3.9	10.3	3.9	100%
		2	10.0	16.1	12.3	19.7	118.4	10.0	100%
		3	12.0	22.1	18.0	35.5	549.5	12.0	100%
	white-box	1	3.9	3.9	3.9	3.9	24.3	3.9	100%
		2	10.0	16.1	12.3	20.1	511.2	10.0	100%
		3	12.0	22.1	18.0	36.9	2864.6	12.0	100%
Trained, $\alpha = 0.5$	visual	1	3.9	3.9	3.9	3.9	10.0	3.9	100%
		2	10.0	16.1	12.3	19.9	117.1	10.0	100%
		3	12.0	22.1	18.0	35.6	552.8	12.0	100%
	white-box	1	3.9	3.9	3.9	3.9	24.3	3.9	100%
		2	10.0	16.1	12.3	20.3	507.0	10.0	100%
		3	12.0	22.1	18.0	36.2	2939.9	12.0	100%
Planted	visual	1	3.9	3.9	3.9	3.9	10.1	3.9	100%
		2	9.2	14.4	12.0	19.0	107.8	10.0	100%
		3	11.4	21.2	17.3	35.1	485.5	12.0	100%
	white-box	1	3.9	3.9	3.9	3.9	24.1	3.9	100%
		2	10.0	16.1	13.5	20.6	529.1	10.0	100%
		3	12.0	22.1	18.4	36.4	3426.7	12.0	100%

E Extended Related Work

This appendix expands the overview in Section 2.

E.1 Structural causal models, interventions, and the causal hierarchy

We use structural causal models (Pearl, 2009; Peters et al., 2017; Spirtes et al., 2000): endogenous variables determined by assignments from parents and exogenous variables, interventions that replace assignments, and counterfactuals evaluated by abduction, action, and prediction. Bareinboim et al. (2022) organize these queries into the Pearl Causal Hierarchy and prove that, for almost all SCMs, lower-layer observational data underdetermine interventional and counterfactual quantities. History-equivalent but interventionally distinct perspectives are a finite, deterministic, individual-level instance of this underdetermination; we quantify *how far* one must move from the observed layer to resolve it. The interventionist account of causation (Woodward, 2003) views causal claims as claims about what happens under interventions. Eberhardt & Scheines (2007) and Korb et al. (2004) analyze the structural versus parametric and hard versus soft varieties on which such claims depend; our admissible family $\mathcal{Q}$ makes this dependence explicit. Halpern & Pearl (2005a;b) and Halpern (2016) define actual causes and explanations relative to an SCM and a context, and explanations relative to an *epistemic state*, a set of contexts the agent considers possible; our exogenous evidence set plays this role. Ibeling & Icard (2023) compare structural, potential-outcome, graphical, and abstraction-based frameworks; we work in the structural one. Shmueli (2010) articulates the gap between explanatory and predictive modeling that motivates our setting.

E.2 Observational and interventional equivalence

The classical notion is Markov equivalence: DAGs entailing the same conditional independences cannot be distinguished from observational distributions (Verma & Pearl, 1990; Andersson et al., 1997; Spirtes et al., 2000). Hauser & Bühlmann (2012) characterize interventional Markov equivalence classes for a fixed family of hard interventions, and Yang et al. (2018) extend this to general, including soft, interventions. These results concern population distributions over many contexts and answer a binary class-membership question. Our notion is individual-level (one observed history), graded (a cost threshold), and relative to an admissible family and cost of departure. Distributional equivalence is neither necessary nor sufficient for history-equivalence, but we share the guiding idea that models should be compared through the interventions that distinguish them.

E.3 Confounding, identification, and exogenous grounding

Confounding and observational ambiguity. Confounding can arise when treatment and outcome share causes, creating an open back-door path such as $A \leftarrow C \rightarrow Y$; consequently, $P(Y \mid A = a)$ can differ from $P(Y \mid \mathrm{do}(A = a))$ even with unlimited observational data (Pearl, 1995; Hernán & Robins, 2020). A variable is not a confounder merely because it is exogenous or correlated with treatment: its causal role and the adjustment set matter. This is one source of observational ambiguity, whereas the paper's history-equivalence only requires agreement at a single realized state. Such agreement can also arise from different mechanisms away from that state, even without a shared cause.

Identification versus separation. Back-door adjustment identifies population effects when suitable observed covariates block the relevant paths and the required support conditions hold (Pearl, 1995; Hernán & Robins, 2020). More generally, Shpitser & Pearl (2008) characterize identification of interventional and counterfactual distributions, including when latent variables prevent identification from lower-layer information. These results ask whether all models compatible with the available distribution and causal assumptions agree on a specified query. We start with specified candidates and search for a query witnessing disagreement at minimum departure cost. Thus, a separating intervention is a proposed discriminating experiment; it does not itself identify a population effect from observational data or establish which candidate is correct.

A common-cause illustration. Consider Boolean models with $C = U_C$ and $A = C$, but with $Y = A$ in M_1 and $Y = C$ in M_2; unused exogenous inputs are ignored. At $U_C = 1$, both perspectives explain $(C, A, Y) = (1, 1, 1)$. In M_2, C is a common cause of A and Y. The query

$(\mathrm{do}(A = 0), Y)$ gives answers 0 and 1, respectively, and hence separation cost 1 under unit costs when this query is admissible. Verifying $U_C = 1$ leaves both perspectives grounded because their disagreement is structural. This illustrates why evidence about a common cause and an intervention on treatment supply different information.

Scope of exogenous grounding. Grounding constrains a realized exogenous coordinate only when its meaning is shared across perspectives and its value is independently verified. Abducting a residual from a fitted model does not verify that value. These constraints neither establish conditional exchangeability nor guarantee that all common causes have been measured; even fully grounded candidates can disagree structurally (Theorem 7). The fixed-context theory specifies no probability law over U, so it makes no population-level independence or identification claim. A probabilistic extension could represent latent common causes through dependence among exogenous disturbances and combine grounding with identification or sensitivity analysis. The present benchmark does not evaluate robustness to latent confounding.

E.4 Experimental design for causal discovery and model discrimination

Two traditions ask which experiments distinguish causal hypotheses. For structure learning, Eberhardt et al. (2005) show that $\log_2 N + 1$ multi-variable experiments suffice, and are worst-case necessary, to identify all causal relations among N variables, while Eberhardt (2008) gives almost optimal intervention sets. Later work optimizes intervention size (Shanmugam et al., 2015), cost (Kocaoglu et al., 2017; Lindgren et al., 2018), fixed budgets (Ghassami et al., 2018), or targeted subgraphs (Agrawal et al., 2019); Hauser & Bühlmann (2014) and Squires et al. (2020) give optimal and near-optimal active strategies, and Bayesian active structure learning dates to Tong & Koller (2001) and Murphy (2001). Closest to us, Choo et al. (2022) study *verification*, the minimum interventions needed to confirm a purported DAG within its equivalence class, characterized through covered edges, with subset versions in Choo & Shiragur (2023). These works identify one structure among all DAGs; we compare two *given* models, possibly learned and possibly wrong, and seek the cheapest intervention on which their answers differ. In statistics and systems biology, Hunter & Reiner (1965) and Box & Hill (1967) design sequential experiments to discriminate between rival mechanistic models, Atkinson & Fedorov (1975) introduce T-optimal designs maximizing rival lack of fit, and Chaloner & Verdinelli (1995) review Bayesian designs based on the expected information of Lindley (1956). For Boolean regulatory networks, Ideker et al. (2000) choose perturbations discriminating among hypotheses consistent with the data, Akutsu et al. (2003) bound the disruption and overexpression experiments needed to identify a network, and Apgar et al. (2008) and Kreutz & Timmer (2009) design stimuli for model selection in signaling networks. Separation cost is a worst-case, budget-indexed relative: rather than maximizing expected divergence under a prior, it finds the minimum departure from the observed state at which rivals are *certain* to disagree, treating departure cost as primary. The Boolean-network methods of Ideker et al. (2000) and Akutsu et al. (2003) are the closest in spirit and differ on three points: they keep the candidate networks as an explicit list and score an experiment by the candidates it eliminates or the divergence it induces, whereas our version space is implicit and separation is decided exactly over all 2^m completions by integer programs; their target is identification of the true network, whereas ours is a certified radius of interventional agreement that is meaningful even when many networks remain; and their guarantee is identification under the class assumption, whereas the certificate is exact, cost-bounded above and below, and explicitly conditional on realizability, with a validation phase for the misspecified case. In the terms of Popper (2002), this is the cheapest potential falsifier of at least one perspective.

E.5 Amortized and cost-constrained active interventional design

The strongest recent methods for choosing interventions are amortized or explicitly cost-constrained, and it is worth saying precisely where ISS stands relative to them. Annadani et al. (2024) train CAASL, a transformer policy, by reinforcement learning on a simulator of the design environment, formalized as a hidden-parameter MDP whose hidden parameters are the graph and the mechanism parameters; the reward is the improvement in the number of correct entries of the adjacency matrix predicted by the amortized posterior of Lorch et al. (2022), which the authors relate to an approximation of multi-step expected information gain, and the design horizon is a fixed number of interventions. Zhang et al. (2023) assume a ladder of experimental oracles—higher fidelity is more precise and more expensive—and choose which variable to intervene on *and at which fidelity*

by a mutual-information criterion, with a cascading model tying the fidelities together and an ε-submodularity argument restoring the greedy guarantee that ordinary submodularity fails to provide here. Guo et al. (2026) give the most explicitly cost-bounded of the three: COPEx pre-trains an amortized posterior and design policy offline and then plans online by multi-step lookahead over a scenario tree of hypothetical outcomes, so that a design sequence can respect a budget, varying costs and design-dependent feasibility at test time. Earlier steps in the same direction include the scalable interventional design of Tigas et al. (2022).

Five differences separate this line of work from ours; none of them is a claim that one target is better than the other.

What is unknown. All three carry uncertainty over the causal *structure* and score a design by how much of it the design resolves; CAASL rewards recovery of the adjacency matrix and explicitly does not reward learning the mechanism parameters. In our setting the parent sets are given and the structure is not in doubt: what is undetermined is the mechanism on the parent configurations the observational regime never realizes, together with the exogenous context at a fixed observed state. CAASL's reward therefore has nothing to gain here: the quantity it optimizes is handed to the learner at the outset. The two targets are sequential rather than rival—structure discovery upstream, certification of the surviving mechanisms downstream—which is why Section 7 lists given parent sets as a limitation and not as an assumption we defend.

What cost means. CAASL has no cost model at all: its budget is a count of interventions, and two designs of the same count are equally affordable however far they are from the observed state. Zhang et al. (2023) price the *oracle* rather than the intervention, so cost indexes measurement quality, not how radical an experiment is. COPEx does price designs and plan under a budget, but cost there constrains the design sequence; it does not index a guarantee. In ISS the cost of an intervention is its distance from the observed state, and the same number indexes what is certified: the radius R is both the budget and the scope of the guarantee (Equation equation 1, Theorem 2).

What is returned, and when to stop. These methods return an estimate—a posterior, or a design sequence—after a fixed horizon, and their guarantees are about expected information or empirical recovery; none of them stops on a property of what survives. The difference is not that an information criterion *cannot* express our stopping condition. Answers here are deterministic and the version space is finite, so a query's expected information gain is zero exactly when it separates nobody, and $\max_{q \in \mathcal{Q}_R} \mathrm{EIG}(q) = 0$ is precisely $\kappa(V) > R$: the exhaustion test and the information criterion agree. The difference is that the equivalence holds only when the expectation is taken over the *whole* version space. Under an amortized posterior, or a sample of one, a vanishing maximum is a statement about the approximation, not about the version space. ISS's test is exact by construction (Proposition 3) and is a statement about the surviving hypotheses alone, hence checkable without knowing the truth (Theorem 2(d)). That is the substantive difference, and it is why the certificate survives replacing our acquisition rule with theirs.

Expectation against worst case. An expected-information criterion averages over a prior. Separation cost is a worst-case, budget-indexed quantity: the minimum departure from the observed state at which two rivals are *certain* to disagree. The two often coincide in practice—in the Boolean MNIST study the information-per-cost rule pays almost exactly what the minimum-cost rule pays (Table 7)—and differ entirely in what they license at the end of a run.

E.6 Underspecification, multiplicity, and disagreement between predictors

That many models fit the same data equally well is the Rashomon effect (Breiman, 2001); Fisher et al. (2019) and Semenova et al. (2022) study Rashomon sets, while Marx et al. (2020)quantify ambiguity and discrepancy among near-equal-accuracy models and Black et al. (2022) survey the opportunities and concerns such multiplicity raises. D'Amour et al. (2022) show that underspecified pipelines yield predictors with equivalent held-out performance but divergent stress-test behavior, arguing that such tests rather than held-out accuracy must decide credibility; we formalize the *minimal* stress test deciding between two candidates. Shortcut learning (Geirhos et al., 2020; Beery et al., 2018) and invariance (Peters et al., 2016; Arjovsky et al., 2019; Sagawa et al., 2020; Veitch et al., 2021) explain why observationally agreeing predictors differ causally and use environments or counterfactual edits to identify causal features. De Haan et al. (2019) show that imitation policies can fit demonstrations while conditioning on the wrong causes, then resolve this confusion through

targeted interventions; Richens & Everitt (2024) prove that agents robust to distribution shift must learn a causal world model, motivating small separation cost as fragility. Disagreement already guides query-by-committee, which shrinks a version space using queries on which hypotheses disagree (Mitchell, 1982; Seung et al., 1992; Freund et al., 1997); Bayesian active learning scores expected disagreement (Houlsby et al., 2011), ensembles use disagreement as uncertainty (Lakshminarayanan et al., 2017), and disagreement across training runs predicts test error (Jiang et al., 2022). We replace the input space by interventions and the disagreement *rate* by the disagreement *threshold*: the cheapest intervention around the observed history on which two candidates disagree.

E.7 Counterfactual explanations, recourse, and minimal perturbations

Counterfactual explanations return the closest input changing a prediction (Wachter et al., 2018; Mothilal et al., 2020), and actionable recourse restricts changes to those an individual can make (Ustun et al., 2019). Karimi et al. (2021) cast recourse as a minimum-cost SCM intervention accounting for downstream effects; Karimi et al. (2022) survey the field. Our departure cost is that of causal recourse, but the objective differs: recourse flips one model's output whereas separation finds the cheapest intervention on which two outputs differ. Adversarial examples minimally perturb an input to flip one model's decision (Szegedy et al., 2014; Moosavi-Dezfooli et al., 2016); separation instead targets *disagreement between models*, with do-semantics rather than pixel perturbations. Behavioral testing manually constructs near-history edits through counterfactually augmented data (Kaushik et al., 2020), contrast sets (Gardner et al., 2020), CheckList (Ribeiro et al., 2020), and the CEBaB benchmark of real-world concept interventions (Abraham et al., 2022). These are empirical low-cost interventions; our theory identifies which are guaranteed informative and how cheap the cheapest can be.

E.8 Differential testing, distinguishing sequences, and query learning

Differential testing searches for inputs on which implementations disagree (McKeeman, 1998); DeepXplore does so for deep networks in a whitebox manner (Pei et al., 2017), and DiffChaser searches for disagreements between network versions (Xie et al., 2019). Distinguishing and checking sequences separate finite-state machines (Lee & Yannakakis, 1996), while in query learning an equivalence query returns a counterexample separating hypothesis from target (Angluin, 1988). Separation cost is the cheapest counterexample to the claim of interventional equivalence, measured by departure from an observed state rather than sequence length or input norm, with interventions on a causal model rather than inputs to a function.

E.9 Causal abstraction and interventional interpretability

Causal abstraction asks when one SCM faithfully coarse-grains another (Rubenstein et al., 2017; Beckers & Halpern, 2019; Beckers et al., 2020). Geiger et al. (2021) test whether a network implements a hypothesized causal model through interchange interventions on internal representations, Geiger et al. (2024) learn alignments between causal variables and distributed representations, and Geiger et al. (2025) give the theoretical foundations. Causal mediation (Vig et al., 2020), tracing and editing (Meng et al., 2022), circuit discovery (Wang et al., 2023; Conmy et al., 2023), and neural-causal models (Xia et al., 2021) likewise probe models interventionally. These methods quantify agreement between a network and a hypothesis *averaged* over interventions; we seek the *minimum-cost* environment intervention distinguishing two hypotheses, a complementary criterion applicable to the abstracted models they produce. Causal representation learning and independent causal mechanisms (Schölkopf et al., 2021; Parascandolo et al., 2018) motivate our benchmark's sparse-mechanism-difference regime, while causal data fusion (Bareinboim & Pearl, 2016) formalizes the combination of observational and interventional evidence, which is what the ISS loop does adaptively and at a cost. Evaluations of causal reasoning in language models (Jin et al., 2023; Kıcıman et al., 2024; Zečević et al., 2023) and causal criteria in fairness (Kusner et al., 2017; Zhang & Bareinboim, 2018) are natural application targets.

E.10 Exogenous variables, abduction, and computation

Counterfactual evaluation begins by abducting the exogenous context (Balke & Pearl, 1994; Pearl, 2009); probabilities of causation require bounds precisely because contexts are underdetermined (Tian & Pearl, 2000), while Gumbel-max SCMs (Oberst & Sontag, 2019) and counterfactual data augmentation (Lu et al., 2020) infer posteriors over exogenous noise from trajectories. Exogenous grounding takes the opposite stance: verified exogenous facts are constraints that eliminate hypotheses from the version space, not latent variables to marginalize. Computationally, Eiter & Lukasiewicz (2002) and Aleksandrowicz et al. (2017) settle the complexity of structure-based actual causation, Ibrahim & Pretschner (2020) solve actual-causality queries in Boolean models with integer linear programs, and Ibeling & Icard (2020) and Mossé et al. (2024) show that satisfiability for causal and counterfactual languages is no harder than for their probabilistic counterparts; Boolean causal models also inherit the Boolean-network tradition (Kauffman, 1969). Our NP-completeness result sits alongside these, and our separation checks follow the integer-programming approach of Ibrahim & Pretschner (2020).