Calibrated, Falsifiable Detection of Sparse Mechanism Shift

Many methods for distribution shift under causal structure assume the sparse mechanism shift (SMS) hypothesis: that across environments only a few causal conditionals change. This assumption drives mechanism-shift scoring, causal discovery in heterogeneous data, and transportable prediction, yet it is almost never tested on the data at hand. This paper makes SMS testable. We first ask why it is hard to tell which mechanisms changed once the causal graph must be estimated rather than assumed known. A controlled ablation locates the cause: the false positives that limit precision arise at the truly invariant nodes, because their parent sets are mis-estimated. We then give a graph-free, label-free detector that flags a node only when no conditioning subset makes its conditional invariant across environments (an inverse use of invariant causal prediction), which matches an oracle that knows the true graph. Building on it, we define a calibrated SMS hypothesis test with a data-driven null floor and a bootstrap three-way verdict (no shift / sparse / dense). On synthetic data the verdict tracks the true sparsity; on real protein-signalling interventions it rejects SMS, and a paired atomic-versus-fat-hand study explains the rejection and predicts when SMS should hold.

AUTHORS
Anonymous

AFFILIATIONS
Anonymous

SUBMITTED
Not submitted yet.

Contents

1. Introduction

2. Related Work

3. Method

3.1 Problem setup
3.2 A mechanism-change test
3.3 Diagnosing the estimated-graph bottleneck
3.4 Graph-free detection via existence of an invariant set
3.5 A calibrated SMS hypothesis test
3.6 Nonlinear extension
3.7 Extension to image data via a fixed factor encoder

4. Experiments

4.1 Controlled synthetic SMS
4.2 Diagnostic ablation and the C10 solution
4.3 Semi-synthetic and real causal structures
4.4 Real interventions: the Sachs data
4.5 The calibrated SMS test
4.6 Atomic versus fat-hand interventions
4.7 Extension to image data: dSprites and 3D Shapes

5. Discussion and Limitations

6. Conclusion

1. Introduction

Distribution shift (a change in the data-generating distribution between training and deployment) is a central obstacle to reliable machine learning [1, 2]. A productive way to reason about shift is through causal structure: if the data are generated by a structural causal model (SCM) whose variables $X_1, \ldots, X_d$ obey mechanisms $P(X_j \mid \mathrm{pa}_j)$, then a shift corresponds to some of these mechanisms changing while others stay invariant [3, 4]. The *independent causal mechanisms* principle further suggests that natural shifts are *localised*: a change to the world typically perturbs only a few mechanisms at once. This is the *sparse mechanism shift* (SMS) hypothesis [4]. To state it precisely, assume the data follow an SCM in which each variable X_j is produced by a *mechanism*, its conditional distribution given its causal parents, $P(X_j \mid X_{\mathrm{pa}_j})$. Comparing a source environment s with a target environment t, the changed set, its size, and the sparsity ratio are

$$S^\star = \{\, j : P^{(s)}(X_j \mid X_{\mathrm{pa}_j}) \neq P^{(t)}(X_j \mid X_{\mathrm{pa}_j}) \,\}, \qquad k = |S^\star|, \qquad \rho = \frac{k}{d}, \tag{1}$$

and SMS posits that only a small fraction of the d mechanisms change across environments, i.e. $\rho \ll 1$; the shift is *dense* when ρ is large (a majority of mechanisms change).

SMS is attractive because it licenses powerful procedures. If only a few mechanisms change, one can score and localise them [5], discover structure from heterogeneous or nonstationary data by treating the environment as an auxiliary variable [6], and build predictors that transport by relying on the invariant part of the model [7, 8, 9]. Yet across this literature SMS is typically an *assumption*: methods presuppose that the shift is sparse and proceed, without testing whether sparsity actually holds in the data at hand. When the assumption silently fails (when a shift is in fact dense), the downstream guarantees quietly fail with it.

As a concrete example, consider a protein-signalling network in which each protein's abundance is generated from its upstream regulators [10]. A targeted drug that inhibits a single protein is an *atomic* intervention: ideally it alters only that protein's mechanism and leaves the others invariant, so ρ is small and SMS holds. A broad-acting drug that perturbs a whole pathway is a *fat-hand* intervention: many conditionals shift at once, ρ is large, and SMS is violated. Whether SMS holds is therefore not a modelling convenience but an empirical property of the specific shift, one that, we argue, should be measured rather than assumed.

Objective. The goal of this paper is to make the SMS assumption *testable* on the data at hand. Concretely, we aim to: (i) characterise when and why localising the changed mechanisms becomes hard once the causal graph must be estimated rather than known; (ii) design a detector that recovers the changed set $S^\star$ without a known graph or labels; and (iii) build a calibrated, falsifiable test that returns an accept/reject verdict for SMS together with a sparsity estimate $\hat{\rho} = |\hat{S}|/d$, the empirical counterpart of the true ratio ρ in Eq. (1), with the unknown changed set $S^\star$ replaced by the detector's estimate $\hat{S}$. Pursuing these objectives, our contributions are:

1. **A diagnosis of mechanism-shift detection under an estimated graph.** Through a controlled ablation we show the precision bottleneck is systematic false positives at the truly *invariant* nodes, induced by mis-estimated parent sets. It is *not* the global skeleton quality, *not* the changed nodes, and *not* fixable by conditioning-set stability selection.

2. **A graph-free, label-free detector** that flags a node only when *no* candidate conditioning subset renders its conditional invariant across environments, an inverse use of invariant causal prediction [7]. It is robust to the diagnosed failure mode and, with a false-discovery-rate (FDR) decision [11], matches an oracle that knows the true graph.

3. **A calibrated SMS hypothesis test**: a sparsity statistic $\hat{\rho} = |\hat{S}|/d$ together with a data-driven null floor (split one environment in half, where no real shift exists) anc a bootstrap three-way verdict (no shift / sparse / dense). Only one semantic cut-point (sparse vs. dense) remains a modelling choice.

4. **Evidence across controlled, semi-synthetic, real-structure and real-intervention data.** The verdict tracks the true sparsity on synthetic data and rejects SMS on real protein-signalling interventions; a paired atomic-versus-fat-hand study explains the rejection mechanistically and yields a falsifiable prediction.

We deliberately work in a linear-Gaussian, two-environment setting to establish the test cleanly, and provide a kernel-based extension for nonlinear mechanisms. All reported numbers come from executed runs.

2. Related Work

Mechanism-shift scoring and heterogeneous discovery. The Mechanism Shift Score (MSS) of Perry et al. [5] ranks mechanisms by how much their conditionals change across environments, explicitly *under* the SMS hypothesis. CD-NOD [6] adds the environment as a surrogate node and detects which mechanisms depend on it. Two limitations follow. First, both *presuppose* that the shift is sparse: they rank or flag mechanisms but never report whether SMS actually holds, so a dense shift is mishandled silently. Second, both lean on a known or reliably discoverable causal graph, and their accuracy degrades once the graph must be estimated from the same shifted data. We use MSS and a CD-NOD-style detector as baselines and supply the missing ingredient, a calibrated test of the premise that is also robust to graph-estimation error.

Invariance and transport. Invariant Causal Prediction (ICP) [7] seeks a predictor whose residual distribution is invariant across environments; IRM [12], anchor regression [9] and the graph-surgery view [8] build robust or transportable predictors from invariant components; DRIG [13] interpolates between empirical-risk minimisation and the causal solution. These methods, however, are oriented toward predicting a *single* target from an invariant set, not toward localising *which* mechanisms changed across the whole graph; and several (notably IRM) are known to be sensitive to environment design and optimisation. Our detector *inverts* ICP: instead of finding one invariant predictor for a target, we ask, for *every* node, whether *any* invariant conditioning set exists, and flag the node if none does.

Sparse shift estimation with FDR. Sparse-Joint-Shift / SEES [14] estimates model performance under simultaneous covariate and label shift, assuming the shift is sparse; SGShift [15] attributes concept shift to a sparse set of features with knockoff-based FDR control [16]. Both, however, operate at the *feature* level and still *assume* sparsity as an identifiability or FDR condition rather than testing it; and a feature whose marginal merely moved cannot be distinguished from a mechanism that genuinely changed versus an ancestor shifting upstream. We instead localise *causal-mechanism* changes (conditioning on causal parents separates a changed mechanism from a propagated one), and test the sparsity assumption itself.

Causal representation learning. A complementary recent line lifts causal structure from low-level data such as images: causal representation learning recovers latent factors and their causal graph, with identifiability established mainly from *interventions across environments* [17, 18, 19, 20, 21]. These are identifiability guarantees under idealised assumptions (e.g. one intervention per latent node, or the infinite-sample limit) and are not aimed at *testing* which mechanisms shifted: they say how to obtain aligned factors but stop short of the downstream sparsity test. We connect the two: such an encoder is exactly the front-end our image extension relies on, and our finding that accurate changed-set recovery needs a *cross-environment-consistent* encoder mirrors the role of multi-environment interventional data for identifiability there; intervention-extrapolation guarantees [22] are especially pertinent to the target-domain bias we isolate.

The common gap. Across these lines three issues recur and motivate this work. (i) Sparsity is *assumed, not tested*: no method reports whether SMS holds on the data at hand, so downstream guarantees fail silently when the shift is in fact dense. (ii) Accurate mechanism localisation is demonstrated only under an *oracle* (a known graph, a designated prediction target, or generative-factor-aligned latents), leaving the realistic regime (estimated graph, no labels, no aligned encoder) largely untreated. (iii) The *failure modes* in that regime are uncharacterised, so it is unclear what to repair. We answer these in turn with a calibrated test of the premise, a graph-free and label-free detector, and a precise diagnosis of where estimated-graph detection breaks.

3. Method

3.1 Problem setup

We observe data from two environments, a source s and a target t, generated by SCMs that share a directed acyclic graph G over variables $X = (X_1, \ldots, X_d)$ with parent sets pa_j:

$$X_j \;:=\; f_j^{(e)}\Big(X_{\mathrm{pa}_j}, N_j^{(e)}\Big), \qquad e \in \{s, t\}, \tag{2}$$

with mutually independent noises $N_j^{(e)}$. The *mechanism* of node j in environment e is the conditional $P^{(e)}(X_j \mid X_{\mathrm{pa}_j})$. The changed set $S^\star$, its size $k = |S^\star|$, and the sparsity ratio $\rho = k/d \in [0, 1]$ are as introduced in Eq. (1); the shift satisfies the sparse mechanism shift hypothesis when ρ is small.

Note that a changed mechanism at node j moves the conditional $P(X_j \mid X_{\mathrm{pa}_j})$, but the *marginal* of a downstream node can move even if its own mechanism is invariant, because the change propagates through the graph. Distinguishing "this mechanism changed" from "an ancestor changed upstream" is exactly what conditioning on the causal parents buys, and is why feature-level (marginal) detection is insufficient.

For the controlled setting we use linear-Gaussian SCMs: a changed node receives an intercept shift, a parent-coefficient shift, and a noise-variance rescaling, so that every unchanged node's conditional stays exactly invariant while its marginal may move. This gives ground-truth $S^\star$ for every run, so detection can be scored.

3.2 A mechanism-change test

Given a candidate conditioning set S (intended to be the parents of j), the atomic question is whether node j's mechanism changed, that is, whether the conditional $P(X_j \mid X_S)$ is the *same* in both environments. We cast this as a likelihood-ratio test (LRT) between two linear-Gaussian models of X_j regressed on X_S. The *reduced* model encodes the null hypothesis H_0 "the mechanism is invariant": it pools the two environments and fits a *single* shared conditional. The *full* model encodes the alternative "the mechanism changed": it fits a *separate* conditional in each environment. Writing ℓ_{reduced} and $\ell_{\mathrm{full}} = \ell^{(s)} + \ell^{(t)}$ for the two maximised log-likelihoods,

$$\Lambda_j(S) \;=\; 2\big(\ell_{\mathrm{full}} - \ell_{\mathrm{reduced}}\big) \;\overset{H_0}{\sim}\; \chi^2_{|S|+2}. \tag{3}$$

The full model can only fit at least as well, so $\Lambda_j(S) \geq 0$ measures how much a per-environment fit improves on the pooled one. Under H_0 it is asymptotically χ^2 with $|S| + 2$ degrees of freedom: the free parameters that the second environment's conditional adds over the shared one ($|S| + 1$ regression coefficients: an intercept and one slope per variable in S, plus one noise variance). Because each environment's conditional is fitted independently in *both* its coefficients and its variance, the statistic detects a mean change (a shifted intercept or parent coefficients) *and* a variance change (rescaled noise), exactly the ways a mechanism can move in our SCM. A small p-value $p = P(\chi^2_{|S|+2} \geq \Lambda_j(S))$ rejects invariance and flags node j as changed. Applying the test with $S = \mathrm{pa}_j$ at every node and controlling the false discovery rate with Benjamini–Hochberg (BH) [11] over the d nodes yields the estimated changed set $\hat{S}$ and the sparsity estimate $\hat{\rho} = |\hat{S}|/d$. With a *known* graph this is accurate; the difficulty, which the next subsection diagnoses, is that in practice G (and hence each parent set pa_j) must be estimated.

3.3 Diagnosing the estimated-graph bottleneck

When the graph is estimated (e.g. by the PC algorithm [23] on pooled or single-environment data), detection precision drops sharply. We localise the cause with an ablation [1] that holds the data-generating process and seeds fixed and changes only one ingredient at a time (Table 1):

- A *better global skeleton* does not help: running the combined detector on a high-quality source-only skeleton does not beat its pooled-graph counterpart, so global skeleton quality is not the bottleneck.
- Repairing the *changed* nodes' parents does not help: oracle-repairing the flagged nodes' parents or doing real local repair there leaves F_1 essentially unchanged.
- Repairing the *unchanged* nodes' parents *does*: oracle-repairing the presumed-stable nodes' parents jumps F_1 to near the oracle. Hence the false positives that cap precision come from mis-estimated parent sets of the *invariant* nodes: a wrong parent set makes a stable conditional look environment-dependent.
- Conditioning-set *voting* cannot fix it: stability selection over resampled plausible parent sets is *worse*, because the stable-node false positives are systematic (most sampled parent sets miss a true parent), so voters share the bias.

Table 1. Diagnostic ablation: changed-set F_1 versus the true number of changed mechanisms k (linear-Gaussian SCM, $d = 10$, $n = 3000$ per environment, 15 seeds, FDR $\alpha = 0.1$). Repairing the *stable* nodes recovers oracle-level F_1; repairing the changed nodes, using a better (source-only) skeleton, or conditioning-set voting does not.

Method	k=0	k=2	k=4	k=6	k=8
Oracle graph + FDR (upper bound) [11]	0.80	0.95	0.96	0.98	1.00
Pooled-PC graph + FDR [23]	0.87	0.57	0.67	0.81	0.91
Combined on source-only skeleton [23]	0.80	0.53	0.69	0.84	0.93
Oracle-repair *changed* nodes	0.80	0.52	0.69	0.84	0.93
Local-repair *changed* nodes	0.80	0.51	0.68	0.83	0.93
Oracle-repair *stable* nodes	0.80	0.92	0.93	0.97	1.00
Conditioning-set stability selection [24]	0.73	0.50	0.63	0.77	0.90

Method	$k{=}0$	$k{=}2$	$k{=}4$	$k{=}6$	$k{=}8$
Exists-invariant-set (ours) [7]	**1.00**	**0.99**	**0.96**	**0.98**	**0.98**

3.4 Graph-free detection via existence of an invariant set

Our detector responds to this diagnosis. Rather than commit to one (possibly wrong) parent set per node, we ask whether *some* conditioning set makes the node's conditional invariant. This inverts invariant causal prediction [7]: a node is declared *changed* only if *no* candidate subset survives the invariance test. Formally, with a candidate pool $\mathcal{C}_j$ (by default the variables most correlated with X_j, so a true parent missed by an estimated graph can still be tried),

$$q_j \;=\; \max_{S \subseteq \mathcal{C}_j,\, |S| \le m} \; p\big(\Lambda_j(S)\big), \tag{4}$$

the largest invariance p-value over candidate subsets. A large q_j means an invariant set exists (the node is saved); a small q_j means every set was rejected (the node changed). We apply BH-FDR to $\{q_j\}_{j=1}^{d}$ to obtain $\hat{S}$. Because a stable node is rescued whenever *any* candidate set (e.g. its true parents) looks invariant, the detector is robust to the parent-set errors that defeat the parent-conditioned test, precisely the stable-node failure mode identified above. We refer to this exists-invariant-set detector as C10. Algorithm 1 summarises it; the per-node subset search is the dominant cost and is evaluated by batched linear algebra.

Algorithm 1 Exists-invariant-set detector (C10) and calibrated SMS verdict

Require: source/target samples $X^{(s)}, X^{(t)} \in \mathbb{R}^{n \times d}$; level α; pool size; max $|S| = m$; threshold τ_{dense}

```
function Detect(X^(s), X^(t), α)
  for j = 1, ..., d do
    C_j ← top variables by |corr(X_j, ·)| on pooled data
    q_j ← max_{S⊆C_j, |S|≤m} p(Λ_j(S))   ▷ Eq. (3), early-stop at p ≥ 0.9
  end for
  Ŝ ← { j : BH-FDR({q_j}, α) rejects j }
  return Ŝ
end function
ρ̂ ← bootstrap mean of |Detect(X^(s), X^(t))|/d   ▷ with 90% CI
ρ_null ← bootstrap mean of |Detect(X^(s)_half 1, X^(s)_half 2)|/d   ▷ no real shift
if CI_lo(ρ̂) ≤ CI_hi(ρ_null) then
  return NO SHIFT
else if ρ̂ < τ_dense then
  return SPARSE (SMS HOLDS)
else
  return DENSE (REJECT SMS)
end if
```

3.5 A calibrated SMS hypothesis test

The detector gives a sparsity estimate; to *test* SMS we need a reference for "how sparse is sparse". We avoid a guessed threshold for the lower end by calibrating a nul floor from the data: split the source environment in half, where *no* real shift exists, and run the same detector to obtain ρ_{null}, the detector's own false-positive floor. The decision uses bootstrap 90% confidence intervals (Algorithm 1): if the observed $\hat{\rho}$ interval overlaps the null floor, there is no detectable shift; if it is separated from the floor but $\hat{\rho} < \tau_{\text{dense}}$, the shift is sparse and SMS holds; if $\hat{\rho} \ge \tau_{\text{dense}}$ (default 0.5, "a majority of mechanisms"), the shift is dense and SMS is rejected. Only τ_{dense} remains a chosen value; the noise floor is data-calibrated, which also certifies that a large observed $\hat{\rho}$ is a real dense shift rather than detector noise.

3.6 Nonlinear extension

The invariance test in Eq. (3) is linear-Gaussian. For nonlinear mechanisms we replace it inside Eq. (4) with a kernel conditional-independence test [25]: node j has an invariant set S iff $X_j \perp E \mid X_S$, where E is the environment indicator. This keeps the exists-invariant-set logic but removes the linear-Gaussian assumption.

3.7 Extension to image data via a fixed factor encoder

Because the test operates on a matrix of variables, it extends to images through a front-end encoder g that maps each image to a low-dimensional factor vector $z = g(x)$; the recovered factors are then fed, unchanged, to the same detector and calibrated test. The encoder is *frozen* after training, a pure preprocessing step that leaves the statistical core untouched, so the SMS machinery is reused verbatim. The catch is that sparsity is *basis-dependent*: a shift that is sparse in the causal factors can appear dense in an entangled representation, so g must output factors aligned with the generative ones rather than generic deep features. We therefore evaluate on datasets with *known* generative factors and study how the choice of g governs recovery.

4. Experiments

Table 2 summarises the datasets used throughout. They span synthetic SCMs with full ground truth, semi-synthetic and published causal graphs, real protein-signalling interventions, and image data with known generative factors. Sample sizes and seed counts are reported with each experiment below.

Table 2. Datasets used in the experiments. d counts the variables (graph nodes or generative factors); n per env. is the per-environment sample size (for Sachs, observational / per-intervention cell counts, 5400 total). "LG" denotes linear-Gaussian mechanisms; the nonlinear synthetic SCM uses $\tanh$ mechanisms for the kernel (KCI) detector. dSprites (binary) and 3D Shapes (RGB) are 64×64 images with LG factor mechanisms.

Dataset	Type and mechanism	d	n / env	Ground truth
Synthetic SCM (controlled)	Synthetic (LG)	12	4000	graph, $S^\star$
Synthetic SCM (ablation)	Synthetic (LG)	10	3000	graph, $S^\star$
Synthetic SCM (nonlinear)	Synthetic ($\tanh$)	6	800	graph, $S^\star$
American Community Survey [26]	Semi-synthetic (LG)	8	4000	planted $S^\star$
ASIA [27]	Real topology (LG)	8	3000	graph, $S^\star$
SACHS [27]	Real topology (LG)	11	3000	graph, $S^\star$
CHILD [27]	Real topology (LG)	20	3000	graph, $S^\star$
ALARM [27]	Real topology (LG)	37	3000	graph, $S^\star$
Sachs protein signalling [10]	Real (nonlinear, non-Gaussian)	11	1800 / 600–1200	targets
Atomic vs. fat-hand SCM	Synthetic (LG)	12	3000	$S^\star$
dSprites [28]	Real images (LG factors)	5	3000	factors, $S^\star$
3D Shapes [29]	Real images (LG factors)	6	2500	factors, $S^\star$

4.1 Controlled synthetic SMS

With a known graph ($d = 12$, $n = 4000$ per environment, 30 seeds, FDR $\alpha = 0.1$), the mechanism-change test recovers $S^\star$ reliably: recall is 1.0 at every k, $F_1 \approx 0.90$–1.0, and $\hat{k}$ tracks the truth (e.g. $k = 4 \to \hat{k} \approx 4.5$, $k = 8 \to \hat{k} \approx 8.2$). The abstain rate (the fraction of runs in which the designated target's own mechanism is flagged, so no invariant predictor of it exists) rises from 0 to 1 with k, tracking k/d, and the abstain decision is correct 97–100% of the time (Figure 1). Certified invariance buys robustness: among target-stable cases the parents predictor's transport error stays small (≈ 0.001–0.03) across all k, whereas an empirical-risk-minimising "ERM-all" predictor degrades sharply as the shift densifies (mean transport error $\approx 0.007 \to 27 \to 219$), a heavy-tailed phenomenon across random graphs (Figure 2).

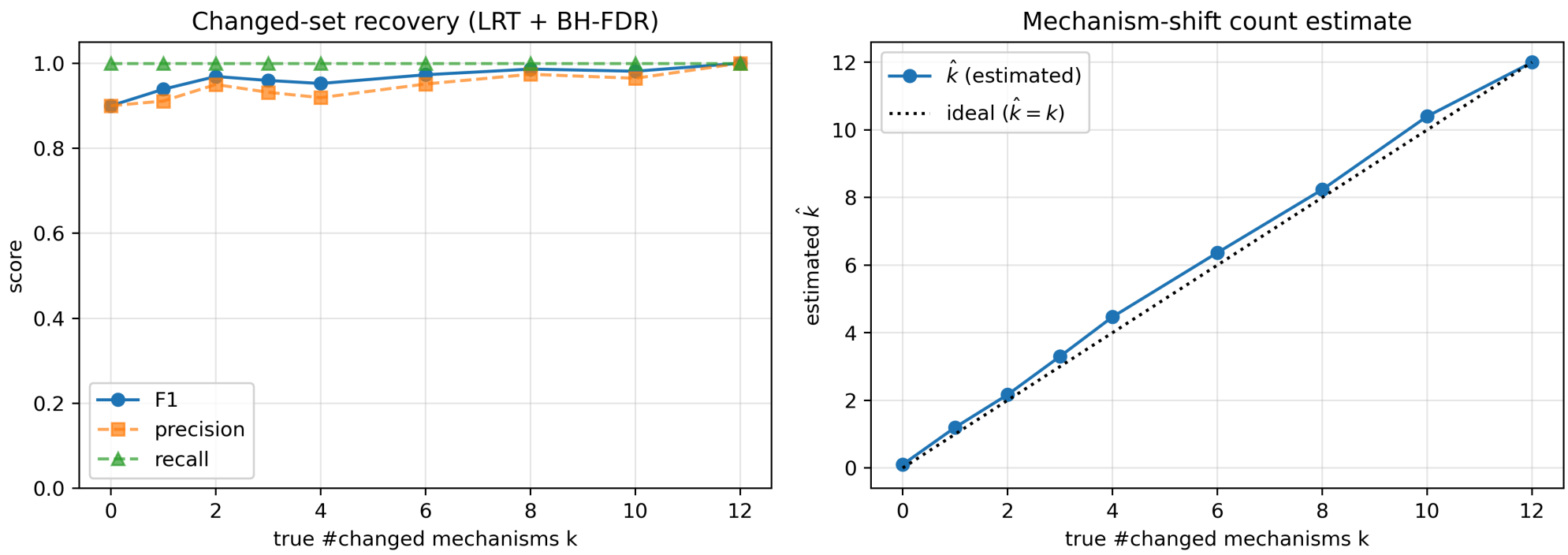

Figure 1. Controlled synthetic SMS with a known graph. Left: changed-set recovery (F1, precision, recall) versus the true number of changed mechanisms k; recall is 1.0 throughout. Right: the estimated count k-hat tracks the ideal k-hat = k.

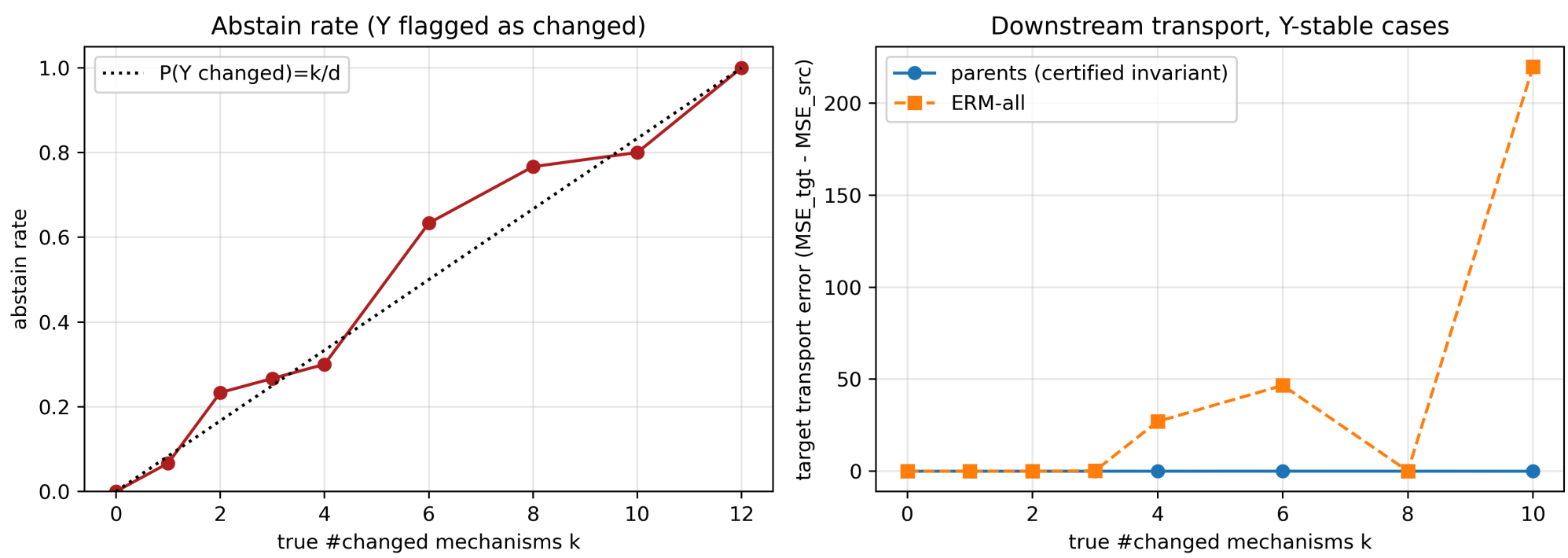

Figure 2. Downstream impact. Left: the abstain rate rises with k (about k/d). Right: among target-stable cases, the certified parents predictor transports with small error while ERM-all degrades sharply as the shift densifies.

4.2 Diagnostic ablation and the C10 solution

Dropping the known-graph assumption ($d = 10, n = 3000$, 15 seeds), Table 1 and Figure 3 establish both the diagnosis and our solution. Two further comparisons sharpen the message. First, FDR control beats fixed-threshold MSS exactly in the sparse regime it targets: at $k = 0$ the MSS baseline has $F_1 \approx 0.53$ versus FDR's 0.80 (the gap closes as the shift densifies and false positives matter less). Second, mechanism-level beats feature-level: a graph-free marginal detector flags any variable whose marginal moved, inflating its estimate far beyond the truth (precision ≈ 0.44 at $k = 2$ versus the oracle's 0.92), because it cannot tell a changed mechanism from a shifted ancestor. The upgraded C10 (global candidate pool + FDR) reaches recall 1.0 at every k and $F_1 = 1.00/0.99/0.96/0.98/0.98$ at $k = 0/2/4/6/8$, matching the oracle and even exceeding it at small k, with no labels and no known graph. Figure 4 shows the complementary discovery view: single-environment (source-only) discovery keeps graph recovery flat as the shift densifies, while pooled and CD-NOD discovery degrade; and CD-NOD's environment adjacency is a high-precision but falling-recall changed-set signal, complementary to the high-recall LRT+FDR.

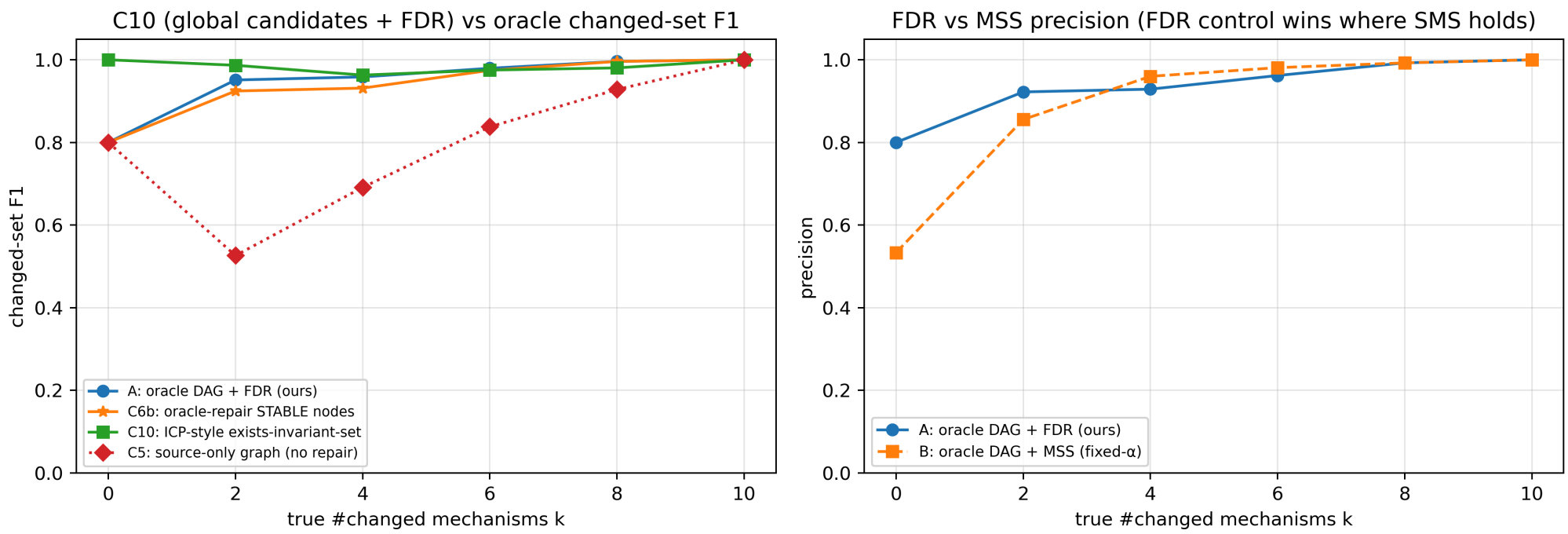

Figure 3. Estimated-graph regime. Left: the upgraded C10 detector matches the oracle on changed-set F1, solving the stable-node false-positive problem. Right: FDR control beats fixed-threshold MSS precision where SMS holds (small k).

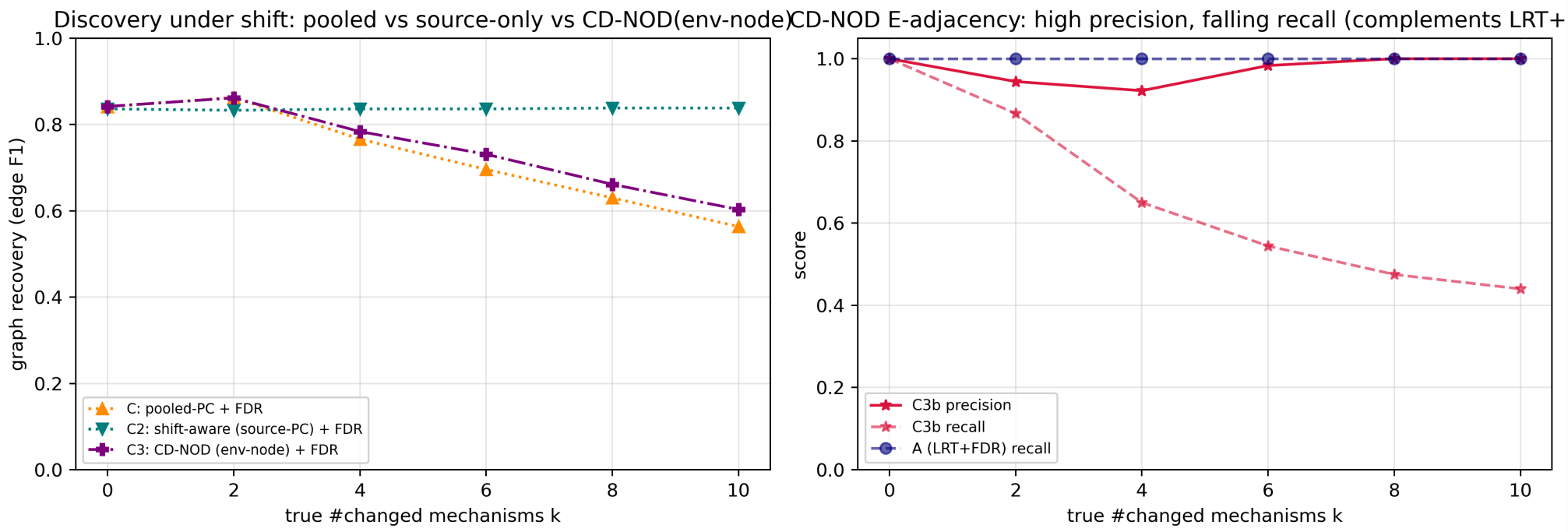

Figure 4. Discovery under shift. Left: graph recovery (edge F1) for pooled, source-only and CD-NOD (environment-node) discovery as the shift densifies. Right: CD-NOD's environment-adjacency changed-set signal has high precision but falling recall, complementing the high-recall LRT+FDR.

4.3 Semi-synthetic and real causal structures

On a linear-Gaussian SCM fit from real American Community Survey covariates [26] ($d = 8$, 15 seeds) with planted shifts, C10 attains $F_1 = 0.97$–1.0 (recall 1.0), tying or beating the oracle and far above an estimated-graph baseline, so the result is not an artefact of a clean synthetic SCM. Table 3 repeats the test on four *published* causal-graph topologies [27] (ASIA, SACHS, CHILD, ALARM; $d = 8$ to 37) with synthetic mechanisms and planted shifts. C10 ties or beats the oracle and stays robust with dimension, whereas a pooled-PC estimated graph collapses on the denser networks (e.g. SACHS F_1 down to 0.48 at $k = 2$). The nonlinear extension behaves as intended: on a $\tanh$ SCM the kernel version of C10 restores F_1 to $1.00/0.98$ at $k = 2/4$ where the linear version is misspecified ($0.81/0.85$)

Table 3. Real causal structures: changed-set F_1 for the exists-invariant-set detector (C10), the oracle (A), and an estimated pooled-PC graph (C), at k changed mechanisms. C10 ties or beats the oracle across $d = 8$–37; the estimated graph collapses on the denser networks. Pooled-PC is omitted for ALARM ($d = 37$), where it is prohibitively slow.

Network	Method	k=0	k=1	k=2	k=4
ASIA (d=8)	A: oracle	1.00	1.00	1.00	1.00
	C10 (ours)	1.00	1.00	1.00	1.00
	C: pooled-PC	1.00	0.87	0.96	0.96
SACHS (d=11)	A: oracle	1.00	0.93	0.98	0.96
	C10 (ours)	1.00	1.00	1.00	0.99
	C: pooled-PC	0.90	0.65	0.48	0.59
CHILD (d=20)	A: oracle	0.70	0.87	0.91	0.91
	C10 (ours)	1.00	1.00	0.91	0.98
	C: pooled-PC	0.70	0.62	0.68	0.71
ALARM (d=37)	A: oracle	0.90	0.95	0.97	0.97
	C10 (ours)	1.00	1.00	1.00	0.88

4.4 Real interventions: the Sachs data

We next apply the detector to real protein-signalling data with known chemical interventions [10]: each intervention is contrasted with the observational condition over 11 phospho-proteins on the published consensus topology. The detector *localises* every intervened target (it is always included in $\hat{S}$, recall $5/5$), but precision is low (≈ 0.09–0.17, $\hat{k} = 8$–11 of 11). This is not a detector failure: real chemical drugs are *fat-hand* (a single drug perturbs an entire pathway), and the observational and interventional conditions differ globally, so the real shift simply is not sparse. In fact, the test correctly diagnoses a *dense* shift on real interventions, exactly the situation SMS-assuming methods would mishandle silently.

4.5 The calibrated SMS test

Table 4 reports the calibrated verdict ($d = 12$, $n = 3000$, 20-bootstrap 90% CIs, $\tau_{\text{dense}} = 0.5$). On controlled synthetic data the verdict anchors all three regimes: $k = 0$ is called *no shift* ($\hat{\rho} \approx \rho_{\text{null}} \approx 0$, so the detector does not hallucinate a shift), $k = 1$–4 are *sparse*, and $k = 8$ is *dense*, with $\hat{\rho}$ tracking the true k/d. On all five Sachs interventions the verdict is *dense / reject SMS*; critically, the null floor is only $\rho_{\text{null}} \leq 0.28$ while $\hat{\rho} = 0.65$–1.0, proving the dense verdict is a real property of the data and not noise from the discretised, observational measurements.

Table 4. Calibrated SMS hypothesis test. $\hat{\rho} = |\hat{S}|/d$ with bootstrap 90% CI; ρ_{null} is the data-calibrated false-positive floor (source split in half), with its CI upper bound in parentheses; $\tau_{\text{dense}} = 0.5$.

Scenario	true ρ	$\hat{\rho}$ [90% CI]	ρ_{null} (hi)	Verdict
synthetic k=0	0.00	0.00 [0.00, 0.00]	0.00 (0.00)	no shift
synthetic k=1	0.08	0.08 [0.08, 0.08]	0.00 (0.00)	sparse (SMS holds)
synthetic k=2	0.17	0.36 [0.25, 0.50]	0.00 (0.00)	sparse (SMS holds)
synthetic k=4	0.33	0.35 [0.33, 0.42]	0.00 (0.00)	sparse (SMS holds)
synthetic k=8	0.67	0.69 [0.67, 0.75]	0.00 (0.00)	**dense (reject)**
Sachs: →Mek	0.09	0.76 [0.73, 0.82]	0.10 (0.27)	**dense (reject)**
Sachs: →PIP2	0.09	0.65 [0.55, 0.73]	0.19 (0.28)	**dense (reject)**
Sachs: →Akt	0.09	0.82 [0.73, 0.91]	0.13 (0.27)	**dense (reject)**
Sachs: →PKA	0.09	0.80 [0.72, 0.91]	0.13 (0.28)	**dense (reject)**
Sachs: →PKC	0.09	1.00 [1.00, 1.00]	0.05 (0.18)	**dense (reject)**

4.6 Atomic versus fat-hand interventions

The Sachs rejection raises a mechanistic question: is a more *atomic* intervention actually sparser? This is the regime of Perturb-seq [30], where a CRISPR knockout

targets a single gene. Lacking the single-cell data here, we test the mechanism-level logic on a controlled SCM ($d = 12$, 6 seeds) by contrasting, on the *same* graph, an *atomic* intervention (change only the target's mechanism) against a *fat-hand* one (change the target and all its descendants, emulating pathway propagation). Table 5 shows the calibrated test calls the atomic intervention sparse in $6/6$ runs ($\hat{\rho} = 0.09$) and the fat-hand one dense in $6/6$ ($\hat{\rho} = 0.83$). This explains the Sachs result (chemical drugs are fat-hand) and yields a falsifiable prediction: a genuinely atomic intervention, such as a CRISPR knockout, should be sparser and more likely to satisfy SMS.

Table 5. Atomic versus fat-hand interventions on the same SCM ($d = 12$, 6 seeds). The calibrated test calls atomic sparse and fat-hand dense, every time.

Intervention	mean #changed	true ρ	$\hat{\rho}$	% sparse	% dense
atomic (target only)	1.0	0.08	0.09	100	0
fat-hand (target + descendants)	9.7	0.81	0.83	0	100

4.7 Extension to image data: dSprites and 3D Shapes

As a proof of concept we apply the image pipeline to dSprites [28], binary 64×64 images with known generative factors. We drive a $d{=}5$ factor SCM (shape, scale, orientation, x, y), render the corresponding real images, and plant k changed factor-mechanisms (known $S^\star$). With the oracle factors the verdict is correct in the sparse/dense sense but noisy at this low dimension (the count statistic $\hat{\rho} = \hat{k}/d$ is coarse at $d{=}5$), so we report changed-set F_1, which is more stable, and use the verdict only qualitatively.

We now examine how the frozen encoder g affects recovery (Table 6 and Figure 5). A CNN trained on the *source* environment only loses F_1 relative to the oracle, despite high reconstruction fidelity (target recon ≈ 0.87). This loss is *not* a fidelity problem: a controlled experiment that adds independent Gaussian noise of the same magnitude to the oracle factors leaves F_1 essentially unchanged at that recon level, so it is the encoder's *cross-environment systematic bias* (extrapolation onto the shifted target), not residual magnitude, that manufactures the false positives. Consistent with this, an encoder trained jointly on both environments (a domain-consistent encoder; recon ≈ 0.98) restores F_1 to the oracle. A frozen encoder thus preserves the coarse verdict, but accurate changed-set recovery requires a representation that is consistent (identifiable) across environments, not merely high-fidelity.

We repeat the comparison on 3D Shapes [29], RGB 64×64 images with six known factors (Table 6). Here the source-only frozen CNN *already* matches the oracle (F_1 within 0.01), leaving no gap for a domain-consistent encoder to close: the clean renderings let even a source-trained CNN reach recon ≈ 0.97–0.99 that transfers to the target. This *sharpens* rather than contradicts the earlier conclusion: the changed-set deficit is governed by the encoder's cross-environment *extrapolation error*, which is large on dSprites but small here, so domain-consistent training helps precisely when that error is large. As on dSprites, the low factor count and discretisation cap the oracle itself (e.g. $F_1 \approx 0.52$ for sparse shifts).

Table 6. Image data (dSprites, 3D Shapes): changed-set F_1 for oracle factors, a source-only frozen CNN, and a domain-consistent (jointly trained) frozen CNN, averaged over seeds.

Dataset	Shift	oracle F_1	source-only CNN F_1	domain-consistent CNN F_1
dSprites ($d{=}5$)	sparse	0.58	0.47	0.54
dSprites ($d{=}5$)	dense	0.80	0.66	0.80
3D Shapes ($d{=}6$)	sparse	0.52	0.51	0.53
3D Shapes ($d{=}6$)	dense	0.89	0.89	0.89

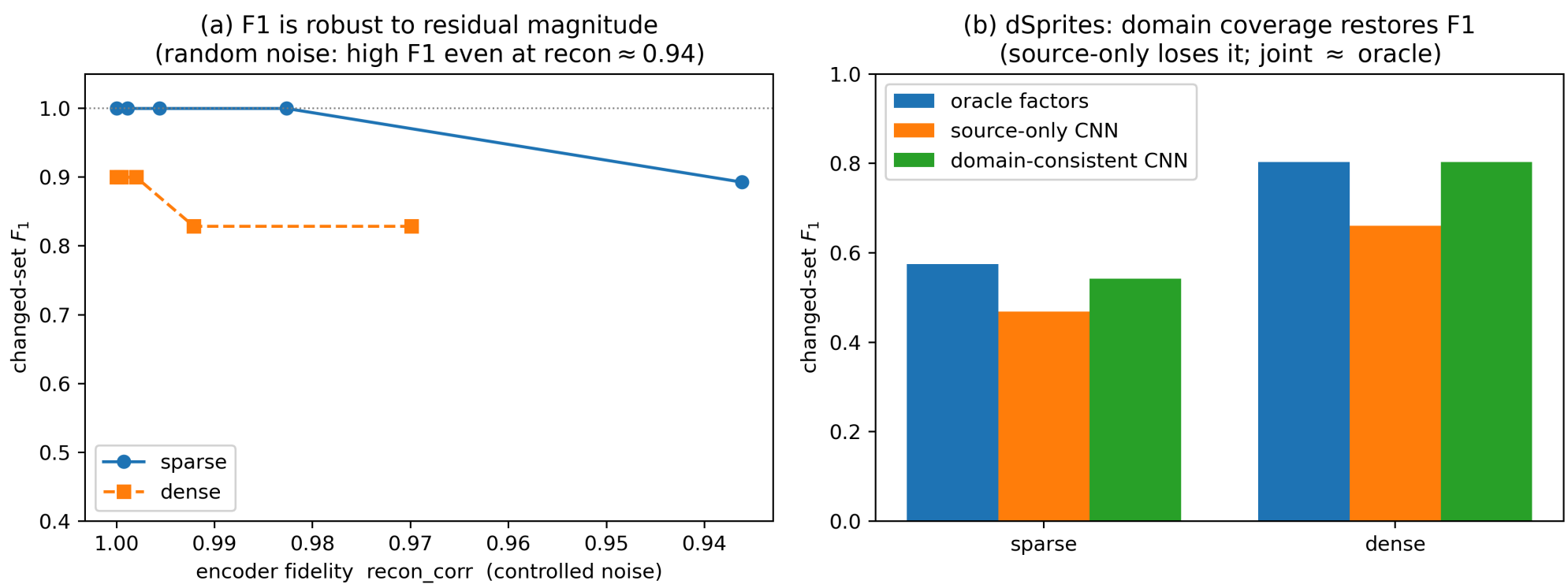

Figure 5. Image extension. (a) Under controlled independent noise on the factors, changed-set F1 stays high even at fidelity recon about 0.94, so the loss is not a residual-magnitude problem. (b) On dSprites, a source-only frozen CNN loses F1 while a domain-consistent (jointly trained) encoder restores it to the oracle: the deficit is a cross-environment bias, removed by domain coverage.

5. Discussion and Limitations

We have made the SMS premise testable: a graph-free, label-free detector that matches an oracle, and a hypothesis test with a data-driven null floor that accepts sparsity when it holds and rejects it when it does not, demonstrated to reject on real interventions and explained mechanistically by intervention atomicity. Several limitations remain. (i) The core test is linear-Gaussian with two environments; the kernel extension addresses nonlinearity but a full multi-environment treatment is

future work. (ii) The exists-invariant-set search is bounded in candidate-pool size and subset order, which caps recovery on very dense, high-dimensional shifts (e.g. ALARM at $k = 4$ dips to 0.88). (iii) The Sachs analysis treats discretised protein levels as continuous and adopts the standard intervention targets and a contested consensus graph. (iv) The sparse/dense cut-point τ_{dense} remains a semantic choice, although the lower null floor is calibrated; stronger FDR-style guarantees on the verdict are an open direction. (v) The atomic-versus-fat-hand study is a synthetic proxy for Perturb-seq; validating the prediction on real single-cell perturbation data is the natural next step. (vi) The image extension is a proof of concept: the disentanglement datasets we use (dSprites, 3D Shapes) have only $d{=}5{-}6$ factors, so the verdict is dimension-limited (we rely on F_1), and the domain-consistent encoder uses target factor labels as a domain-coverage upper bound. A label-free realisation we attempted with an identifiable VAE (with BCE/KL-annealing tuning, $K{=}16$ pseudo-environments, and a continuous-factor diagnostic) did not reach the labelled upper bound: its disentanglement (MCC) plateaus near 0.5, and the continuous-factor test indicates the bottleneck is the iVAE itself rather than discretisation; a stronger, intervention-tailored method (interventional causal representation learning [17, 20, 21]) is the natural route, which we leave to future work.

6. Conclusion

The sparse mechanism shift hypothesis underwrites a large body of causal approaches to distribution shift, yet is rarely tested. We provide the missing test: a precise diagnosis of why mechanism-shift detection fails under an estimated graph, an exists-invariant-set detector that resolves it, and a calibrated three-way SMS verdict. Across controlled, semi-synthetic, real-structure and real-intervention data the test behaves correctly, tracking sparsity when it holds and rejecting it when, as for real chemical interventions, the shift is dense. Making SMS falsifiable lets practitioners check the assumption their methods rely on, rather than assume it.

Footnotes

1. The diagnostic variants are named descriptively here. In our code and released results they correspond to labels A (oracle graph), C (pooled-PC), C5–C8 (the ablations), and C10 (the exists-invariant-set detector, used as a shorthand in later sections). ↩

References

1. **Dataset Shift in Machine Learning**
   Quinonero-Candela, J., Sugiyama, M., Schwaighofer, A. and Lawrence, N.D., 2009. MIT Press.
2. **A unifying view on dataset shift in classification**
   Moreno-Torres, J.G., Raeder, T., Alaiz-Rodriguez, R., Chawla, N.V. and Herrera, F., 2012. Pattern Recognition, Vol 45(1), pp. 521--530. DOI: 10.1016/j.patcog.2011.06.019
3. **Causality: Models, Reasoning, and Inference**
   Pearl, J., 2009. Cambridge University Press.
4. **Toward Causal Representation Learning**
   Scholkopf, B., Locatello, F., Bauer, S., Ke, N.R., Kalchbrenner, N., Goyal, A. and Bengio, Y., 2021. Proceedings of the IEEE, Vol 109(5), pp. 612--634. DOI: 10.1109/JPROC.2021.3058954
5. **Causal Discovery in Heterogeneous Environments Under the Sparse Mechanism Shift Hypothesis**
   Perry, R., von Kugelgen, J. and Scholkopf, B., 2022. Advances in Neural Information Processing Systems (NeurIPS), Vol 35.
6. **Causal Discovery from Heterogeneous/Nonstationary Data**
   Huang, B., Zhang, K., Zhang, J., Ramsey, J., Sanchez-Romero, R., Glymour, C. and Scholkopf, B., 2020. Journal of Machine Learning Research, Vol 21(89), pp. 1--53.
7. **Causal inference by using invariant prediction: identification and confidence intervals**
   Peters, J., Buhlmann, P. and Meinshausen, N., 2016. Journal of the Royal Statistical Society: Series B (Statistical Methodology), Vol 78(5), pp. 947--1012. DOI: 10.1111/rssb.12167
8. **Preventing Failures Due to Dataset Shift: Learning Predictive Models That Transport**
   Subbaswamy, A., Schulam, P. and Saria, S., 2019. Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS), pp. 3118--3127.
9. **Anchor Regression: Heterogeneous Data Meet Causality**
   Rothenhausler, D., Meinshausen, N., Buhlmann, P. and Peters, J., 2021. Journal of the Royal Statistical Society: Series B (Statistical Methodology), Vol 83(2), pp. 215--246. DOI: 10.1111/rssb.12398
10. **Causal Protein-Signaling Networks Derived from Multiparameter Single-Cell Data**
    Sachs, K., Perez, O., Pe'er, D., Lauffenburger, D.A. and Nolan, G.P., 2005. Science, Vol 308(5721), pp. 523--529. DOI: 10.1126/science.1105809
11. **Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing**
    Benjamini, Y. and Hochberg, Y., 1995. Journal of the Royal Statistical Society: Series B (Methodological), Vol 57(1), pp. 289--300. DOI: 10.1111/j.2517-6161.1995.tb02031.x
12. **Invariant Risk Minimization**
    Arjovsky, M., Bottou, L., Gulrajani, I. and Lopez-Paz, D., 2019. arXiv preprint arXiv:1907.02893.
13. **Causality-oriented robustness: exploiting general additive interventions**
    Shen, X., Buhlmann, P. and Taeb, A., 2023. arXiv preprint arXiv:2307.10299.
14. **Estimating and Explaining Model Performance When Both Covariates and Labels Shift**
    Chen, L., Zaharia, M. and Zou, J., 2022. Advances in Neural Information Processing Systems (NeurIPS), Vol 35.
15. **Explaining Concept Shift with Interpretable Feature Attribution**
    Lyu, R., Turcan, A. and Wilder, B., 2025. arXiv preprint arXiv:2505.20634.
16. **Panning for Gold: Model-X Knockoffs for High-Dimensional Controlled Variable Selection**
    Candes, E., Fan, Y., Janson, L. and Lv, J., 2018. Journal of the Royal Statistical Society: Series B (Statistical Methodology), Vol 80(3), pp. 551--577. DOI: 10.1111/rssb.12265
17. **Interventional Causal Representation Learning**
    Ahuja, K., Mahajan, D., Wang, Y. and Bengio, Y., 2022. Advances in Neural Information Processing Systems (NeurIPS).

18. **Weakly Supervised Causal Representation Learning**
    Brehmer, J., De Haan, P., Lippe, P. and Cohen, T., 2022. Advances in Neural Information Processing Systems (NeurIPS).

19. **Linear Causal Disentanglement via Interventions**
    Squires, C., Seigal, A., Bhate, S.S. and Uhler, C., 2023. Proceedings of the 40th International Conference on Machine Learning (ICML), Vol 202, pp. 32540--32560.

20. **Learning Linear Causal Representations from Interventions under General Nonlinear Mixing**
    Buchholz, S., Rajendran, G., Rosenfeld, E., Aragam, B., Scholkopf, B. and Ravikumar, P., 2023. Advances in Neural Information Processing Systems (NeurIPS).

21. **Nonparametric Identifiability of Causal Representations from Unknown Interventions**
    von Kugelgen, J., Besserve, M., Wendong, L., Gresele, L., Kekic, A., Bareinboim, E., Blei, D.M. and Scholkopf, B., 2023. Advances in Neural Information Processing Systems (NeurIPS).

22. **Identifying Representations for Intervention Extrapolation**
    Saengkyongam, S., Rosenfeld, E., Ravikumar, P., Pfister, N. and Peters, J., 2024. International Conference on Learning Representations (ICLR).

23. **Causation, Prediction, and Search**
    Spirtes, P., Glymour, C. and Scheines, R., 2000. MIT Press.

24. **Stability Selection**
    Meinshausen, N. and Buhlmann, P., 2010. Journal of the Royal Statistical Society: Series B (Statistical Methodology), Vol 72(4), pp. 417--473. DOI: 10.1111/j.1467-9868.2010.00740.x

25. **Kernel-based Conditional Independence Test and Application in Causal Discovery**
    Zhang, K., Peters, J., Janzing, D. and Scholkopf, B., 2011. Proceedings of the 27th Conference on Uncertainty in Artificial Intelligence (UAI), pp. 804--813.

26. **Retiring Adult: New Datasets for Fair Machine Learning**
    Ding, F., Hardt, M., Miller, J. and Schmidt, L., 2021. Advances in Neural Information Processing Systems (NeurIPS), Vol 34, pp. 6478--6490.

27. **Learning {B}ayesian Networks with the {bnlearn} {R} Package**
    Scutari, M., 2010. Journal of Statistical Software, Vol 35(3), pp. 1--22. DOI: 10.18637/jss.v035.i03

28. **{dSprites}: Disentanglement Testing Sprites Dataset**
    Matthey, L., Higgins, I., Hassabis, D. and Lerchner, A., 2017.

29. **{3D} {S}hapes Dataset**
    Burgess, C. and Kim, H., 2018.

30. **Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens**
    Dixit, A., Parnas, O., Li, B., Chen, J., Fulco, C.P., Jerby-Arnon, L., Marjanovic, N.D. and Regev, A., 2016. Cell, Vol 167(7), pp. 1853--1866. DOI: 10.1016/j.cell.2016.11.038

For attribution in academic contexts, please cite this work as

```
Anonymous, "Calibrated, Falsifiable Detection of Sparse Mechanism Shift", TMLR Beyond PDF, 2026.
```

BibTeX citation

```
@inproceedings{anonymous2026calibratedfalsifiabledetection,
  author = {Anonymous, },
  title = {Calibrated, Falsifiable Detection of Sparse Mechanism Shift},
  booktitle = {TMLR Beyond PDF},
  year = {July 5, 2026}
}
```