Partial Causal Structure Learning for Valid Selective Conformal Inference under Interventions

Amir Asiaee[1] **Kaveh Aryan**[2] **James P. Long**[3]

[1]Department of Biostatistics, Vanderbilt University Medical Center, TN 37203, USA
[2]Department of Informatics, King's College London, London, UK
[3]Department of Biostatistics, MD Anderson Cancer Center, Houston, TX, USA

Abstract

Selective conformal prediction can yield substantially tighter uncertainty sets when we can identify calibration examples that are exchangeable with the test example. In interventional settings, such as perturbation experiments in genomics, exchangeability often holds only within subsets of interventions that leave a target variable "unaffected" (e.g., non-descendants of an intervened node in a causal graph). We study the practical regime where this invariance structure is unknown and must be estimated from data. Our main result quantifies how coverage degrades when the estimated safe calibration set accidentally includes interventions that affect the target, and gives a conservative correction when an upper bound on this error is available. Rather than learning a full causal graph, we learn only the intervention–target relationships needed to choose calibration interventions. We give algorithms for this partial learning task and evaluate them on synthetic structural equation models and Replogle K562 CRISPR-interference data, where the experiments illustrate synthetic gains from selective calibration and finite-sample tradeoffs on real perturbation screens.

1 INTRODUCTION

Conformal prediction (CP) provides distribution-free uncertainty quantification: given any black-box predictor, CP constructs prediction sets with finite-sample marginal coverage guarantees under exchangeability [Vovk et al., 2005, Shafer and Vovk, 2008, Angelopoulos and Bates, 2023]. Split conformal prediction, in particular, requires only a single pass over a held-out calibration set and scales to modern high-dimensional problems.

However, the marginal coverage guarantee of standard CP can be loose when the data exhibit heterogeneity. In many scientific applications, data are generated under multiple interventions or environments, and exchangeability holds *only within* certain subsets of these conditions. If we can identify such subsets, *selective* (or *Mondrian*) calibration [Vovk, 2012, Boström et al., 2021] provides the same coverage guarantee but with prediction sets that are often substantially tighter, because calibration is restricted to the relevant stratum.

Motivating application: gene perturbation experiments. Single-cell perturbation screens such as Perturb-seq [Dixit et al., 2016, Replogle et al., 2022] now routinely profile the transcriptional consequences of hundreds to thousands of genetic interventions (e.g., CRISPRi knockdowns) in a single experiment. A central scientific question is: for a given target gene i, which interventions affect its expression and by how much? Under a causal model, intervening on gene a changes the distribution of gene i if and only if i is a causal descendant of a in the gene regulatory network.

For non-descendant targets, the residual distribution under intervention a is the same as under control, creating a natural exchangeability structure. Exploiting this structure via selective conformal inference can provide tighter, more informative prediction intervals for the effects of unseen interventions, enabling better prioritization in experimental design and more confident in silico differential expression calls.

The challenge is that the descendant structure is rarely known. Full causal graph learning in high dimensions is notoriously difficult and computationally expensive [Peters et al., 2017, Hauser and Bühlmann, 2012], and the resulting errors propagate to the selective calibration procedure in ways that are poorly understood. We quantify the statistical cost of imperfect causal knowledge for conformal inference and develop algorithms that learn only the partial causal structure needed to choose calibration interventions.

Contributions. Our contributions are:

Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence (UAI 2026), PMLR 337:161–184.

1. **Coverage under imperfect stratum selection.** We prove a finite-sample coverage bound for selective conformal prediction when some selected calibration interventions come from the wrong causal stratum, and we give a conservative level correction when the amount of such error can be bounded.
2. **Partial causal learning for calibration.** Instead of recovering the full causal graph, we reduce the learning problem to deciding which calibration interventions can safely be used for each target. This makes the relevant error asymmetric: discarding a safe calibration point mainly costs efficiency, while admitting an intervention that affects the target can hurt coverage.
3. **Algorithms and empirical validation.** We give a descendant-discovery procedure based on intersections of differentially affected sets and a local invariant-causal-prediction procedure for estimating distance to an intervention, with recovery conditions for both. On synthetic linear SEMs, coverage degrades monotonically as the selected calibration set is contaminated, and the corrected procedure restores conservative coverage when supplied with the realized contamination level. On Replogle K562 CRISPRi data, the corrected method exceeds nominal coverage among the causal-selector methods on its feasible cases, while a weighted CP baseline using control-cell coexpression provides a conservative full-feasibility comparison.

Organization. Section 2 discusses related work. Section 3 formalizes the setup. Section 4 formulates the partial causal learning objective. Section 5 presents the δ-robustness theorem. Section 6 describes the algorithms. Section 7 states recovery conditions. Section 8 details the experimental evaluation. Section 9 concludes. All proofs are deferred to the Supplementary Material.

2 RELATED WORK

Conformal prediction: foundations and efficiency. Conformal prediction originated in the algorithmic learning theory literature [Vovk et al., 2005] and has become a standard approach to distribution-free uncertainty quantification [Shafer and Vovk, 2008, Angelopoulos and Bates, 2023]. Conformalized quantile regression [Romano et al., 2019] improves efficiency by adapting interval widths to the input, while maintaining the finite-sample coverage guarantee.

Conditional and Mondrian conformal prediction. Standard CP provides only marginal coverage; conditional coverage (coverage conditional on X) is generally impossible without strong assumptions [Vovk, 2012]. Mondrian conformal prediction [Boström et al., 2021] provides a practical middle ground: it stratifies calibration points into categories and applies CP within each stratum, achieving stratum-conditional coverage. Gibbs et al. [2025] develop a framework interpolating between marginal and conditional validity, showing how to achieve coverage guarantees over finite collections of subgroups. Our selective calibration is Mondrian-like, with strata defined by causal invariances rather than observable features.

Conformal prediction under distribution shift. When exchangeability fails, weighted conformal prediction can restore validity under covariate shift if the density ratio is known [Tibshirani et al., 2019]. Barber et al. [2023] develop conformal prediction beyond exchangeability, providing a general framework for weighted and non-exchangeable settings with explicit finite-sample coverage bounds. The contamination model we study is complementary: rather than a continuous distributional shift, we analyze the discrete setting where a fraction of calibration points come from the wrong stratum.

Robust conformal prediction under contamination. Einbinder et al. [2024] study label noise robustness, showing that conformal prediction is conservative when noise is dispersive and providing corrections for adversarial noise of bounded size. Clarkson et al. [2024] study split conformal prediction under a Huber contamination model, providing coverage bounds in terms of the contamination fraction and the Kolmogorov–Smirnov distance. Gendler et al. [2022] address adversarial input perturbations via randomized smoothing. Our Theorem 1 is tailored to the *selective* (Mondrian) setting where contamination arises from misclassification of calibration strata, which is the natural error mode when causal structure is learned. The key distinction is that we bound coverage loss as a function of the fraction of miscategorized calibration points, without assumptions on the contaminating distribution.

Conformal prediction and causal inference. Lei and Candès [2021] develop conformal inference of counterfactuals and individual treatment effects in the potential outcomes framework, providing finite-sample coverage under unconfoundedness. Alaa et al. [2023] extend this to conformal meta-learners for predictive inference of individualized treatment effects. By contrast, we use causal structure to identify approximately exchangeable calibration subsets across *multiple simultaneous interventions*, enabling tighter predictive intervals for post-intervention outcomes rather than performing inference on a treatment effect.

Causal discovery from interventional data. Hauser and Bühlmann [2012, 2015] characterize interventional Markov equivalence classes and develop greedy algorithms for learning directed acyclic graphs (DAGs) from mixed data. Eberhardt et al. [2005] establish that $\lceil \log_2(N) \rceil + 1$ experiments suffice to identify the full causal graph over N variables in the worst case, motivating efficient experimental design. Squires et al. [2020] propose active structure learning via directed clique trees, achieving near-optimal intervention counts for full graph identification. Brouillard et al. [2020]

develop differentiable causal discovery methods that scale to larger graphs using interventional data. These works aim to recover the complete DAG; our approach instead learns the partial structure needed for selective conformal calibration.

Invariant causal prediction. Invariant causal prediction (ICP) [Peters et al., 2016] identifies parents of a target by exploiting the stability of the conditional distribution $P(Y_i \mid X_{\mathrm{Pa}(i)})$ across environments. Nonlinear extensions have been developed by Heinze-Deml et al. [2018]. We adapt ICP ideas locally to estimate parent sets and thereby distance-to-intervention, without attempting full graph recovery.

Perturbation biology and causal gene networks. Single-cell perturbation screens provide a natural setting for interventional causal inference. Perturb-seq [Dixit et al., 2016] combines CRISPR perturbations with single-cell RNA-seq readouts. Replogle et al. [2022] scaled this to genome-wide screens. Peidli et al. [2024] provide harmonized perturbation datasets across studies. CausalBench [Chevalley et al., 2025] benchmarks gene network inference from perturbation data. Differential expression testing in perturbation screens raises statistical challenges, including the need for appropriate aggregation and multiple testing correction [Squair et al., 2021, Barry et al., 2024]. Differentially expressed gene (DEG) sets from perturbation screens serve as the inputs for our descendant discovery algorithm (Algorithm 1). Zhu et al. [2025] argue that in-silico perturbation models should be evaluated not only by global response accuracy but also by their ability to identify differentially expressed genes, proposing AUPRC as a biologically relevant metric for DE-gene recovery. This motivates uncertainty-quantified post-intervention predictions, since downstream biological interpretation often depends on confidence in DE calls rather than on point predictions alone.

3 INTERVENTIONAL PREDICTION AND SELECTIVE CONFORMAL

We consider a collection of observed interventions (environments) indexed by $a \in \mathcal{A}$. For split conformal prediction, we partition these observed interventions into disjoint training and calibration sets, $\mathcal{A} = \mathcal{A}_{\mathrm{train}} \cup \mathcal{A}_{\mathrm{cal}}$ and $\mathcal{A}_{\mathrm{train}} \cap \mathcal{A}_{\mathrm{cal}} = \emptyset$. For each observed intervention a we observe i.i.d. samples of outcomes $Y^{(a)} \in \mathbb{R}^p$ (e.g., gene expression vector) and possibly covariates $X^{(a)}$. Typically X represents shared sample-level information that does not depend on a (e.g., cell type, batch indicator, or technical covariates), though in some designs $X^{(a)}$ may include intervention-specific features such as dosage or timing. We focus on predicting a scalar target component $Y_i^{(a)}$ (or a derived effect size) for each $i \in [p]$, e.g., the expression of a single gene i under intervention a.

Let $\hat{f}_i(a, X)$ be a fitted predictor of Y_i in intervention a, learned by an arbitrary predictor-learning algorithm $\mathcal{L}_i$ from the training interventions $\mathcal{A}_{\mathrm{train}}$. We use a calibration score (nonconformity score)

$$R_i^{(a)} = S_i\big(X^{(a)}, Y_i^{(a)}\big) \equiv \big|Y_i^{(a)} - \hat{f}_i(a, X^{(a)})\big|.$$

Calibration scores are computed only for interventions $a \in \mathcal{A}_{\mathrm{cal}}$. The test intervention satisfies $a^\star \notin \mathcal{A}$ and plays the same role as a new test point in ordinary split conformal prediction. For example, if $\mathcal{A}_{\mathrm{train}} = \{a_1, a_2\}$, $\mathcal{A}_{\mathrm{cal}} = \{a_3, a_4, a_5\}$, and $a^\star = a_6$, then $\hat{f}_i$ is learned using a_1, a_2, conformal scores are computed only from a_3, a_4, a_5, and a_6 is the held-out test intervention.

Running genomics example. In a single-cell perturbation screen, an intervention label a denotes a perturbation condition, such as CRISPR interference (CRISPRi) knockdown, CRISPR activation (CRISPRa), or CRISPR-Cas9 knockout of a gene; the perturbation modality can be included as part of the label a. For target gene i, $Y_i^{(a)}$ is the expression of gene i measured after applying intervention a, and $X^{(a)}$ contains sample-level covariates such as cell type, batch, donor, time point, or dose. The prediction model $\hat{f}_i(a, x)$ answers the question: if we apply intervention a in context x, what expression should we expect for gene i? It is learned from prior interventional data in $\mathcal{A}_{\mathrm{train}}$, while $\mathcal{A}_{\mathrm{cal}}$ is reserved for estimating the residual scale $R_i^{(a)} = |Y_i^{(a)} - \hat{f}_i(a, X^{(a)})|$. The new intervention $a^\star$ is then a held-out perturbation condition, for example a CRISPRi knockdown not used for either training or calibration. If we fix a cell type or context $X = c$, the notation reduces to asking how the predicted expression of gene i changes as only the perturbation label a varies. Our theory and experiments treat a as a single intervention condition; extending the descendant-learning and coverage analysis to combinatorial perturbations, such as knockdown of two genes, is left for future work.

3.1 SELECTIVE INVARIANCE SETS

Suppose that for each intervention $a \in \mathcal{A} \cup \{a^\star\}$ and target i there exists an unknown indicator $Z_{a,i} = \mathbf{1}\{i \in \mathrm{desc}(a)\}$, where $\mathrm{desc}(a)$ denotes causal descendants of a in an underlying causal graph $G = (V, E)$. We assume $a \in \mathrm{desc}(a)$, i.e., $Z_{a,a} = 1$. We assume a single graph G shared across all samples; extending to covariate-dependent graphs (e.g., cell-type-specific regulatory networks) is an important direction for future work. Define the *safe calibration set* for target i as $\mathcal{A}^\star(i) = \{a \in \mathcal{A}_{\mathrm{cal}} : Z_{a,i} = 0\}$ (i.e., calibration interventions that do not affect i).

Assumption 1 (Selective exchangeability)**.** Fix a target gene i and a test intervention $a^\star$ with $Z_{a^\star,i} = 0$ (i.e., i is unaffected by $a^\star$). The calibration scores $\{R_i^{(a)} : a \in \mathcal{A}^\star(i)\} \cup \{R_i^{(a^\star)}\}$ are exchangeable.

If $Z_{a^\star,i} = 0$, we calibrate using interventions where i is unaffected, yielding valid, typically tighter intervals than

a pooled baseline mixing interventions that do and do not affect target i. Pooling is justified only for a coarser mixture-level target, e.g., when the test intervention is drawn exchangeably from the same mixture, or when pooled scores are otherwise exchangeable. When $Z_{a^\star,i} = 1$, selective calibration on the safe set does not apply; one needs an exchangeable affected stratum, or else pooled CP requires an explicit mixture-exchangeability assumption. **Our framework focuses on $Z_{a^\star,i} = 0$, where selective calibration offers the clearest gains; descendant-target strata are left for future work.**

Example 1 (Why selective calibration tightens intervals). Consider a gene regulatory network with 100 genes and 50 interventions. For a target gene i, suppose only 5 of the 50 interventions are ancestors of i (i.e., affect i). Pooled conformal prediction uses all 50 residuals, mixing 5 residuals from interventions that affect i with 45 residuals from safe interventions. Selective conformal uses only the 45 safe residuals. When the residuals from interventions that affect i are larger, as can occur for predictors that do not fully model intervention effects, they inflate the pooled quantile; the selective interval is then tighter while maintaining exact $1 - \alpha$ coverage.

3.2 SELECTIVE SPLIT CONFORMAL INTERVAL (ORACLE)

Given a set of calibration scores $\{R_i^{(a)} : a \in \mathcal{A}^\star(i)\}$ of size m, define the oracle conformal quantile $k = \lceil (m+1)(1-\alpha) \rceil$ as:

$$\hat{q}_i^{\text{oracle}}(1-\alpha) = \text{the } k\text{-th smallest value in } \{R_i^{(a)} : a \in \mathcal{A}^\star(i)\} \cup \{\infty\}.$$

The prediction interval is

$$C_{i,1-\alpha}^{\text{oracle}}(a^\star, x) = [\, \hat{f}_i(a^\star, x) - \hat{q}_i^{\text{oracle}}(1-\alpha), \hat{f}_i(a^\star, x) + \hat{q}_i^{\text{oracle}}(1-\alpha)\,].$$

Under Assumption 1, this interval achieves $1 - \alpha$ marginal coverage by the standard conformal argument [Shafer and Vovk, 2008].

4 PARTIAL CAUSAL LEARNING FOR CALIBRATION

The oracle selective interval of Section 3 assumes the exchangeable set $\mathcal{A}^\star(i)$ is known. In practice it is unknown and must be estimated. Rather than learn the full causal graph, we estimate only *the binary descendant indicators* needed to select calibration interventions, and keep the resulting selection error small. Section 5 then quantifies exactly how that error translates into a coverage cost, closing the loop.

4.1 BINARY CLASSIFICATION VIEW

For each intervention a and target i, we estimate $\widehat{Z}_{a,i} \approx Z_{a,i} = \mathbf{1}\{i \in \text{desc}(a)\}$. We then select calibration interventions for $(i, a^\star)$ as $\widehat{\mathcal{A}}(i, a^\star) = \{a \in \mathcal{A}_{\text{cal}} : \widehat{Z}_{a,i} = 0\}$, possibly intersected with design strata within which exchangeability is more plausible (for example, the same cell type, experimental context, or batch).

4.2 FROM CLASSIFIER ERRORS TO CONTAMINATION

Let π_0 be the fraction of truly safe interventions in $\mathcal{A}_{\text{cal}}$ for target i, meaning interventions that do not affect i. We use the *descendant-positive* convention throughout, so a "positive" is a pair where intervention a affects target i, i.e., $Z_{a,i} = \mathbf{1}\{i \in \text{desc}(a)\} = 1$. Let $\text{FNR}_i = \mathbb{P}(\widehat{Z}_{a,i} = 0 \mid Z_{a,i} = 1)$ be the fraction of calibration interventions that truly affect i but are misclassified as safe (the error that *contaminates* the selected calibration set), and let $\text{FPR}_i = \mathbb{P}(\widehat{Z}_{a,i} = 1 \mid Z_{a,i} = 0)$ be the fraction of truly safe interventions misclassified as affecting i, which only reduces the calibration set size. Then the contamination fraction δ (the share of the selected set that is not exchangeable, made precise in Section 5) satisfies

$$\delta = \frac{(1-\pi_0)\text{FNR}_i}{\pi_0(1-\text{FPR}_i) + (1-\pi_0)\text{FNR}_i}.$$

This expression shows that *controlling the false-negative rate* FNR_i *is critical*: a missed descendant admits a non-exchangeable score from an intervention that affects i into the calibration set. A classifier that conservatively labels borderline cases as "affecting i" (accepting higher FPR_i, which only reduces the calibration set size) is preferable to one that aggressively labels cases as safe (low FPR_i but higher FNR_i, which increases contamination). For the hard safe-set selector above, this contamination fraction is primarily target-specific, so one may write δ_i; we keep the notation δ after fixing the query. It becomes genuinely $(i, a^\star)$-specific when the selected calibration set also depends on the test intervention, for example through same-context restrictions or distance-based weighting.

Remark 1 (Asymmetric error costs). In sparse networks, most interventions do not affect a given target (π_0 is large), so even a moderate FNR_i can lead to small δ. Indeed, when π_0 is sufficiently large, a naive strategy of labeling *all* calibration interventions as safe (i.e., $\text{FNR}_i = 1, \text{FPR}_i = 0$) yields $\delta = (1-\pi_0)/1 = 1 - \pi_0$; for example, with $\pi_0 \geq 0.95$ this gives $\delta \leq 0.05$, requiring no learning algorithm at all. Algorithm 1 improves upon this baseline by further reducing δ while maintaining calibration set size. However, δ grows rapidly with FNR_i when π_0 is small (dense networks or hub genes). This partial-label formulation makes this tradeoff explicit and motivates conservative classification strategies.

4.3 REDUCTION IN COMPLEXITY

Full causal graph learning over p variables requires estimating $O(p^2)$ edge parameters (or $O(2^p)$ DAG structures in the worst case). This objective reduces full graph learning to estimating $|\mathcal{A}| \times p$ binary labels, and in practice only a small fraction of these are needed (those for which the corresponding $(a^\star, i)$ pairs will be queried at test time). Furthermore, the labels $Z_{a,i}$ have combinatorial structure that can be exploited: if $a \to b \to c$ in G, then $Z_{a,c} = 1$ and $Z_{b,c} = 1$, enabling information sharing across interventions.

5 δ-ROBUSTNESS: COVERAGE UNDER CONTAMINATED SELECTIVE CALIBRATION

The partial-label selector of Section 4 produces the estimated calibration set $\widehat{\mathcal{A}}(i, a^\star) = \{a \in \mathcal{A}_{\text{cal}} : \widehat{Z}_{a,i} = 0\}$ (recall $\mathcal{A}_{\text{cal}}$ is the held-out calibration split, with $a^\star \notin \mathcal{A}_{\text{cal}}$). Because $\widehat{Z}$ is imperfect, some selected interventions violate exchangeability (they truly affect target i but are misclassified as safe), contaminating the calibration scores. We now quantify the coverage cost of this contamination as a function of its magnitude.

Definition 1 (Contamination fraction)**.** Let $\mathcal{A}^\star = \mathcal{A}^\star(i)$ be the (unknown) set of truly exchangeable safe calibration interventions for target i. Define the contamination fraction

$$\delta \equiv \frac{|\widehat{\mathcal{A}}(i, a^\star) \setminus \mathcal{A}^\star|}{|\widehat{\mathcal{A}}(i, a^\star)|}.$$

That is, δ is the fraction of calibration interventions in the selected set that actually affect target i (i.e., $Z_{a,i} = 1$) and therefore not exchangeable with the test score.

The selected set $\widehat{\mathcal{A}}$ naturally partitions into three regions (Figure 1): *good* interventions $\widehat{\mathcal{A}} \cap \mathcal{A}^\star$ (m truly safe and exchangeable with $a^\star$); *bad* interventions $\widehat{\mathcal{A}} \setminus \mathcal{A}^\star$ ($n - m$ contaminants that break exchangeability); and *missed* interventions $\mathcal{A}^\star \setminus \widehat{\mathcal{A}}$ (truly safe but not selected: wasted data that reduces calibration set size but does not harm coverage). The contamination fraction is $\delta = (n - m)/n$.

5.1 MAIN ROBUSTNESS THEOREM

We prove a finite-sample coverage lower bound that depends explicitly on δ and does *not* assume anything about the score distribution on contaminated interventions.

Theorem 1 (δ-robust selective conformal coverage)**.** *Fix $(i, a^\star)$ with $Z_{a^\star,i} = 0$, and suppose that scores from the "good" set $\mathcal{A}^\star$ are exchangeable with the test score $R_i^{(a^\star)}$ (Assumption 1). Let $\widehat{\mathcal{A}}$ be any selected calibration set of size $n = |\widehat{\mathcal{A}}|$, and let $m = |\widehat{\mathcal{A}} \cap \mathcal{A}^\star|$ be the number of good calibration interventions. Construct the split conformal interval using all scores in $\widehat{\mathcal{A}}$: let $\hat{q}$ be the $\lceil (n+1)(1-\alpha) \rceil$-th order statistic of $\{R_i^{(a)} : a \in \widehat{\mathcal{A}}\} \cup \{\infty\}$ and output $C_{i,1-\alpha} = [\hat{f}_i \pm \hat{q}]$. Then, for all possible distributions of contaminated scores,*

$$\mathbb{P}\big(R_i^{(a^\star)} \le \hat{q}\big) \;\ge\; 1 - \alpha - \frac{n-m}{m+1}.$$

In terms of the contamination fraction $\delta = (n-m)/n$,

$$\mathbb{P}\big(Y_i^{(a^\star)} \in C_{i,1-\alpha}\big) \ge 1 - \alpha - g(\delta, n),$$

where $g(\delta, n) \equiv \frac{\delta n}{(1-\delta)n+1}$ & $\lim_{n\to\infty} g(\delta, n) \approx \delta/(1-\delta)$.

Remark 2 (Interpretation and how to use it). Theorem 1 turns structure learning error into a direct statistical price for selective inference. If δ is small, coverage is close to nominal. The same bound can be used prospectively: one can design the learning procedure to keep δ below a threshold $\delta_{\max}$ such that $g(\delta_{\max}, n) \le \epsilon$ for a desired coverage slack ϵ. The formal α-correction based on an upper bound $\hat{\delta}$ is stated separately in Corollary 1.

Remark 3 (Comparison with Huber contamination bounds). Unlike the Huber model of Clarkson et al. [2024], where an ϵ-fraction of *all* calibration points is corrupted, here contamination arises from misclassification of the selective/Mondrian stratum: a calibration intervention that truly belongs outside the safe stratum for target i is incorrectly admitted into $\widehat{\mathcal{A}}$. For the hard safe-set selector, δ is target-specific (depends on i given the split); it becomes $(i, a^\star)$-specific only if selection also uses test-dependent strata such as cell type, batch, or distance to $a^\star$. Exploiting this selective structure yields a bound tailored to the actual selected calibration set rather than a generic Huber corruption model.

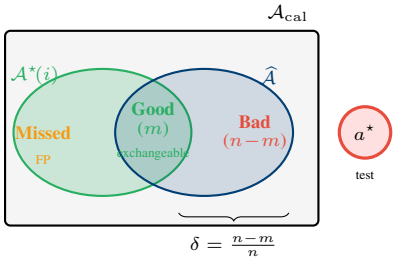
Remark 4 (Benign contamination). In many perturbation applications, interventions misclassified as safe will actually induce *larger* residuals (because the intervention does change the target), making $\hat{q}$ larger and the interval conservative. Theorem 1 is worst-case over contaminating distributions; empirical coverage is often better than the bound. Corollary 1 below concerns how to correct the nominal conformal level once an upper bound on δ is available.

Corollary 1 (Corrected selective conformal)**.** *If $\hat{\delta}$ is an upper bound on the contamination fraction δ (possibly estimated or given by the recovery conditions of Propositions 1–2), then running split conformal with nominal level $\alpha' = \alpha - g(\hat{\delta}, n)$ yields coverage at least $1 - \alpha$ for the selective interval (when $\alpha' \le 0$, the interval is $(-\infty, \infty)$ and coverage is trivially 1).*

6 ALGORITHMS

We present two practical estimators from interventional data: one for $\widehat{Z}_{a,i}$ and one for a distance-to-intervention score

(a) Partition of calibration interventions



(b) Concrete example on a causal DAG

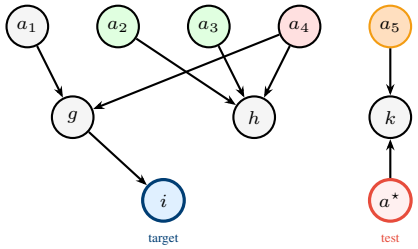


Figure 1: Intervention-set geometry for target i. **(a)** Within the calibration split, $\mathcal{A}^\star(i)$ is the truly safe set and $\widehat{\mathcal{A}}$ is the selected set; their overlap gives good scores, contaminants, and missed safe scores. Contaminants drive $\delta = (n-m)/n$ and reduce coverage, while missed safe scores only reduce calibration size. **(b)** Example where $\widehat{\mathcal{A}} = \{a_2, a_3, a_4\}$: a_2, a_3 are good, a_4 is a contaminant, and a_5 is missed.

$\widehat{d}(a, i)$, defined below as an estimated path length from intervention a to target i. Both are designed for scalability in high dimensions and many interventions.

6.1 DESCENDANT DISCOVERY VIA PERTURBATION INTERSECTION PATTERNS

The first algorithm estimates descendant sets $\widehat{\text{desc}}(a)$ from differentially affected variable sets across interventions, using set intersections to prune false positives.

Input: differentially affected sets. Assume we have a differentially affected set $\mathcal{S}_a \subset [p]$ for each intervention $a \in \mathcal{A}$, that is, an estimate of the variables whose distribution changes under a. In genomics these are differentially expressed gene (DEG) sets, computed by standard statistical tests (e.g., two-sample t-tests or Wilcoxon rank-sum tests with Benjamini–Hochberg correction [Squair et al., 2021, Barry et al., 2024]); in other domains any test for distributional change applies. $\mathcal{S}_a$ is an estimate of $\{i : Z_{a,i} = 1\} = \text{desc}(a)$, but may contain false positives and false negatives.

Intersection estimator. For each intervention a, identify interventions that are "upstream" of a, $\mathcal{U}(a) = \{b \in \mathcal{A} : a \in \mathcal{S}_b\}$ (i.e., interventions b such that a is affected by b). We estimate the descendant set of a by intersecting its own affected set with those of its upstream interventions:

$$\widehat{\text{desc}}(a) = \begin{cases} \mathcal{S}_a, & \mathcal{U}(a) = \emptyset, \\ \mathcal{S}_a \cap \bigcap_{b \in \mathcal{U}(a)} \mathcal{S}_b, & \text{otherwise.} \end{cases}$$

The intuition is: if b is upstream of a (i.e., $a \in \text{desc}(b)$), then every descendant of a is also a descendant of b, so $\text{desc}(a) \subseteq \text{desc}(b)$. Thus descendants of a should appear consistently across $\mathcal{S}_a$ and across upstream affected sets, while spurious entries will not. Equivalently, one can view this as starting from $\mathcal{S}_a$ and pruning any candidate variable that fails to appear in some upstream affected set; we present the set-intersection form for clarity.

Complexity. Algorithm 1 runs in $O(|\mathcal{A}|^2 \cdot p)$ time in the worst case (for each of $|\mathcal{A}|$ interventions, iterating over $O(|\mathcal{A}|)$ upstream interventions and $O(p)$ candidate genes). In practice, the estimator is implemented with set operations; since affected sets are typically sparse, the intersections are much cheaper than the worst-case bound. The expected cost is $O(|\mathcal{A}|^2\, \bar{s})$, where $\bar{s} = \mathbb{E}[|\mathcal{S}_a|]$ is the expected affected-set size; under the Replogle top-10% rule $\bar{s} \approx 500$ against $p \approx 5{,}000$, roughly a $10\times$ reduction over the worst case.

6.2 LOCAL ICP FOR DISTANCE ESTIMATION

Beyond binary descendant labels, selective inference can benefit from a *distance-to-intervention* notion $\hat{d}(a, i)$ that prioritizes closer calibration interventions or enables weighted conformal calibration. We propose a local search inspired by ICP [Peters et al., 2016] that traces back from the target i through estimated parents to obtain a coarse path-length estimate $\widehat{d}(a, i)$ without learning the full graph. The full algorithm (Algorithm 2) and details on using distance for kernel-weighted calibration are given in the Supplementary Material. The parent-pool construction and runtime accounting for Local ICP are deferred to Appendix F.

7 CONDITIONS FOR RECOVERY

This section states concrete, checkable conditions under which Algorithm 1 controls the contamination fraction δ.

Assumption 2 (Interventional faithfulness and detectability)**.** There is an underlying causal DAG $G = (V, E)$ such that intervening on node a changes (in distribution) precisely the descendants of a, i.e., for every $i \in \text{desc}(a)$ the marginal $P(Y_i \mid \text{do}(a))$ differs from $P(Y_i)$ (no path cancellations), and for every $i \notin \text{desc}(a)$ it is unchanged. Moreover, each descendant effect is detectable with probability at least $1 - \epsilon_{\text{fn}}$ by the testing procedure used to form $\mathcal{S}_a$. Formally: for each $a \in \mathcal{A}$ and $i \in \text{desc}(a)$, $\mathbb{P}(i \in \mathcal{S}_a) \geq 1 - \epsilon_{\text{fn}}$, and for each $i \notin \text{desc}(a)$, $\mathbb{P}(i \in \mathcal{S}_a) \leq \epsilon_{\text{fp}}$.

Algorithm 1 Descendant Discovery via Differentially Affected Set Intersections

Require: Differentially affected sets $\{\mathcal{S}_a\}_{a\in\mathcal{A}}$, interventions $\mathcal{A}$
Ensure: Estimated descendant sets $\{\widehat{\mathrm{desc}}(a)\}_{a\in\mathcal{A}}$

```
for each intervention a ∈ A do
    U(a) ← {b ∈ A : a ∈ S_b} {Upstream interventions of a}
    if U(a) = ∅ then
        Cand(a) ← S_a {Fallback to own affected set}
    else
        Cand(a) ← S_a ∩ ⋂_{b∈U(a)} S_b {Intersection with upstream affected sets}
    end if
    desc^(a) ← Cand(a)
end for
return {desc^(a)}_{a∈A}
```

Proposition 1 (Estimated descendant sets contain the true descendants)**.** *Under Assumption 2, for any intervention $a \in \mathcal{A}$, if $\mathcal{U}(a) \neq \emptyset$ and for every $b \in \mathcal{U}(a)$ we have $a \in \mathrm{desc}(b)$ (i.e., all identified upstream interventions are true ancestors), then with probability at least $1 - |\mathrm{desc}(a)| \cdot (|\mathcal{U}(a)|+1) \cdot \epsilon_{\mathrm{fn}}$ (by a union bound), $\mathrm{desc}(a) \subseteq \widehat{\mathrm{desc}}(a)$.*

Assumption 3 (Upstream diversity for intersections)**.** For any intervention $a \in \mathcal{A}$ and non-descendant $i \notin \mathrm{desc}(a)$, there exists an intervention $b \in \mathcal{A}$ such that $a \in \mathrm{desc}(b)$ but $i \notin \mathrm{desc}(b)$. Moreover, for such a b the testing procedure satisfies $\mathbb{P}(a \in \mathcal{S}_b \text{ and } i \notin \mathcal{S}_b) \geq 1 - \epsilon_{\mathrm{cx}}$.

Assumption 3 requires "upstream diversity": for each non-descendant i, there exists an ancestor of a whose descendants include a but not i. This is plausible when ancestors of a have sufficiently diverse descendant sets. Here ϵ_{cx} is a cross-screening failure probability: for the auxiliary intervention b, it is the chance that either a is not detected in $\mathcal{S}_b$ or the non-descendant i is spuriously included in $\mathcal{S}_b$. Under Assumption 2, a union bound gives $\epsilon_{\mathrm{cx}} \leq \epsilon_{\mathrm{fn}} + \epsilon_{\mathrm{fp}}$ for any fixed b satisfying the structural condition; we keep ϵ_{cx} separate because the joint event can be estimated or bounded directly. The final false-positive probability for the estimated descendant set $\widehat{\mathrm{desc}}(a)$ is bounded by both ϵ_{fp} and ϵ_{cx}, so we write $\epsilon_{\mathrm{sel}} = \min\{\epsilon_{\mathrm{fp}}, \epsilon_{\mathrm{cx}}\}$ for the sharper of these two bounds.

Proposition 2 (False positive control via upstream intersections)**.** *Under Assumptions 2–3, for any non-descendant $i \notin \mathrm{desc}(a)$, the intersection estimator in Algorithm 1 excludes i with probability at least $1 - \epsilon_{\mathrm{sel}}$. Consequently, the false positive rate satisfies $\mathbb{P}(\widehat{Z}_{a,i} = 1 \mid Z_{a,i} = 0) \leq \epsilon_{\mathrm{sel}}$.*

Corollary 2 (Contamination control via recovery conditions)**.** *Under Assumptions 2–3 and the upstream-ancestor condition of Proposition 1, the expected contamination fraction for target i and the calibration set $\widehat{A}(i, a^\star)$ satisfies*

$$\mathbb{E}[\delta] \leq \frac{(1-\pi_0)(\bar{u}+1)\epsilon_{\mathrm{fn}}}{\pi_0(1-\epsilon_{\mathrm{sel}}) + (1-\pi_0)(\bar{u}+1)\epsilon_{\mathrm{fn}}},$$

where π_0 is the fraction of safe calibration interventions for target i and $\bar{u} = \max_a |\mathcal{U}(a)|$ is the maximum upstream set size across interventions; the factor $\bar{u}+1$ counts the own affected set $\mathcal{S}_a$ together with the $|\mathcal{U}(a)|$ upstream sets in the union bound. Under the descendant-positive convention the contamination-driving false-negative rate $(\bar{u}+1)\epsilon_{\mathrm{fn}}$ appears in the numerator and the efficiency-cost false-positive rate ϵ_{sel} in the denominator; a full derivation is in Appendix D. In sparse networks where π_0 is near 1 and ϵ_{fn} is small, $\mathbb{E}[\delta]$ is small.

Together, Propositions 1–2 and Corollary 2 imply a tunable tradeoff between spurious descendants (false positives, which reduce calibration set size but do not violate coverage) and missed descendants (false negatives, which contaminate the calibration set), directly controlling the coverage loss via Theorem 1.

Assumption 3 (upstream diversity) is the strongest of the three, but Algorithm 1 degrades gracefully when it is only partially satisfied. When no upstream interventions are identified for a (i.e., $\mathcal{U}(a) = \emptyset$), the algorithm sets $\widehat{\mathrm{desc}}(a) = \mathcal{S}_a$ rather than failing, so Theorem 1 remains in force: a missing intersection step raises the required contamination upper bound $\hat{\delta}$ (and hence the width of the corrected interval) rather than silently breaking coverage. The false-positive control of Proposition 2 weakens for such a, but validity is preserved through the $\hat{\delta}$-correction of Corollary 1. The FNR bound is conditional on the identified upstream interventions being true ancestors; if spurious upstream interventions enter $\mathcal{U}(a)$, they can prune true descendants, so their probability must be included in any unconditional end-to-end recovery bound.

8 EXPERIMENTS

We evaluate the framework on synthetic and real data, focusing on three questions: (Q1) Does selective conformal prediction with learned $\widehat{Z}$ maintain valid coverage? (Q2) Does coverage degrade with contamination δ as predicted by Theorem 1? (Q3) Does the corrected procedure (Corollary 1) restore nominal coverage when supplied with realized or proxy contamination? In the experiments, Corrected uses the realized contamination fraction computed from true synthetic labels or real-data proxy labels, so these results test the correction formula given such a bound rather than a fully deployable $\hat{\delta}$ estimator.

Throughout, *Coverage* denotes the empirical fraction of test points whose true outcome falls within the prediction interval, and *Width* denotes the average interval length across test points. Both are computed over held-out (test intervention, target gene) pairs with $Z_{a^\star,i} = 0$, i.e. pairs the test intervention does not affect, for which a selective interval is meaningful. We treat coverage as a validity constraint rather than a monotone leaderboard score: values above 0.9

are acceptable, but higher coverage can simply mean more conservative intervals. In tables, bold coverage values meet or exceed the nominal 0.9 level. Among entries that meet this coverage threshold, bold width marks the narrowest interval and underlined width marks the second narrowest; widths for methods below nominal coverage are not treated as wins. Appendix H.1 gives procedural definitions of the table methods and additional baselines.

8.1 SYNTHETIC INTERVENTIONAL SEM

Setup. We generate random DAGs using the Erdős–Rényi model on $p = 200$ nodes with expected degree $d_{\text{avg}} = 2.0$. Edge weights are sampled uniformly from $[-1, -0.3] \cup [0.3, 1]$. The linear Gaussian SEM is $V = B^\top V + \varepsilon$ with $\varepsilon \sim \mathcal{N}(0, I)$, where $V \in \mathbb{R}^p$ is the vector of all node values. Interventions are hard (do-operator), setting $V_a := 0$. We use $n_{\text{obs}} = 200$ observational and $n_a = 200$ interventional samples per node, with $|\mathcal{A}| = 150$ randomly selected intervention targets. Ground-truth descendant sets $\text{desc}(a)$ are computed from the transitive closure of B.

Conformal scores. To isolate calibration-set selection from predictor quality, we use synthetic scores: safe pairs ($Z_{a,i} = 0$) receive $R_i^{(a)} = |N(0,1)|$, while pairs where intervention a affects target i receive $R_i^{(a)} = |N(0, 0.15)|$. These smaller contaminating scores pull the conformal quantile downward, causing under-coverage when the calibration set is contaminated by these affected interventions.

DEG procedure and descendant estimation. Gene-wise two-sample t-tests comparing interventional vs. observational data with Benjamini–Hochberg correction at $q = 0.05$ yield DEG sets $\mathcal{S}_a$, and Algorithm 1 estimates $\widehat{Z}_{a,i}$. Interventions are split 10%/81%/9% into training/calibration/test; $\widehat{Z}$ is fit without test interventions, and coverage is evaluated on held-out tests using calibration interventions. No outcome predictor is trained here: the training split feeds only descendant estimation (the DEG tests and Algorithm 1). Because synthetic scores are generated independently of DEG-based descendant estimation, selecting $\widehat{\mathcal{A}}$ does not create dependence between $\widehat{Z}$ and calibration scores.

Results. Table 1 summarizes results over 20 random seeds (220,132 evaluations). All four methods achieve coverage near the nominal $1 - \alpha = 0.9$ level. The realized contamination fraction, computed from the true DAG after Algorithm 1 selects the calibration set, is very low ($\delta = 0.018$), indicating that the intersection-based procedure keeps contamination well-controlled in the linear SEM setting. Because π_0 is high in this sparse network (most interventions do not affect a given target), the learned selector labels nearly all calibration interventions as safe, so $\widehat{\mathcal{A}} \approx \mathcal{A}_{\text{cal}}$ and the Estimated and Pooled methods coincide; their differences emerge under controlled contamination (Section 8.2). Corrected selective CP is the only non-oracle method above nominal coverage in this table, using the realized δ for the level correction; wider intervals reflect the cost of the α-correction even when contamination is small.

Table 1: Main synthetic experiment ($p = 200$, $|\mathcal{A}| = 150$, $\alpha = 0.1$). Coverage and width are averaged over 20 seeds. The realized δ is not applicable to Pooled, which applies no selection.

Method	Coverage	Width	n_{cal}	realized δ
Oracle	**0.901**	**3.35**	118.8	0.000
Estimated	0.899	3.32	121.0	0.018
Pooled	0.899	3.32	121.0	–
Corrected	**0.918**	<u>3.58</u>	121.0	0.018

8.2 CONTROLLED δ ABLATION

To test Theorem 1, we replace a target fraction $\delta_{\text{inject}} \in \{0, 0.05, 0.1, 0.15, 0.2, 0.3\}$ of scores in the true safe calibration set with scores from interventions that affect the target, resampling to achieve the target contamination rate. We use $p = 200$ nodes, $|\mathcal{A}| = 150$ interventions, and repeat over 100 random seeds (1,065,408 total evaluations). For finite-width reporting, we clip the corrected level below at $\alpha' = 0.01$; if $\alpha - g(\delta, n) \leq 0$, Corollary 1 would instead return the infinite interval.

Results. Figure 2 shows the results (full numerical breakdown in Table 3 in the Supplementary Material). Estimated coverage drops monotonically from 0.905 at $\delta = 0$ to 0.867 at $\delta = 0.30$, as predicted by Theorem 1. When supplied with the injected contamination level, the Corrected method empirically achieves coverage ≥ 0.95 at all nonzero contamination levels ($\delta \geq 0.05$), well above the nominal 0.9, at the cost of 1.2–1.8× wider intervals under this clipped finite-width implementation. Oracle and Pooled coverage remain flat, while empirical coverage always exceeds the theoretical lower bound $1 - \alpha - g(\delta, n)$ (Figure 3 in the Supplementary Material).

8.3 REAL PERTURBATION DATA: REPLOGLE K562 CRISPRI SCREEN

We apply the complete method to real data from the Replogle K562 CRISPRi genome-scale screen [Replogle et al., 2022], accessed via Zenodo. We select the 50 perturbations with the largest cell counts (≥ 200 cells each), retaining $p \approx 5{,}000$ genes expressed in $\geq 10\%$ of cells. Prior perturbation-screen benchmarks often use restricted intervention or gene subsets for scale [Sethuraman et al., 2023, Rohbeck et al., 2024, Roohani et al., 2024, Lopez et al., 2022], and we follow the same practical convention using a fixed cell-count-selected Replogle subset of 50 perturbations and ∼5,000 expressed genes. Log-fold-change (LFC) vectors are computed per perturbation against the

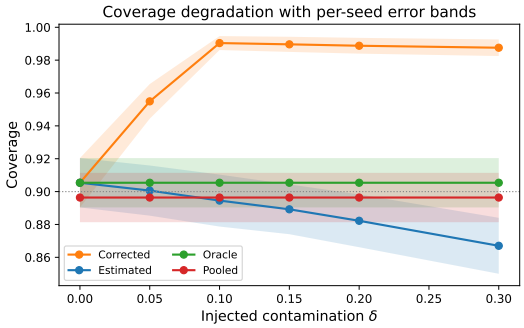


Figure 2: Coverage under injected contamination. Estimated degrades with δ, Corrected remains above nominal when supplied with the injected δ under the clipped finite-width implementation, and Oracle/Pooled are flat. Dashed line: $1 - \alpha = 0.9$; bands show ± 1 seed-level standard deviation.

Table 2: Real-data results on Replogle K562 CRISPRi. Coverage and width are averaged over all evaluations except Corrected, where they are averaged over feasible finite-interval cases, so its width is not directly comparable to full-feasibility rows. Corrected uses the proxy contamination fraction derived from LFC-based labels. Feasible is the fraction of finite intervals; only Corrected produces infinite intervals. For WCP, n_{cal} is the Kish effective sample size.

Method	Coverage	Width	n_{cal}	Feasible
Oracle (proxy)	0.864	0.306	36.7	100%
Estimated	0.888	0.349	40.0	100%
Pooled	0.888	0.349	40.0	100%
Corrected	**0.906**	0.329	40.0	59.8%
Ctrl-coexpr WCP	**0.925**	0.391	40.0	100%
Full-graph proxy	0.888	0.354	39.7	100%
Obs.-only sel.	0.888	0.349	40.0	100%

non-targeting control. Since ground-truth descendant sets are unavailable, we define proxy "oracle" affected sets as the top 10% of genes by absolute LFC per perturbation, and DEG sets analogously. Perturbations are split into 10% training, 81% calibration, and 9% test (5 test perturbations, ∼40 calibration perturbations). We fit $\widehat{Z}$ using only the training and calibration perturbations (holding out test perturbations from descendant discovery), and evaluate coverage on the held-out test perturbations using calibration perturbations for conformal calibration.

Results. Table 2 reports the primary four methods and additional baselines on the same split; Appendix H.1 defines all table methods procedurally. Among the causal-selector methods, Corrected is the only one above nominal coverage (0.906, averaged over feasible evaluations, those yielding a finite interval), using the proxy contamination fraction derived from LFC-based labels. However, it is finite for only 59.8% of evaluations; the rest are infinite because the clipped corrected level is often $\alpha' = 0.01$, requiring roughly 99 calibration scores, while this split has $n_{\text{cal}} \approx 40$. The proxy oracle undercovers (0.864), indicating proxy safe-set exchangeability violations. Weighted CP [Tibshirani et al., 2019] is implemented leakage-free: weights use only control-cell coexpression of perturbation targets, not held-out LFC response profiles. It reaches 0.925 coverage with full feasibility but wider intervals (0.391), so it trades width for feasibility relative to Corrected. The full-graph proxy and observational-only heuristic collapse to Estimated/Pooled behavior (0.888), supporting Algorithm 1 as the selector of choice in this sparse regime. Corrected is therefore above nominal coverage on the feasible proxy-labeled cases but not dominant on this real-data proxy.

Limitations of the real-data evaluation. The proxy oracle is constructed from LFC quantiles, not from true causal knowledge. Its coverage shortfall (0.864 vs. the nominal 0.9) indicates that the LFC-based proxy does not perfectly capture the exchangeability structure, likely because (i) indirect downstream effects create correlated residuals even among "unaffected" genes, and (ii) batch effects and technical noise break the i.i.d. assumption within strata. Moreover, on real data the proxy safe sets and residual scores are both computed from the same expression matrix, so the selective-exchangeability assumption should be read as an approximate diagnostic rather than a verified condition. A genuine oracle would require independently validated causal annotations (for example, curated knockout-effect databases or matched perturbation atlases) which are not available at the scale of 50 perturbations $\times$ 5,000 genes for K562; the controlled-δ synthetic experiment (Section 8.2) therefore remains the cleanest validation of the method itself. Despite these violations, within the causal-selector methods, the qualitative ordering (Corrected $>$ Estimated $\approx$ Pooled $>$ Oracle) is consistent with the theoretical predictions, and Corrected is the only one of these methods to exceed nominal coverage. The 59.8% feasibility is a finite-sample artifact of the small calibration set rather than a methodological limit: a scaling experiment on the same split (Appendix H, Figure 4) shows feasibility climbing monotonically to 100% by $n_{\text{cal}} = 160$ while proxy δ stays near 0.083. A more thorough real-data evaluation with larger calibration sets (more perturbations) and biologically validated oracle sets remains open.

9 CONCLUSION

We show that selective conformal calibration can exploit partial causal structure without requiring full graph recovery, and that the coverage cost of imperfect selection can be quantified directly through the contamination fraction of the selected calibration set. The synthetic experiments provide the cleanest validation of the theory, while the Replogle K562 experiment illustrates both the promise and the finite-sample limitations of applying the method with proxy causal labels. Open directions are collected in Appendix J.1.

Acknowledgements

A.A. was partially supported by the Patient-Centered Outcomes Research Institute (PCORI) award ME-2023C1-32148 and by the National Institutes of Health under award R01MH139379. J.P.L. was partially supported by the National Cancer Institute and the National Center for Advancing Translational Sciences of the National Institutes of Health under awards P50CA127001-16, CCSG P30CA016672-46, and CCTS UM1TR004906.

Reproducibility. All experiments, tables, and figures in this paper can be reproduced with the code, data-processing scripts, and notebooks available at `https://github.com/AsiaeeLab/selective-conformal-interventions`.

References

Ahmed M. Alaa, Zaid Ahmad, and Mark van der Laan. Conformal meta-learners for predictive inference of individual treatment effects. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Anastasios N. Angelopoulos and Stephen Bates. Conformal prediction: A gentle introduction. *Foundations and Trends in Machine Learning*, 16(4):494–591, 2023. doi: 10.1561/2200000101.

Rina Foygel Barber, Emmanuel J. Candès, Aaditya Ramdas, and Ryan J. Tibshirani. Conformal prediction beyond exchangeability. *The Annals of Statistics*, 51(2):816–845, 2023. doi: 10.1214/23-AOS2276.

Timothy Barry, Kaishu Mason, Kathryn Roeder, and Eugene Katsevich. Robust differential expression testing for single-cell CRISPR screens at low multiplicity of infection. *Genome Biology*, 25(1):124, 2024. doi: 10.1186/s13059-024-03254-2.

Henrik Boström, Ulf Johansson, and Tuwe Löfström. Mondrian conformal predictive distributions. In *Conformal and Probabilistic Prediction and Applications (COPA)*, volume 152 of *Proceedings of Machine Learning Research*, pages 24–38, 2021.

Philippe Brouillard, Sébastien Lachapelle, Alexandre Lacoste, Simon Lacoste-Julien, and Alexandre Drouin. Differentiable causal discovery from interventional data. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, 2020.

Mathieu Chevalley, Yusuf H. Roohani, Arash Mehrjou, Jure Leskovec, and Patrick Schwab. A large-scale benchmark for network inference from single-cell perturbation data. *Communications Biology*, 8(1):412, 2025. doi: 10.1038/s42003-025-07764-y.

Jason Clarkson, Wenkai Xu, Mihai Cucuringu, and Gesine Reinert. Split conformal prediction under data contamination. In *Conformal and Probabilistic Prediction and Applications (COPA)*, volume 230 of *Proceedings of Machine Learning Research*, pages 5–27, 2024.

Atray Dixit, Oren Parnas, Biyu Li, Jenny Chen, Charles P. Fulco, Livnat Jerby-Arnon, Nemanja D. Marjanovic, Danielle Dionne, Tyler Burks, Raktima Raychowdhury, Britt Adamson, Thomas M. Norman, Eric S. Lander, Jonathan S. Weissman, Nir Friedman, and Aviv Regev. Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens. *Cell*, 167(7):1853–1866.e17, 2016. doi: 10.1016/j.cell.2016.11.038.

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 (UAI)*, pages 178–184, 2005.

Bat-Sheva Einbinder, Shai Feldman, Stephen Bates, Anastasios N. Angelopoulos, Asaf Gendler, and Yaniv Romano. Label noise robustness of conformal prediction. *Journal of Machine Learning Research*, 25(1):1–66, 2024.

Asaf Gendler, Tsui-Wei Weng, Luca Daniel, and Yaniv Romano. Adversarially robust conformal prediction. In *International Conference on Learning Representations (ICLR)*, 2022.

Isaac Gibbs, John J. Cherian, and Emmanuel J. Candès. Conformal prediction with conditional guarantees. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 87(4):1100–1126, 2025. doi: 10.1093/jrsssb/qkaf008.

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

Alain Hauser and Peter Bühlmann. Jointly interventional and observational data: Estimation of interventional Markov equivalence classes of directed acyclic graphs. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 77(1):291–318, 2015. doi: 10.1111/rssb.12071.

Christina Heinze-Deml, Jonas Peters, and Nicolai Meinshausen. Invariant causal prediction for nonlinear models. *Journal of Causal Inference*, 6(2), 2018. doi: 10.1515/jci-2017-0016.

Lihua Lei and Emmanuel J. Candès. Conformal inference of counterfactuals and individual treatment effects. *Journal of the Royal Statistical Society: Series*

B (Statistical Methodology), 83(5):911–938, 2021. doi: 10.1111/rssb.12445.

Romain Lopez, Jan-Christian Hütter, Jonathan K. Pritchard, and Aviv Regev. Large-scale differentiable causal discovery of factor graphs. In *Advances in Neural Information Processing Systems 35 (NeurIPS 2022)*, volume 35, 2022.

Stefan Peidli, Tessa D. Green, Ciyue Shen, Torsten Gross, Joseph Min, Samuele Garda, Bo Yuan, Linus J. Schumacher, Jake P. Taylor-King, Debora S. Marks, Augustin Luna, Nils Blüthgen, and Chris Sander. scPerturb: harmonized single-cell perturbation data. *Nature Methods*, 21 (3):531–540, 2024. doi: 10.1038/s41592-023-02144-y.

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. doi: 10.1111/rssb.12167.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of Causal Inference: Foundations and Learning Algorithms*. MIT Press, 2017.

Joseph M. Replogle, Reuben A. Saunders, Angela N. Pogson, Jeffrey A. Hussmann, Alexander Lenail, Alina Guna, Lauren Mascibroda, Eric J. Wagner, Karen Adelman, Gila Lithwick-Yanai, Nika Iremadze, Florian Oberstrass, Doron Lipson, Jessica L. Bonnar, Marco Jost, Thomas M. Norman, and Jonathan S. Weissman. Mapping information-rich genotype-phenotype landscapes with genome-scale Perturb-seq. *Cell*, 185(14):2559–2575.e28, 2022. doi: 10.1016/j.cell.2022.05.013.

Martin Rohbeck, Brian Clarke, Katharina Mikulik, Alexandra Pettet, Oliver Stegle, and Kai Ueltzhöffer. Bicycle: Intervention-based causal discovery with cycles. In *Proceedings of the Third Conference on Causal Learning and Reasoning (CLeaR)*, volume 236 of *Proceedings of Machine Learning Research*, pages 209–242. PMLR, 2024.

Yaniv Romano, Evan Patterson, and Emmanuel J. Candès. Conformalized quantile regression. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019.

Yusuf Roohani, Kexin Huang, and Jure Leskovec. Predicting transcriptional outcomes of novel multigene perturbations with GEARS. *Nature Biotechnology*, 42(6):927–935, 2024. doi: 10.1038/s41587-023-01905-6.

Muralikrishnna G. Sethuraman, Romain Lopez, Rahul Mohan, Faramarz Fekri, Tommaso Biancalani, and Jan-Christian Hütter. NODAGS-Flow: Nonlinear cyclic causal structure learning. *arXiv preprint arXiv:2301.01849*, 2023.

Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9 (Mar):371–421, 2008.

Jordan W. Squair, Matthieu Gautier, Claudia Kathe, Mark A. Anderson, Nicholas D. James, Thomas H. Hutson, Rémi Hudelle, Taha Qaiser, Kaya J. E. Matson, Quentin Barraud, Ariel J. Levine, Gioele La Manno, Michael A. Skinnider, and Grégoire Courtine. Confronting false discoveries in single-cell differential expression. *Nature Communications*, 12(1):5692, 2021. doi: 10.1038/s41467-021-25960-2.

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 (NeurIPS)*, volume 33, 2020.

Ryan J. Tibshirani, Rina Foygel Barber, Emmanuel J. Candès, and Aaditya Ramdas. Conformal prediction under covariate shift. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019.

Vladimir Vovk. Conditional validity of inductive conformal predictors. In *Proceedings of the Asian Conference on Machine Learning (ACML)*, volume 25 of *Proceedings of Machine Learning Research*, pages 475–490, 2012.

Vladimir Vovk, Alex Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, New York, NY, 2005.

Hongxu Zhu, Amir Asiaee, Leila Azinfar, Jun Li, Han Liang, Ehsan Irajizad, Kim-Anh Do, and James P. Long. AUPRC: A metric for evaluating the performance of in-silico perturbation methods in identifying differentially expressed genes. *Briefings in Bioinformatics*, 26(5): bbaf426, 2025.

Partial Causal Structure Learning for Valid Selective Conformal Inference under Interventions (Supplementary Material)

Amir Asiaee[1] **Kaveh Aryan**[2] **James P. Long**[3]

[1]Department of Biostatistics, Vanderbilt University Medical Center, TN 37203, USA
[2]Department of Informatics, King's College London, London, UK
[3]Department of Biostatistics, MD Anderson Cancer Center, Houston, TX, USA

A PROOF OF THEOREM 1

Proof of Theorem 1. **Setup.** Fix target i and test intervention $a^\star$ with $Z_{a^\star,i} = 0$. Let $\widehat{\mathcal{A}}$ be the selected calibration set of size n. Partition $\widehat{\mathcal{A}}$ into $G = \widehat{\mathcal{A}} \cap \mathcal{A}^\star$ (good, $|G| = m$) and $B = \widehat{\mathcal{A}} \setminus \mathcal{A}^\star$ (bad, $|B| = n - m$). Let $k = \lceil (n+1)(1-\alpha) \rceil$ and let $\hat{q}$ be the k-th order statistic of all n calibration scores.

Step 1 (Worst-case bad scores). The adversary controls the values of the $n - m$ bad scores. To minimize $\hat{q}$ (and hence minimize coverage), the adversary places all bad scores at $-\infty$. Then the bottom $n - m$ positions in the sorted order are occupied by bad scores, so

$$\hat{q} \geq R^G_{k-(n-m)},$$

where $R^G_1 \leq \cdots \leq R^G_m$ are the good score order statistics. (If $k - (n-m) \leq 0$, then $\hat{q} \geq -\infty$, which is trivially true; if $k - (n-m) > m$, then $\hat{q} \geq \infty$, and coverage is 1.)

Step 2 (Exchangeability of good scores). The $m+1$ values $\{R^G_1, \ldots, R^G_m, R^{(a^\star)}_i\}$ are exchangeable by Assumption 1. By the fundamental property of exchangeable sequences (see Shafer and Vovk [2008, Lemma 1] or Vovk et al. [2005, Proposition 1]), the test score $R^{(a^\star)}_i$ is equally likely to occupy any of the $m+1$ ranks among these values. Therefore, for any integer $1 \leq r \leq m$,

$$\mathbb{P}\big(R^{(a^\star)}_i \leq R^G_r\big) \geq \frac{r}{m+1}.$$

Step 3 (Combining). Setting $r = k - (n-m)$ (assuming $1 \leq r \leq m$; the bound is trivial otherwise):

$$\mathbb{P}(R^{(a^\star)}_i \leq \hat{q}) \geq \mathbb{P}(R^{(a^\star)}_i \leq R^G_r) \geq \frac{k-(n-m)}{m+1}.$$

Step 4 (Algebra). Since $k = \lceil (n+1)(1-\alpha) \rceil \geq (n+1)(1-\alpha)$:

$$\begin{aligned}\frac{k-(n-m)}{m+1} &\geq \frac{(n+1)(1-\alpha)-(n-m)}{m+1} \\ &= \frac{m+1-(n+1)\alpha}{m+1} \\ &= 1 - \frac{(n+1)\alpha}{m+1}.\end{aligned}$$

The coverage gap beyond $1-\alpha$ is:

$$\begin{aligned}1-\frac{(n+1)\alpha}{m+1}-(1-\alpha) &= \alpha-\frac{(n+1)\alpha}{m+1}\\ &= \alpha\left(1-\frac{n+1}{m+1}\right)\\ &= -\alpha\cdot\frac{n-m}{m+1}.\end{aligned}$$

Since $\alpha \le 1$, we bound $\alpha\cdot\frac{n-m}{m+1} \le \frac{n-m}{m+1}$, giving

$$\mathbb{P}(R_i^{(a^\star)} \le \hat{q}) \ge 1-\alpha-\frac{n-m}{m+1}.$$

Substituting $n-m=\delta n$ and $m=(1-\delta)n$:

$$\frac{n-m}{m+1}=\frac{\delta n}{(1-\delta)n+1}=g(\delta,n).$$ □

Tightness. The bound is tight in the following sense: for any δ, n, α, there exist score distributions (with the bad scores placed adversarially at $-\infty$) such that the coverage equals exactly $\frac{k-(n-m)}{m+1}$, which matches the lower bound up to the ceiling in k.

B PROOF OF COROLLARY 1

Proof. By Theorem 1 applied with nominal level α', coverage is at least

$$\mathbb{P}(\ldots) \ge 1-\alpha'-g(\delta,n).$$

Since $\delta \le \hat{\delta}$ and $g(\delta,n)$ is nondecreasing in δ, we have $g(\delta,n) \le g(\hat{\delta},n)$, hence

$$\begin{aligned}\mathbb{P}(\ldots) &\ge 1-\alpha'-g(\hat{\delta},n)\\ &= 1-(\alpha-g(\hat{\delta},n))-g(\hat{\delta},n)\\ &= 1-\alpha.\end{aligned}$$

□

C PROOF OF PROPOSITIONS 1 AND 2

Proof of Proposition 1. Fix $a \in \mathcal{A}$ and a true descendant $i \in \mathrm{desc}(a)$. Since Algorithm 1 sets $\widehat{\mathrm{desc}}(a) = \mathrm{Cand}(a)$, it suffices to show $i \in \mathrm{Cand}(a) = \mathcal{S}_a \cap \bigcap_{b\in\mathcal{U}(a)} \mathcal{S}_b$.

By Assumption 2, $\mathbb{P}(i \in \mathcal{S}_a) \ge 1-\epsilon_{\mathrm{fn}}$, hence $\mathbb{P}(i \notin \mathcal{S}_a) \le \epsilon_{\mathrm{fn}}$.

For each $b \in \mathcal{U}(a)$: if b is a true ancestor of a (i.e., $a \in \mathrm{desc}(b)$), then $\mathrm{desc}(a) \subseteq \mathrm{desc}(b)$, so $i \in \mathrm{desc}(b)$. By Assumption 2, $\mathbb{P}(i \in \mathcal{S}_b) \ge 1-\epsilon_{\mathrm{fn}}$, hence $\mathbb{P}(i \notin \mathcal{S}_b) \le \epsilon_{\mathrm{fn}}$.

The event $\{i \notin \mathrm{Cand}(a)\}$ requires either $i \notin \mathcal{S}_a$ or $\exists\, b \in \mathcal{U}(a)$ with $i \notin \mathcal{S}_b$. By a union bound:

$$\mathbb{P}(i \notin \mathrm{Cand}(a)) \le (|\mathcal{U}(a)|+1)\cdot\epsilon_{\mathrm{fn}}.$$

A further union bound over all $|\mathrm{desc}(a)|$ true descendants gives the result. Let $E_a = \{\exists\, i \in \mathrm{desc}(a) : i \notin \mathrm{Cand}(a)\}$.

$$\mathbb{P}(E_a) \le |\mathrm{desc}(a)|\cdot(|\mathcal{U}(a)|+1)\cdot\epsilon_{\mathrm{fn}}.$$

Hence $\mathrm{desc}(a) \subseteq \widehat{\mathrm{desc}}(a)$ with probability at least $1-|\mathrm{desc}(a)|\cdot(|\mathcal{U}(a)|+1)\cdot\epsilon_{\mathrm{fn}}$. □

Proof of Proposition 2. Fix $a \in \mathcal{A}$ and $i \notin \mathrm{desc}(a)$. First, because $i \notin \mathrm{desc}(a)$, Assumption 2 gives $\mathbb{P}(i \in \mathcal{S}_a) \le \epsilon_{\mathrm{fp}}$. Since Algorithm 1 can include i in $\widehat{\mathrm{desc}}(a)$ only if $i \in \mathcal{S}_a$, this already gives $\mathbb{P}(i \in \widehat{\mathrm{desc}}(a)) \le \epsilon_{\mathrm{fp}}$. By Assumption 3, there exists $b \in \mathcal{A}$ such that $a \in \mathrm{desc}(b)$ but $i \notin \mathrm{desc}(b)$ and

$$\mathbb{P}(a \in \mathcal{S}_b \text{ and } i \notin \mathcal{S}_b) \ge 1 - \epsilon_{\mathrm{cx}}.$$

On this event, b is identified as upstream of a (since $a \in \mathcal{S}_b$), but $i \notin \mathcal{S}_b$. Thus $i \notin \bigcap_{b' \in \mathcal{U}(a)} \mathcal{S}_{b'}$ and therefore $i \notin \mathrm{Cand}(a)$. Hence $\mathbb{P}(i \in \mathrm{Cand}(a)) \le \epsilon_{\mathrm{cx}}$. Since $\widehat{\mathrm{desc}}(a) = \mathrm{Cand}(a)$ in Algorithm 1, combining the direct false-positive bound with the upstream-pruning bound gives $\mathbb{P}(i \in \widehat{\mathrm{desc}}(a)) \le \epsilon_{\mathrm{sel}} = \min\{\epsilon_{\mathrm{fp}}, \epsilon_{\mathrm{cx}}\}$. Since $\widehat{Z}_{a,i} = \mathbf{1}\{i \in \widehat{\mathrm{desc}}(a)\}$,

$$\begin{aligned}\mathbb{P}(\widehat{Z}_{a,i} = 1 \mid Z_{a,i} = 0) &= \mathbb{P}(i \in \widehat{\mathrm{desc}}(a) \mid i \notin \mathrm{desc}(a)) \\ &\le \epsilon_{\mathrm{sel}}.\end{aligned}$$

□

D FPR/FNR CONVENTION FOR ALGORITHM 1

We restate the recovery consequences of Algorithm 1 using the descendant-positive convention

$$Z_{a,i} = \mathbf{1}\{i \in \mathrm{desc}(a)\}, \qquad \widehat{Z}_{a,i} = \mathbf{1}\{i \in \widehat{\mathrm{desc}}(a)\}.$$

Thus

$$\mathrm{FNR}_{a,i} = \mathbb{P}(\widehat{Z}_{a,i} = 0 \mid Z_{a,i} = 1)$$

is the probability that a calibration intervention that truly affects target i is selected as safe; this is the error that contaminates selective conformal calibration. Conversely,

$$\mathrm{FPR}_{a,i} = \mathbb{P}(\widehat{Z}_{a,i} = 1 \mid Z_{a,i} = 0)$$

is the probability of discarding a truly safe calibration intervention, which affects efficiency but not validity.

Proposition 3 (FNR bound for Algorithm 1)**.** *Under Assumption 2, fix $a \in \mathcal{A}$ and suppose that $\mathcal{U}(a) \neq \emptyset$ and every identified upstream intervention is a true ancestor of a, i.e.,*

$$b \in \mathcal{U}(a) \quad \Longrightarrow \quad a \in \mathrm{desc}(b).$$

Then for any fixed $i \in \mathrm{desc}(a)$,

$$\mathbb{P}(i \notin \widehat{\mathrm{desc}}(a)) \le \big(|\mathcal{U}(a)| + 1\big)\epsilon_{\mathrm{fn}},$$

and hence under the descendant-positive convention,

$$\mathrm{FNR}_{a,i} \le \big(|\mathcal{U}(a)| + 1\big)\epsilon_{\mathrm{fn}}.$$

A further union bound over $i \in \mathrm{desc}(a)$ gives

$$\mathbb{P}\big(\mathrm{desc}(a) \subseteq \widehat{\mathrm{desc}}(a)\big) \ge 1 - |\mathrm{desc}(a)|\big(|\mathcal{U}(a)| + 1\big)\epsilon_{\mathrm{fn}}.$$

Proof. Fix $i \in \mathrm{desc}(a)$. In order for i to belong to

$$\mathrm{Cand}(a) = \mathcal{S}_a \cap \bigcap_{b \in \mathcal{U}(a)} \mathcal{S}_b,$$

we need $i \in \mathcal{S}_a$ and $i \in \mathcal{S}_b$ for every $b \in \mathcal{U}(a)$. By Assumption 2,

$$\mathbb{P}(i \notin \mathcal{S}_a) \le \epsilon_{\mathrm{fn}}.$$

If $b \in \mathcal{U}(a)$ is a true ancestor of a, then $\mathrm{desc}(a) \subseteq \mathrm{desc}(b)$, hence $i \in \mathrm{desc}(b)$ and again

$$\mathbb{P}(i \notin \mathcal{S}_b) \le \epsilon_{\mathrm{fn}}.$$

A union bound over $\mathcal{S}_a$ and the $|\mathcal{U}(a)|$ upstream sets gives the per-pair bound

$$\mathbb{P}(i \notin \mathrm{Cand}(a)) \le \big(|\mathcal{U}(a)|+1\big)\epsilon_{\mathrm{fn}}.$$

Since $\widehat{\mathrm{desc}}(a) = \mathrm{Cand}(a)$ in Algorithm 1,

$$\mathbb{P}(i \notin \widehat{\mathrm{desc}}(a)) = \mathbb{P}(i \notin \mathrm{Cand}(a)) \le \big(|\mathcal{U}(a)|+1\big)\epsilon_{\mathrm{fn}},$$

and

$$\mathrm{FNR}_{a,i} = \mathbb{P}(\widehat{Z}_{a,i}=0 \mid Z_{a,i}=1) = \mathbb{P}(i \notin \mathrm{Cand}(a) \mid i \in \mathrm{desc}(a)) \le \big(|\mathcal{U}(a)|+1\big)\epsilon_{\mathrm{fn}}.$$

Taking a further union bound over all $i \in \mathrm{desc}(a)$ yields

$$\mathbb{P}\big(\exists\, i \in \mathrm{desc}(a) : i \notin \mathrm{Cand}(a)\big) \le \sum_{i \in \mathrm{desc}(a)} \mathbb{P}(i \notin \mathrm{Cand}(a)) \le |\mathrm{desc}(a)|\big(|\mathcal{U}(a)|+1\big)\epsilon_{\mathrm{fn}},$$

which is equivalent to the displayed lower bound for $\mathbb{P}(\mathrm{desc}(a) \subseteq \widehat{\mathrm{desc}}(a))$. □

Proposition 4 (FPR bound for Algorithm 1)**.** *Under Assumptions 2–3, for every $a \in \mathcal{A}$ and every non-descendant $i \notin \mathrm{desc}(a)$,*

$$\mathbb{P}(\widehat{Z}_{a,i}=1 \mid Z_{a,i}=0) \le \epsilon_{\mathrm{sel}}.$$

Equivalently, under the descendant-positive convention,

$$\mathrm{FPR}_{a,i} \le \epsilon_{\mathrm{sel}}, \qquad \epsilon_{\mathrm{sel}} = \min\{\epsilon_{\mathrm{fp}}, \epsilon_{\mathrm{cx}}\}.$$

Proof. Fix $i \notin \mathrm{desc}(a)$. Since $i \notin \mathrm{desc}(a)$, Assumption 2 gives $\mathbb{P}(i \in \mathcal{S}_a) \le \epsilon_{\mathrm{fp}}$, and Algorithm 1 can include i only if $i \in \mathcal{S}_a$. Hence $\mathbb{P}(i \in \widehat{\mathrm{desc}}(a)) \le \epsilon_{\mathrm{fp}}$. By Assumption 3, there exists $b \in \mathcal{A}$ such that $a \in \mathrm{desc}(b)$, $i \notin \mathrm{desc}(b)$, and

$$\mathbb{P}(a \in \mathcal{S}_b \text{ and } i \notin \mathcal{S}_b) \ge 1 - \epsilon_{\mathrm{cx}}.$$

On this event, b is included in the upstream set $\mathcal{U}(a)$ while i is absent from $\mathcal{S}_b$. Therefore i is removed by the intersection step and $i \notin \mathrm{Cand}(a)$. Since $\widehat{\mathrm{desc}}(a) = \mathrm{Cand}(a)$ in Algorithm 1, this gives the second bound $\mathbb{P}(i \in \widehat{\mathrm{desc}}(a)) \le \epsilon_{\mathrm{cx}}$. Combining the two bounds yields $\mathbb{P}(i \in \widehat{\mathrm{desc}}(a)) \le \epsilon_{\mathrm{sel}}$.

$$\mathbb{P}(\widehat{Z}_{a,i}=1 \mid Z_{a,i}=0) = \mathbb{P}(i \in \widehat{\mathrm{desc}}(a) \mid i \notin \mathrm{desc}(a)) \le \epsilon_{\mathrm{sel}}.$$

□

Corollary 3 (Expected contamination under corrected labels)**.** *Fix target i and calibration split $\mathcal{A}_{\mathrm{cal}}$. Let π_0 be the fraction of truly safe interventions in $\mathcal{A}_{\mathrm{cal}}$, meaning interventions that do not affect target i:*

$$\pi_0 = \frac{|\{a \in \mathcal{A}_{\mathrm{cal}} : Z_{a,i}=0\}|}{|\mathcal{A}_{\mathrm{cal}}|}.$$

Define the worst-case bounds

$$\mathrm{FNR}_{\max} = \max_{a \in \mathcal{A}_{\mathrm{cal}}} \big(|\mathcal{U}(a)|+1\big)\epsilon_{\mathrm{fn}} = (\bar{u}+1)\epsilon_{\mathrm{fn}}, \qquad \mathrm{FPR}_{\max} = \epsilon_{\mathrm{sel}}.$$

Here $\bar{u} = \max_{a \in \mathcal{A}_{\mathrm{cal}}} |\mathcal{U}(a)|$. Under Assumptions 2–3 and the upstream-ancestor condition in Proposition 3,

$$\mathbb{E}[\delta] \le \frac{(1-\pi_0)\mathrm{FNR}_{\max}}{\pi_0(1-\mathrm{FPR}_{\max}) + (1-\pi_0)\mathrm{FNR}_{\max}}.$$

Equivalently,

$$\mathbb{E}[\delta] \le \frac{(1-\pi_0)(\bar{u}+1)\epsilon_{\mathrm{fn}}}{\pi_0(1-\epsilon_{\mathrm{sel}}) + (1-\pi_0)(\bar{u}+1)\epsilon_{\mathrm{fn}}}.$$

Proof. Under the descendant-positive convention, contamination occurs exactly when $Z_{a,i} = 1$ but $\widehat{Z}_{a,i} = 0$. Let A denote the expected selected mass among interventions that truly affect target i and S the expected selected mass among safe interventions. By the per-pair FNR bound in Proposition 3 and the FPR bound in Proposition 4,

$$A \le (1 - \pi_0)\mathrm{FNR}_{\max}, \qquad S \ge \pi_0(1 - \mathrm{FPR}_{\max}).$$

Using the contamination expression from Section 4, $\mathbb{E}[\delta] = A/(S + A)$ at the level of expected selected masses. This expression is increasing in A and decreasing in S, so

$$\mathbb{E}[\delta] \le \frac{(1 - \pi_0)\mathrm{FNR}_{\max}}{\pi_0(1 - \mathrm{FPR}_{\max}) + (1 - \pi_0)\mathrm{FNR}_{\max}}.$$

Equivalently, at the level of expected or population rates, the selected safe mass is at least

$$\pi_0(1 - \mathrm{FPR}_{\max})$$

and the contaminating selected mass is at most $(1 - \pi_0)\mathrm{FNR}_{\max}$. The bounds on $\mathrm{FNR}_{\max}$ and $\mathrm{FPR}_{\max}$ are exactly Propositions 3 and 4. The factor $\bar{u} + 1$ counts the own affected set $\mathcal{S}_a$ together with the $|\mathcal{U}(a)|$ upstream affected sets in the union bound. □

Remark 5 (Validity versus efficiency). With these labels, Proposition 3 is the validity-relevant statement: false negatives under the descendant-positive convention are interventions that truly affect target i but are incorrectly admitted into $\widehat{\mathcal{A}}(i, a^\star)$, and hence they drive δ and the coverage loss in Theorem 1. Proposition 4 is efficiency-relevant: false positives discard safe calibration interventions and can widen intervals, but they do not introduce non-exchangeable scores into the selected calibration set.

E A HIGH-PROBABILITY BOUND ON THE CONTAMINATION FRACTION

The correction in Corollary 1 requires an upper bound on the realized contamination fraction

$$\delta = \frac{|\widehat{\mathcal{A}}(i, a^\star) \setminus \mathcal{A}^\star(i)|}{|\widehat{\mathcal{A}}(i, a^\star)|}.$$

Corollary 2 controls this quantity in expectation. We next give a high-probability version under an explicit weak-dependence condition on the classification errors across calibration interventions.

Assumption 4 (Fixed selected size and weak dependence)**.** Fix a target i and test intervention $a^\star$ with $Z_{a^\star,i} = 0$. Assume that $|\widehat{\mathcal{A}}(i, a^\star)| = n$ is fixed by design, or condition on an event on which $|\widehat{\mathcal{A}}(i, a^\star)| = n$ and the error bounds below hold conditionally. For each $a \in \mathcal{A}_{\mathrm{cal}}$, define

$$W_a = \mathbf{1}\{a \in \widehat{\mathcal{A}}(i, a^\star),\ Z_{a,i} = 1\}.$$

There is a dependency graph H_i on $\mathcal{A}_{\mathrm{cal}}$ with maximum degree Δ_i such that any two subcollections of $\{W_a : a \in \mathcal{A}_{\mathrm{cal}}\}$ indexed by vertex sets with no edge between them are independent.

Assumption 4 is automatic with $\Delta_i = 0$ if the descendant-classification errors are independent across interventions. The dependency-graph form allows the upstream-intersection step in Algorithm 1 to induce local dependence through shared differentially affected sets. When the selected set size is random and no conditioning is imposed, the same argument applies with n replaced by any deterministic lower bound on $|\widehat{\mathcal{A}}(i, a^\star)|$.

Proposition 5 (High-probability contamination bound)**.** *Under Assumption 4, let*

$$C_i = \sum_{a \in \mathcal{A}_{\mathrm{cal}}} W_a, \qquad \mu_i = \mathbb{E}[C_i], \qquad \sigma_i^2 = \mathrm{Var}(C_i).$$

Then $\delta = C_i/n$ and, for every $t > 0$,

$$\mathbb{P}\left(\delta > \frac{\mu_i}{n} + t\right) \le \frac{\sigma_i^2}{n^2 t^2}.$$

Moreover, if $q_a = \mathbb{P}(W_a = 1)$, then

$$\sigma_i^2 \le \sum_{a \in \mathcal{A}_{\mathrm{cal}}} q_a(1-q_a) + 2 \sum_{\{a,b\} \in E(H_i)} \sqrt{q_a q_b} \le (\Delta_i + 1)\mu_i.$$

Consequently, for any $\eta \in (0,1)$, with probability at least $1-\eta$ over the realization of $\widehat{Z}$,

$$\delta \le \widehat{\delta}_\eta \equiv \frac{\mu_i}{n} + \frac{\sigma_i}{n\sqrt{\eta}} \le \frac{\mu_i}{n} + \frac{\sqrt{(\Delta_i+1)\mu_i}}{n\sqrt{\eta}}.$$

Under the descendant-positive convention, if

$$\mathrm{FNR}_a = \mathbb{P}(\widehat{Z}_{a,i} = 0 \mid Z_{a,i} = 1),$$

then

$$\mu_i = \sum_{a : Z_{a,i}=1} \mathbb{P}(\widehat{Z}_{a,i} = 0) \le \sum_{a : Z_{a,i}=1} \mathrm{FNR}_a.$$

Proof. By definition, $W_a = 1$ exactly when calibration intervention a is selected but truly affects target i. Hence

$$C_i = \sum_{a \in \mathcal{A}_{\mathrm{cal}}} W_a = |\widehat{\mathcal{A}}(i, a^\star) \setminus \mathcal{A}^\star(i)|.$$

Conditioning on $|\widehat{\mathcal{A}}(i,a^\star)| = n$ gives $\delta = C_i/n$.

Chebyshev's inequality gives, for every $t > 0$,

$$\mathbb{P}\left(\delta > \frac{\mu_i}{n} + t\right) = \mathbb{P}(C_i - \mu_i > nt) \le \mathbb{P}(|C_i - \mu_i| > nt) \le \frac{\sigma_i^2}{n^2 t^2}.$$

It remains to bound σ_i^2. Since W_a is Bernoulli,

$$\mathrm{Var}(W_a) = q_a(1-q_a).$$

By the dependency-graph assumption, $\mathrm{Cov}(W_a, W_b) = 0$ whenever $\{a,b\} \notin E(H_i)$. For edges of H_i, Cauchy–Schwarz gives

$$|\mathrm{Cov}(W_a, W_b)| \le \sqrt{\mathrm{Var}(W_a)\mathrm{Var}(W_b)} \le \sqrt{q_a q_b}.$$

Therefore

$$\mathrm{Var}(C_i) \le \sum_a q_a(1-q_a) + 2 \sum_{\{a,b\} \in E(H_i)} \sqrt{q_a q_b}.$$

Using $2\sqrt{q_a q_b} \le q_a + q_b$ and the fact that each vertex has degree at most Δ_i,

$$2 \sum_{\{a,b\} \in E(H_i)} \sqrt{q_a q_b} \le \sum_{\{a,b\} \in E(H_i)} (q_a + q_b) \le \Delta_i \sum_a q_a = \Delta_i \mu_i.$$

Since $\sum_a q_a(1-q_a) \le \sum_a q_a = \mu_i$, we obtain $\sigma_i^2 \le (\Delta_i + 1)\mu_i$. The displayed $1-\eta$ bound follows by taking $t = \sigma_i/(n\sqrt{\eta})$. Finally, if $Z_{a,i} = 0$ then $W_a = 0$ deterministically; if $Z_{a,i} = 1$ then $W_a = 1$ implies $\widehat{Z}_{a,i} = 0$, so $\mathbb{P}(W_a = 1) \le \mathbb{P}(\widehat{Z}_{a,i} = 0 \mid Z_{a,i} = 1) = \mathrm{FNR}_a$. □

Corollary 4 (High-probability corrected selective conformal)**.** *Fix $\eta \in (0,1)$ and suppose the assumptions of Proposition 5 hold. Let $\widehat{\delta}_\eta$ be any deterministic quantity satisfying*

$$\mathbb{P}(\delta \le \widehat{\delta}_\eta) \ge 1 - \eta.$$

For example, one may use the bound in Proposition 5, with μ_i further upper bounded by the FNR bounds from Proposition 3. Then, with probability at least $1-\eta$ over the realized descendant estimator $\widehat{Z}$, the split conformal interval computed at nominal miscoverage

$$\alpha' = \alpha - g(\widehat{\delta}_\eta, n)$$

has conditional coverage at least $1-\alpha$ for $R_i^{(a^\star)}$. If $\alpha' \le 0$, the corrected interval is taken to be $(-\infty, \infty)$ and the claim is trivial.

Proof. On the event $\{\delta \le \widehat{\delta}_\eta\}$, Corollary 1 applies with upper bound $\widehat{\delta}_\eta$ and gives conditional coverage at least $1-\alpha$. Proposition 5 gives this event probability at least $1-\eta$. □

Remark 6 (When is the bound informative?). The Chebyshev correction is useful when the additive term $\sigma_i/(n\sqrt{\eta})$ is small compared with the target slack in δ. For a relative bound of the form $\delta \le (1+c)\mathbb{E}[\delta]$, Chebyshev gives failure probability at most

$$\frac{\mathrm{Var}(\delta)}{c^2\mathbb{E}[\delta]^2} = \frac{\sigma_i^2}{c^2\mu_i^2} \le \frac{\Delta_i + 1}{c^2\mu_i}.$$

Thus relative concentration requires $\mu_i \gg \Delta_i$, i.e., the expected number of contaminating selected interventions must dominate the local dependence degree. For coverage correction, however, the relevant requirement is often absolute rather than relative:

$$\frac{\sqrt{(\Delta_i+1)\mu_i}}{n\sqrt{\eta}} \ll 1.$$

This can hold even when $\mathbb{E}[\delta] = \mu_i/n$ is small, provided the selected calibration set is large and the dependency graph is sparse. If the upstream-intersection errors are highly coupled across many interventions, so that Δ_i is of the same order as $|\mathcal{A}_{\mathrm{cal}}|$, this Chebyshev bound becomes conservative; sharper concentration would require a more detailed model of the dependence induced by Algorithm 1.

F ALGORITHM 2: LOCAL ICP FOR DISTANCE ESTIMATION

Algorithm 2 Local ICP for Distance Estimation

Require: Target gene i, interventions $\mathcal{A}$, data $\{Y^{(a)}\}_{a\in\mathcal{A}}$, maximum depth D
Ensure: Estimated distance $\widehat{d}(a,i)$ for each $a \in \mathcal{A}$

$\mathcal{P}_0 \leftarrow$ candidate parent pool via correlation screening or prior network
$\widehat{\mathrm{Pa}}(i) \leftarrow \mathrm{ICP}(i, \mathcal{P}_0, \mathcal{A})$ {Invariant parent set}
$\mathcal{B}_0 \leftarrow \{i\}$
for $t = 1, \dots, D$ **do**
  $\mathcal{B}_t \leftarrow \mathcal{B}_{t-1}$
  $\{\mathcal{B}_{-1} = \emptyset\}$
  for each $g \in \mathcal{B}_{t-1} \setminus \mathcal{B}_{t-2}$ **do**
    $\mathcal{P}_0^{(g)} \leftarrow$ candidate parent pool for g
    $\widehat{\mathrm{Pa}}(g) \leftarrow \mathrm{ICP}(g, \mathcal{P}_0^{(g)}, \mathcal{A})$
    $\mathcal{B}_t \leftarrow \mathcal{B}_t \cup \widehat{\mathrm{Pa}}(g)$
  end for
end for
for each $a \in \mathcal{A}$ **do**
  $\widehat{d}(a,i) \leftarrow \min\{t : a \in \mathcal{B}_t\}$ {∞ if $a \notin \mathcal{B}_D$}
end for
return $\{\widehat{d}(a,i)\}_{a\in\mathcal{A}}$

The ICP subroutine works as follows: for a target gene g and candidate parent pool $\mathcal{P}_0^{(g)}$, test subsets $S \subseteq \mathcal{P}_0^{(g)}$ for invariance of the regression residual $Y_g - \hat{\beta}_S^\top X_S$ across environments (interventions), and return a minimal invariant set as $\widehat{\mathrm{Pa}}(g)$. The frontier expansion traces back from i through estimated parents, yielding a coarse path-length estimate without learning the full graph. For runtime accounting, $\mathcal{P}_0^{(g)}$ denotes the screened candidate parent pool used when ICP is called for gene g; the initial pool $\mathcal{P}_0$ in Algorithm 2 is $\mathcal{P}_0^{(i)}$. Let $K_0 = \max_g |\mathcal{P}_0^{(g)}|$, let k_{ICP} be the maximum parent-set size searched by the ICP subroutine, let D be the maximum backward search depth, and let $M_D(i) = |\cup_{t=0}^{D} \mathcal{B}_t|$ be the number of genes visited during the local expansion from target i. Each ICP call examines $O(K_0^{k_{\mathrm{ICP}}})$ candidate subsets, so the subset-search cost is $O(M_D(i) K_0^{k_{\mathrm{ICP}}})$, up to the cost of the individual invariance tests. In the single-gene perturbation setting with the expansion restricted to intervention-indexed genes, the shorthand bound is $O(|\mathcal{A}| D K_0^{k_{\mathrm{ICP}}})$. Thus correlation screening makes the runtime depend on the screened pool size K_0 and the local visited set, rather than directly on the ambient number of genes p.

Using distance for weighted calibration. Given distance estimates $\widehat{d}(a,i)$, one can define weighted conformal calibration:

$$w(a) = K\left(\frac{\widehat{d}(a,i)}{h}\right),$$

where K is a kernel function and h is a bandwidth parameter. This provides a smooth interpolation between selective calibration (hard thresholding on $\widehat{Z}$) and pooled calibration, potentially improving efficiency when the binary classification is uncertain.

G CONSISTENCY OF ALGORITHM 2

We give a recovery guarantee for the Local ICP distance estimator in Algorithm 2. Recall that the algorithm starts from $\mathcal{B}_0 = \{i\}$ and recursively expands backward through estimated invariant parent sets. It returns

$$\widehat{d}(a, i) = \min\{t : \ a \in \mathcal{B}_t\},$$

with $\widehat{d}(a, i) = \infty$ if $a \notin \mathcal{B}_D$.

For a node g, let $\mathrm{Pa}(g)$ denote its true parent set in the underlying DAG G, and let $d(a, i)$ be the length of the shortest directed path from a to i, with $d(a, i) = \infty$ if no such path exists. Define

$$\mathrm{Anc}^{\leq D}(i) = \{g : \ d(g, i) \leq D\},$$

where $i \in \mathrm{Anc}^{\leq D}(i)$ with $d(i, i) = 0$.

Proposition 6 (Local ICP distance recovery)**.** *Assume:*

1. *the data are generated from a causally sufficient DAG G;*
2. *interventional faithfulness holds, so invariant causal prediction identifies the true invariant parent set in the population;*
3. *for every gene g that can be visited by Algorithm 2, the candidate parent pool covers the true parents,*

$$\mathcal{P}_0^{(g)} \supseteq \mathrm{Pa}(g);$$

4. *the ICP subroutine satisfies the finite-sample recovery guarantee*

$$\mathbb{P}\big(\widehat{\mathrm{Pa}}(g) = \mathrm{Pa}(g)\big) \geq 1 - \beta_g$$

for every $g \in \mathrm{Anc}^{\leq D}(i)$.

Then, with probability at least

$$1 - \sum_{g \in \mathrm{Anc}^{\leq D}(i)} \beta_g,$$

Algorithm 2 recovers the exact truncated ancestor balls:

$$\mathcal{B}_t = \{g : \ d(g, i) \leq t\} \qquad \textit{for all } t = 0, \ldots, D.$$

Consequently, for every $a \in \mathcal{A}$ with $d(a, i) \leq D$,

$$\mathbb{P}\big(\widehat{d}(a, i) = d(a, i)\big) \geq 1 - \sum_{g \in \mathrm{Anc}^{\leq D}(i)} \beta_g,$$

and for every $a \in \mathcal{A}$ with $d(a, i) > D$,

$$\mathbb{P}\big(\widehat{d}(a, i) = \infty\big) \geq 1 - \sum_{g \in \mathrm{Anc}^{\leq D}(i)} \beta_g.$$

Proof. Let

$$E_i = \bigcap_{g \in \mathrm{Anc}^{\leq D}(i)} \{\widehat{\mathrm{Pa}}(g) = \mathrm{Pa}(g)\}.$$

By a union bound,

$$\mathbb{P}(E_i) \geq 1 - \sum_{g \in \mathrm{Anc}^{\leq D}(i)} \beta_g.$$

It is enough to prove the deterministic claim on E_i.

We proceed by induction on t. For $t = 0$, Algorithm 2 initializes $\mathcal{B}_0 = \{i\}$, which equals $\{g : d(g, i) \leq 0\}$. Assume for some $t - 1 < D$ that

$$\mathcal{B}_{t-1} = \{g : \ d(g, i) \leq t - 1\}.$$

The new frontier is $\mathcal{B}_{t-1} \setminus \mathcal{B}_{t-2}$, which by the induction hypothesis is exactly the set of nodes at distance $t - 1$ from i. For each such node g, on the event E_i the ICP call returns

$$\widehat{\mathrm{Pa}}(g) = \mathrm{Pa}(g).$$

Therefore the update

$$\mathcal{B}_t = \mathcal{B}_{t-1} \cup \bigcup_{g \in \mathcal{B}_{t-1} \setminus \mathcal{B}_{t-2}} \widehat{\mathrm{Pa}}(g)$$

equals

$$\{g : \ d(g, i) \leq t - 1\} \cup \bigcup_{h: \ d(h,i)=t-1} \mathrm{Pa}(h).$$

Every parent of a node at distance $t - 1$ has a directed path of length t to i, so the right-hand side is contained in $\{g : d(g, i) \leq t\}$. Conversely, if u satisfies $d(u, i) = t$, then there exists a shortest directed path

$$u \to h \to \cdots \to i$$

where $d(h, i) = t - 1$ and $u \in \mathrm{Pa}(h)$. Since h is in the frontier at depth $t - 1$, the algorithm adds u at step t. Thus $\mathcal{B}_t = \{g : d(g, i) \leq t\}$.

The induction proves the exact recovery of all truncated ancestor balls through depth D on E_i. If $d(a, i) \leq D$, then a first appears in $\mathcal{B}_{d(a,i)}$ and does not appear in any earlier $\mathcal{B}_t$, so $\widehat{d}(a, i) = d(a, i)$. If $d(a, i) > D$, then $a \notin \mathcal{B}_D$, so the algorithm returns $\widehat{d}(a, i) = \infty$. Combining these deterministic implications with the lower bound on $\mathbb{P}(E_i)$ proves the result. □

Corollary 5 (Finite-sample consistency)**.** *Suppose the assumptions of Proposition 6 hold and that, for each fixed $g \in \mathrm{Anc}^{\leq D}(i)$, the ICP error probability satisfies*

$$\beta_g = \beta_g(n_{\min}) \to 0 \qquad \textit{as} \qquad n_{\min} = \min_{a \in \mathcal{A}} n_a \to \infty,$$

where n_a is the sample size under intervention a. If $|\mathrm{Anc}^{\leq D}(i)| < \infty$, then for every $a \in \mathcal{A}$,

$$\widehat{d}(a, i) \xrightarrow{p} \begin{cases} d(a, i), & d(a, i) \leq D, \\ \infty, & d(a, i) > D. \end{cases}$$

Proof. By Proposition 6, the probability of any truncated-distance recovery error is at most

$$\sum_{g \in \mathrm{Anc}^{\leq D}(i)} \beta_g(n_{\min}).$$

The ancestor set is finite and each summand converges to zero, so the sum converges to zero. □

Remark 7 (Implication for weighted conformal calibration). Appendix F proposes kernel weights

$$w(a) = K\left(\frac{\widehat{d}(a, i)}{h}\right)$$

for distance-weighted calibration. Proposition 6 shows that, under the stated ICP recovery conditions, these estimated weights equal the oracle weights $K(d(a, i)/h)$ with probability tending to one as the ICP subroutine becomes consistent. Therefore any exchangeability or weighted-exchangeability guarantee proved for the oracle distance weights is inherited asymptotically by the Local ICP weights. This does not by itself give a new finite-sample weighted-conformal theorem; it reduces the additional error from estimating $d(a, i)$ to the ICP recovery probability in Proposition 6.

H ADDITIONAL EXPERIMENTAL RESULTS

H.1 PROCEDURAL DEFINITION OF TABLE METHODS

For a fixed test intervention–target pair $(a^\star, i)$, let $r_a = R_i^{(a)}$ be the calibration score for intervention $a \in \mathcal{A}_{\text{cal}}$ and let $r^\star = R_i^{(a^\star)}$ be the test score. For any selected calibration set $S \subseteq \mathcal{A}_{\text{cal}}$, split conformal uses the order statistic

$$\widehat{q}_\alpha(S) = \text{the } \lceil (|S|+1)(1-\alpha) \rceil \text{-th smallest value in } \{r_a : a \in S\} \cup \{\infty\}.$$

Coverage for the score-based experiments is the indicator $\mathbf{1}\{r^\star \le \widehat{q}_\alpha(S)\}$ and width is $2\widehat{q}_\alpha(S)$. Here S is only a local shorthand for a selected intervention set: $S_{\text{oracle}}(i) = \mathcal{A}^\star(i)$ and $S_{\text{est}}(i) = \widehat{\mathcal{A}}(i, a^\star)$ in the notation of the main text.

Oracle. The oracle selector uses the truly safe calibration set $S_{\text{oracle}}(i) = \{a \in \mathcal{A}_{\text{cal}} : Z_{a,i} = 0\}$. In synthetic experiments Z is known from the simulated DAG. In the real-data experiment this row is only a proxy oracle, with Z defined by the top-10% absolute log-fold-change rule.

Estimated. The estimated selector uses Algorithm 1 to form $\widehat{Z}$ and selects $S_{\text{est}}(i) = \{a \in \mathcal{A}_{\text{cal}} : \widehat{Z}_{a,i} = 0\}$. If the selected set is too small to yield a finite split-conformal quantile, the implementation falls back to pooled calibration.

Pooled. The pooled baseline ignores causal strata and uses all calibration interventions, $S_{\text{pooled}} = \mathcal{A}_{\text{cal}}$. It is included as a coarse baseline rather than as the target exchangeable set for fixed safe test pairs.

Corrected. The corrected method uses S_{est} but replaces α by $\alpha' = \alpha - g(\widehat{\delta}, |S_{\text{est}}|)$, where $\widehat{\delta}$ is treated as an externally supplied upper bound; in the experiments we plug in realized synthetic contamination or proxy real-data contamination. The formal correction returns an infinite interval when $\alpha' \le 0$; the experiments report finite-width behavior under the clipped implementation described in the main text.

Control-coexpression weighted CP. For perturbation a, let $g(a)$ be its target gene and let $x_a = (\log(1 + X_{c,g(a)}))_{c \in \text{control}}$ be the vector of control-cell expression values for that gene. We compute a correlation distance between x_a and $x_{a^\star}$ and assign Gaussian kernel weights to calibration interventions, with bandwidth set by the tenth nearest calibration perturbation. This baseline is leakage-free because it uses only control-cell expression of the perturbed target genes, not the held-out perturbation's response vector. The reported n_{cal} is the Kish effective sample size of the weights.

Full-graph proxy. This baseline constructs a proxy gene graph from correlations among perturbation-level LFC profiles, thresholds the top 1% of absolute correlations to form a skeleton, orients edges by perturbation-level variance, and uses graph descendants as affected labels before applying the same split-conformal procedure as Estimated.

Observational-only selection. This heuristic also uses the correlation matrix but does not run Algorithm 1. For each perturbed target, it labels the most correlated genes as affected, with the number of selected genes matched to the average descendant-set size produced by Algorithm 1, and then applies split conformal to the remaining calibration interventions.

Table 3: Controlled δ ablation. Coverage and width by injected contamination level. Corrected maintains ≥ 0.95 coverage at all nonzero contamination levels ($\delta \ge 0.05$) when supplied with the injected δ under the clipped finite-width implementation. Oracle and Pooled widths (constant at 3.38 and 3.29 respectively) are omitted, so the displayed width rows are interpreted as the uncorrected–corrected tradeoff rather than a complete width ranking.

	δ_{inject}					
Method	0.00	0.05	0.10	0.15	0.20	0.30
Coverage						
Oracle	**.905**	**.905**	**.905**	**.905**	**.905**	**.905**
Estimated	**.905**	**.901**	.895	.889	.882	.867
Pooled	.896	.896	.896	.896	.896	.896
Corrected	**.905**	**.955**	**.990**	**.990**	**.989**	**.988**
Width						
Estimated	3.38	3.33	3.28	3.22	3.16	3.03
Corrected	3.38	4.09	5.52	5.48	5.44	5.35

Feasibility versus calibration-set size. The corrected procedure's 59.8% feasibility on the real-data split (Table 2) reflects the small calibration set rather than a methodological limit. Re-running the Replogle K562 evaluation while scaling the

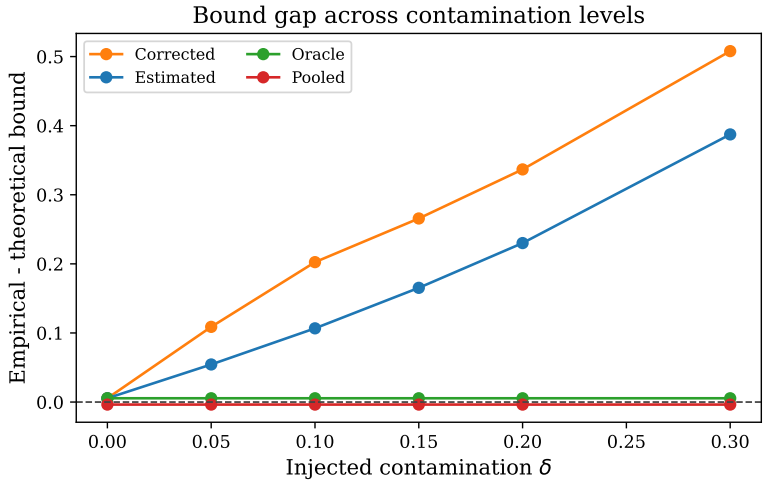


Figure 3: Gap between empirical coverage and the theoretical lower bound from Theorem 1. All values are non-negative for the selective methods (Oracle, Estimated, Corrected), confirming the bound is valid. Pooled shows a small negative gap (≈ -0.004) because it uses all calibration points without selection, so the selective coverage theorem does not apply. The gap grows with δ because worst-case adversarial contamination does not occur in practice.

calibration-set size shows feasibility climbing to 100% by $n_{\text{cal}} = 160$, with the proxy contamination fraction δ stable near 0.083 (Table 4, Figure 4). All rows in this scaling study are produced by bootstrap-resampling the selected real calibration residuals to the target n_{cal}, rather than by adding real perturbations, so their coverage and width should be read as indicative; feasibility, a function of $g(\delta, n)$ using this proxy value versus α, is robust to this expansion.

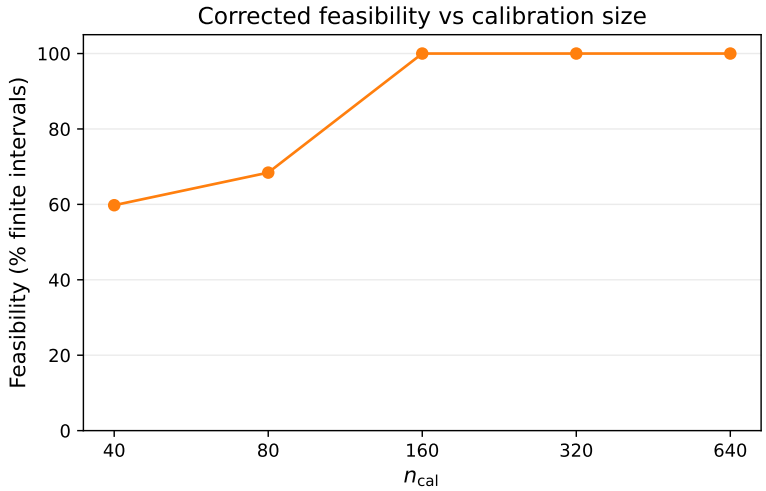


Figure 4: Feasibility (fraction of finite intervals) of the Corrected procedure versus calibration-set size n_{cal} on the Replogle K562 split. Feasibility rises from 0.598 at $n_{\text{cal}} = 40$ to 1.0 by $n_{\text{cal}} = 160$.

I ADDITIONAL ROBUSTNESS EXPERIMENTS

Extended contamination range. We extend the controlled-δ ablation of Section 8.2 to injected contamination up to $\delta = 0.70$ (same setting: $p = 200$, $|\mathcal{A}| = 150$, 100 seeds). Table 5 reports the Corrected procedure. Coverage stays at or above the nominal level across the entire range, and the intervals remain finite (feasibility ≈ 1) even at $\delta = 0.70$: at the synthetic calibration size ($n_{\text{cal}} \approx 120$) the corrected level is clipped at $\alpha' = 0.01$ for finite-width reporting, still admitting a finite conformal quantile. Without this clipping, rows where $\alpha - g(\delta_{\text{inject}}, n) \leq 0$ would correspond to the formal infinite-interval correction. The infinite-interval behavior seen on the real data (Table 2) is therefore a consequence of the small calibration

Table 4: Corrected procedure versus calibration-set size n_{cal} on the Replogle K562 split. Rows use bootstrap-resampled calibration residuals at the displayed target sizes. Feasibility reaches 100% by $n_{\mathrm{cal}} = 160$ while proxy δ stays near 0.083. Width highlighting ranks only rows that also exceed nominal coverage.

n_{cal}	Coverage	Width	Feasibility	proxy δ
40	0.890	0.316	0.598	0.083
80	0.898	0.338	0.684	0.083
160	**0.928**	0.520	1.000	0.083
320	**0.925**	**0.517**	1.000	0.083
640	**0.926**	<u>0.519</u>	1.000	0.083

set: at $n_{\mathrm{cal}} \approx 40$ the nonzero contamination lowers the corrected level enough that too few scores remain for a finite quantile, and increasing n_{cal} at the same δ restores feasibility (Figure 4).

Table 5: Corrected procedure under extended injected contamination (δ up to 0.70; $p = 200$, $|\mathcal{A}| = 150$, 100 seeds). Finite-width rows use the clipped level $\alpha' \geq 0.01$. Rows are contamination levels rather than competing methods, so widths are not ranked.

δ_{inject}	Coverage	Width	Feasibility
0.00	**0.905**	3.38	100.0%
0.05	**0.955**	4.09	100.0%
0.10	**0.990**	5.52	99.9%
0.15	**0.989**	5.48	99.9%
0.20	**0.989**	5.44	99.9%
0.30	**0.988**	5.35	99.9%
0.40	**0.986**	5.25	99.9%
0.50	**0.984**	5.12	99.9%
0.60	**0.979**	4.96	99.9%
0.70	**0.971**	4.75	99.9%

Nonlinear mechanisms and residual scores. We replace the linear SEM $V = B^\top V + \varepsilon$ with additive nonlinear mechanisms $V_v = \sum_{u \in \mathrm{pa}(v)} B_{uv} \tanh(V_u) + \varepsilon_v$, and evaluate under two nonconformity scores: the synthetic score used in the main experiments and a genuine ridge-regression residual score $R_i^{(a)} = |V_i - \hat{f}_i(V_{-i})|$ with $\hat{f}_i$ fit on observational data ($p = 200$, $|\mathcal{A}| = 150$, 20 seeds). Table 6 shows that Corrected remains the most conservative method while Oracle, Estimated, and Pooled stay close to nominal coverage, confirming that the method's behavior does not depend on the linear-Gaussian form of the synthetic generator.

Table 6: Nonlinear SEM with additive $\tanh$ mechanisms, under the synthetic score and a real ridge-residual score ($p = 200$, $|\mathcal{A}| = 150$, 20 seeds). Coverage / width per method. Ties at displayed precision are marked together.

	Synthetic score		Ridge-residual score	
Method	Coverage	Width	Coverage	Width
Oracle	**0.903**	<u>3.34</u>	**0.907**	**4.35**
Estimated	**0.901**	**3.32**	**0.906**	**4.35**
Pooled	**0.901**	**3.32**	**0.906**	**4.35**
Corrected	**0.920**	3.59	**0.924**	4.39

J ADDITIONAL DISCUSSION

J.1 FUTURE WORK

Calibration for descendant targets. The main theory focuses on test pairs with $Z_{a^\star,i} = 0$, where interventions leaving target i unchanged define a natural exchangeable stratum. Developing valid and efficient strata for descendant targets, where $Z_{a^\star,i} = 1$, remains open.

Weighted selective calibration. Algorithm 2 gives a distance estimator that can define kernel weights over interventions, but a finite-sample weighted-conformal guarantee with estimated causal distances remains to be proved.

Deployable contamination bounds. Our correction uses an upper bound on the selected-set contamination fraction δ; in the experiments this bound is supplied by true synthetic labels or real-data proxy labels. A fully deployable version would need data-driven high-confidence bounds on δ, for example by combining descendant-learner uncertainty, held-out validation perturbations, or external biological annotations.

Context-aware Mondrian strata. Real perturbation screens may violate exchangeability because of cell type, batch, tissue, dose, or donor effects. A natural extension is batch- or cell-type-stratified Mondrian calibration, potentially using multi-context resources such as scPerturb [Peidli et al., 2024].

Combinatorial perturbations and active design. Our notation treats a as a single intervention condition. Extending descendant learning and contamination control to combinatorial perturbations, and combining the method with active experimental design for causal discovery [Squires et al., 2020, Eberhardt et al., 2005], are important next steps.

Larger real-data validation. The real-data experiment uses proxy descendant labels and a small calibration set. Larger perturbation atlases with biologically validated causal annotations would allow a sharper test of both coverage and interval efficiency.