CERTIFYING QUERY-LOCAL CALIBRATION CONTAMINATION FOR SAFE-PAIR PREDICTION UNDER INTERVENTIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

This paper aims to construct reliable prediction intervals for intervention–outcome pairs asserted to be causally unaffected. Although working graphs identify candidate negative controls, graph errors can contaminate query-local calibration pools, making global error corrections insufficient. To address this problem, we propose Certified Graph-Local Conformal Prediction (CGLCP), which couples limited random verification with query-specific conformal calibration. For each fixed safe query, we invert the exact hypergeometric law to upper-bound the selected pool's unsafe count and translate this certificate into a conformal rank. Under safe-score exchangeability and exact independent verification, the procedure guarantees finite-sample marginal coverage and uses the smallest certificate-valid rank against arbitrary contaminating scores. We further quantify the efficiency loss relative to an oracle that knows the local contamination count. Simulations and a K562 Perturb-seq study illustrate the resulting trade-offs among coverage, interval width, finite-output availability, and verification cost, supporting the use of calibrated intervals as negative-control diagnostics under interventions.

1 INTRODUCTION

Large-scale intervention experiments increasingly ask not only which outcomes respond to a perturbation, but also whether outcomes that *should not respond* remain within a normal range. If a scientist knocks down gene a and prior pathway knowledge places gene i outside its downstream pathway, then (a, i) is a natural negative-control pair. A reliable interval for the expression of i describes variation compatible with that claim. An observation outside the interval can trigger investigation of guide-specific off-target activity, batch or cell-state shift, or an omitted path in the regulatory network. The same logic applies to a targeted drug and a biomarker that its known mechanism should leave unaffected. Perturbational atlases are valuable in this role because unexpected responses can reveal incomplete mechanisms and unanticipated treatment effects (Dixit et al., 2016; Replogle et al., 2022; Subramanian et al., 2017; Srivatsan et al., 2020).

The problem also arises outside biology. A production, trade, or financial network may predict that a policy intervention should not reach employment, asset risk, or output in a nominally unexposed sector. Since shocks can propagate through unmodeled network links, an abnormal response in such an indicator can signal spillovers, information leakage, or structural change (Acemoglu et al., 2012; 2016). In either setting, crossing the interval is a calibrated alarm that prioritizes follow-up; it is not, by itself, proof of a new causal edge. This interpretation follows the broader use of negative-control outcomes as diagnostics for model failure and unanticipated bias (Lipsitch et al., 2010).

We formalize this task as *safe-pair prediction*. Let

$$Z^{\circ}_{a,i} = \mathbf{1}\{i \in \mathrm{desc}_{\mathcal{G}^{\circ}}(a)\}$$

indicate whether outcome i is downstream of intervention a in a latent causal graph. Here $\mathrm{desc}_{\mathcal{G}^{\circ}}(a)$ denotes the descendants of a in the latent graph $\mathcal{G}^{\circ}$, and $\mathbf{1}\{\cdot\}$ is the indicator. A pair is safe when $Z^{\circ}_{a,i} = 0$. For a prespecified held-out test pair $(a^{\star}, i)$ asserted to be safe, we seek a prediction interval for $Y_i^{(a^{\star})}$, the outcome of target i under intervention $a^{\star}$. The method does not decide whether an arbitrary test pair is safe; it makes uncertainty quantification reliable under that scientific assertion.

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

If every historical safety label were known, one should retain only truly safe calibration observations and apply ordinary split conformal prediction (CP) within that stratum (Shafer & Vovk, 2008; Angelopoulos & Bates, 2023). This oracle regime is rarely available. A low-cost working graph supplies predicted-safe labels for all candidate pairs, while reliable verification is available for only a limited number. False-safe members may enter a selected calibration pool, and their scores need not be exchangeable with the safe test score (Asiaee et al., 2026).

The central issue is query locality. A common calibration data set may be reused across many queries, but a test intervention $a^\star$, target i, and graph distance define a different local subset for each query. In practice, graph distance can measure the shortest-path length between perturbed genes or drug targets in a working biological network, a proxy for mechanistic proximity (Guney et al., 2016). A prespecified radius τ limits how far this neighborhood extends. Consequently, the ordered calibration scores, the hidden unsafe count, its verification certificate, the selected conformal rank, and the final interval are all query-specific. A small global graph-error rate cannot protect a query whose local neighborhood contains a concentrated pathway omission; conversely, applying that global rate to a clean neighborhood may be unnecessarily conservative.

We introduce Certified Graph-Local Conformal Prediction (CGLCP). For every fixed query, CGLCP first selects a local predicted-safe pool without inspecting the final calibration or test scores. It verifies a simple-random subset of that pool, inverts the exact without-replacement sampling law to obtain an upper bound on the realized local unsafe count, and converts the certified safe mass into a conformal order-statistic rank. The prediction threshold is then the corresponding order statistic of the *observed calibration scores only*. The unknown test score is never inserted into the operational calibration sequence; it appears only in the rank argument that establishes coverage.

Our contributions are:

1. We formulate query-local safe-pair prediction and show why a global graph-error rate alone need not certify the contamination of an individual selected pool. This identifies local causal-membership contamination as the verification target, separately from observable heterogeneity.
2. Under conditional safe-score exchangeability, we derive an attainable worst-case coverage bound for arbitrary unsafe calibration scores and the smallest certificate-valid conformal rank. Minimality holds among deterministic order-statistic ranks measurable with respect to the certificate.
3. We turn limited random verification into an exact finite-population contamination certificate, establish end-to-end finite-sample coverage under the stated assumptions, and quantify certificate and rank error relative to a local oracle. Simulations and K562 experiments assess coverage, finite-output availability, verification cost, and negative-control alarm behavior.

2 Related Work

Selective and Causal Conformal Inference. Asiaee et al. (2026) introduce safe-pair selective conformal inference and characterize the coverage loss caused by false-safe calibration, establishing the problem background of prediction with imperfect causal structure. Motivated by this problem, we study how limited verification can certify the realized contamination of a query-specific calibration pool. CGLCP combines an exact finite-population certificate with an attainable contaminated-rank bound and its minimal valid rank, yielding a query-wise end-to-end coverage guarantee. Recent causal conformal methods (Schröder et al., 2026) instead address observational-to-interventional shifts under continuous treatments via generalized propensity scores; our focus is local causal-membership contamination induced by a fallible working graph.

Localized and Nonexchangeable Conformal Methods. Localized and nonexchangeable conformal procedures adapt to distributional heterogeneity or bound deviations from exchangeability using observable covariates, spatial localization, or weighted calibration (Guan, 2023; Hore & Barber, 2025; Tibshirani et al., 2019; Barber et al., 2023; Oliveira et al., 2024). For instance, our GraphLCP (graph-localized CP) baseline instantiates the exact localizer of Guan (2023) using graph distance. However, while these methods accommodate covariate-dependent heterogeneity, they do not account for latent causal-stratum membership contamination when graph-selected negative controls are misclassified.

Robust Conformal Prediction and Contamination. Robust conformal prediction under Huber contamination (Clarkson et al., 2024) or privileged side information (Feldman & Romano, 2024)

models corruption mechanisms or requires privileged covariates that render corruption conditionally ignorable. Similarly, label-cleaning methods (Bashari et al., 2025) aim to remove suspected outliers. In contrast, CGLCP requires neither a stochastic model for graph errors nor privileged covariates to model corruption. Instead, via simple random verification, it constructs a finite-sample upper certificate for the realized number of graph-misclassified unsafe pairs, coupling classical hypergeometric confidence limits (Wang, 2015; Bartroff et al., 2023) to a sharp causal-stratum rank rule and an end-to-end coverage guarantee.

3 Problem Setup

3.1 Global data roles and the score table

Partition interventions as $\mathcal{A} = \mathcal{A}_{\text{train}} \dot\cup \mathcal{A}_{\text{cal}} \dot\cup \mathcal{A}_{\text{test}}$, keeping all targets of each intervention in one split. The held-out pairs satisfy $\mathcal{T}_{\text{test}} \subseteq \mathcal{A}_{\text{test}} \times \mathcal{I}$, where $\mathcal{I}$ indexes targets and $\star$ identifies the current test intervention $a^\star$, not an oracle quantity. Training yields a predictor $\widehat{f}_i(a)$ and a learned or externally supplied working graph $\widehat{\mathcal{G}}$, both frozen before final calibration and test outcomes are inspected. The shared calibration score table contains, for $a \in \mathcal{A}_{\text{cal}}$,

$$R_i^{(a)} = \left| Y_i^{(a)} - \widehat{f}_i(a) \right|. \tag{1}$$

For a test pair $(a^\star, i)$, define the unobserved realized score

$$R_i^\star = \left| Y_i^{(a^\star)} - \widehat{f}_i(a^\star) \right|. \tag{2}$$

3.2 A different local calibration sequence for every query

The working graph supplies predicted descendant labels $\widehat{Z}_{a,i}$ (one for predicted unsafe, zero for predicted safe) and a fixed intervention distance $\widehat{d}(a, a^\star)$. For a prespecified radius τ, the test pair $(a^\star, i)$ selects the local predicted-safe pool

$$\mathcal{S}_i^\star = \mathcal{S}_\tau(a^\star, i) = \left\{ a \in \mathcal{A}_{\text{cal}} : \widehat{Z}_{a,i} = 0,\ \widehat{d}(a, a^\star) \le \tau \right\}, n_i^\star = |\mathcal{S}_i^\star|.$$

For shortest-path distance, τ limits network hops (or weighted path length). Here undirected shortest-path distance measures proximity between intervention genes in the working regulatory network, not effect-size similarity. A fixed binary inclusion rule is needed to identify the population whose unsafe count is certified; the distance threshold is our chosen rule, although other prespecified finite subsets are possible. Soft weights alone do not specify that local population. We separately examine distance weighting inside a fixed radius in Appendix E.6.

The selected score multiset is $\mathcal{R}_i^\star = \left\{ R_i^{(a)} : a \in \mathcal{S}_i^\star \right\}$, with ordered calibration values $R_{i,[j]}^\star$, distinct from the unobserved test score $R_i^\star$:

$$R_{i,[1]}^\star \le R_{i,[2]}^\star \le \cdots \le R_{i,[n_i^\star]}^\star, \qquad R_{i,[n_i^\star+1]}^\star := +\infty. \tag{3}$$

For a nonempty pool $n_i^\star \ge 1$, define its latent unsafe count $B_i^\star = \sum_{a \in \mathcal{S}_i^\star} Z_{a,i}^\circ$, safe count $M_i^\star = n_i^\star - B_i^\star$, and contamination fraction $\delta_i^\star = B_i^\star / n_i^\star$. Empty pools are handled in Appendix C.

Assumption 1 (Non-score query construction) *The predictor, working graph, distance, radius, local pool, and verification design for each query are fixed without inspecting the final scores in Eqs.* (1)–(2)*.*

Assumption 2 (Query-wise safe-score invariance) *Fix a test pair $(a^\star, i)$ with $Z_{a^\star,i}^\circ = 0$. Conditional on all non-score construction information, the latent memberships of $\mathcal{S}_i^\star$, and the independent verification information, the $M_i^\star$ truly safe calibration scores in $\mathcal{R}_i^\star$ and the test score $R_i^\star$ are exchangeable. The $B_i^\star$ unsafe calibration scores are arbitrary.*

Assumption 3 (Exact independent verification) *Every test pair covered by the main theorem has $n_i^\star \ge 1$ and a fixed budget $1 \le s_i^\star \le n_i^\star$. Conditional on the non-score construction information and*

the complete latent membership vector $(Z^\circ_{a,i} : a \in \mathcal{S}^\star_i)$, the subset $V^\star_i$ is drawn uniformly without replacement and independently of all calibration and test scores. The verification oracle returns the exact labels $Z^\circ_{a,i}$ for sampled members.

Remark 1 (Interpretation of Assumption 3) *With unsafe as positive, false positives inflate the upper certificate and can only widen intervals; false negatives can lower it and invalidate nominal coverage. If certificate failure has probability p, coverage remains at least $(1-p)(1-\beta)$ provided Assumption 2 holds given verification, with β from Eq.* (9)*. Exact labels can be cached for unchanged pairs and contexts: draw each query's uniform subset independently of scores and cached label values, then reuse answers. This preserves query-wise validity; total assay cost counts distinct sampled pairs rather than summed query budgets. Caching does not correct label noise. We separately test false-positive and false-negative verification errors in Appendix E.5.*

Graph proximity and safe membership do not ensure score exchangeability across batch, donor, dose, time, or cell state; these differences require predictor or design adjustments. We propagate a conditional swap-defect bound through Theorems 1 and 2 in Appendix B, distinguishing exchangeable heavy tails from context mismatch.

3.3 Conformal quantiles from calibration scores only

For any multiset $\mathcal{R} = \{r_1, \ldots, r_n\}$ with $n \geq 1$, ordered values $r_{[1]} \leq \cdots \leq r_{[n]}$ and sentinel $r_{[n+1]} = +\infty$, define the finite-sample split-conformal quantile operator, for $0 \leq \gamma < 1$, by

$$\mathrm{Quantile}^{\mathrm{conf}}_{1-\gamma}(\mathcal{R}) := r_{[\lceil (n+1)(1-\gamma) \rceil]}.$$

The factor $n+1$ follows from exchangeability of the calibration and test scores; the quantile uses calibration scores only.

3.4 Why global contamination cannot certify a local query

Suppose the global predicted-safe collection has N members and B_{all} unsafe members. A query-local subset of size n can have any contamination rate in the exact envelope

$$\max\left\{0, \frac{B_{\mathrm{all}} - (N-n)}{n}\right\} \leq \delta_{\mathrm{local}} \leq \min\left\{1, \frac{B_{\mathrm{all}}}{n}\right\}. \tag{4}$$

Both endpoints are attainable. Thus a global correction can under-correct concentrated local contamination and over-correct a clean neighborhood.

4 Sharp Effect of Query-Local Contamination

Theorem 1 (Sharp contaminated-rank coverage) *Fix a truly safe test pair $(a^\star, i)$. Under Assumption 2, for every $k \in \{1, \ldots, n^\star_i\}$, writing $[x]_+ = \max(x, 0)$,*

$$\mathbb{P}\Big\{R^\star_i \leq R^\star_{i,[k]}\Big\} \geq \frac{[k - B^\star_i]_+}{M^\star_i + 1}, \tag{5}$$

and coverage is one for $k = n^\star_i + 1$. With a continuous positive safe-score distribution, the bound is attained by placing every unsafe calibration score at zero.

The denominator in Eq. (5) counts the $M^\star_i$ safe calibration scores and the test score in the exchangeability argument.

For target miscoverage $0 < \alpha < 1$, consider the uncorrected rank $k^\star_{\alpha,i} = \lceil (n^\star_i + 1)(1-\alpha) \rceil$, Theorem 1 implies

$$\mathbb{P}\Big\{R^\star_i \leq R^\star_{i,[k^\star_{\alpha,i}]}\Big\} \geq 1 - \alpha - \frac{\alpha B^\star_i}{M^\star_i + 1}, \tag{6}$$

whenever the right side is positive. The bound accommodates both low-rank contaminants that narrow intervals and high-rank contaminants that widen them.

5 CERTIFIED GRAPH-LOCAL CONFORMAL PREDICTION

5.1 A TEST-PAIR-SPECIFIC CONTAMINATION CERTIFICATE

Under Assumption 3, draw $V_i^\star \subset \mathcal{S}_i^\star$ uniformly without replacement with size $s_i^\star$. The verified unsafe count $X_i^\star = \sum_{a \in V_i^\star} Z_{a,i}^\circ$ satisfies $X_i^\star \mid B_i^\star \sim \mathrm{Hypergeom}(n_i^\star, B_i^\star, s_i^\star)$, where Hypergeom denotes the hypergeometric distribution. Coverage requires an upper certificate, not the plug-in fraction $X_i^\star / s_i^\star$. Our verification-budget experiments show diminishing marginal gains around a verified fraction of 0.2–0.4, motivating 0.30 as the default unless otherwise specified (Appendices E.2 and F.2). Prespecify $s_i^\star / n_i^\star$ according to available reliable labels, assay capacity, and cost, while retaining uniform sampling from the full local pool. Let $H_{n,b,s}(x) = \mathbb{P}_b(X \le x)$ be the hypergeometric cumulative distribution function (CDF). Here n is the finite population size, b is its number of successes (unsafe members), and s is the sample size drawn uniformly without replacement; X counts sampled unsafe members. For certificate-failure allowance $0 < \rho < \alpha$, define

$$\overline{B}_i^\star = U_\rho(X_i^\star; n_i^\star, s_i^\star) := \max\{b \in \{0, \dots, n_i^\star\} : H_{n_i^\star, b, s_i^\star}(X_i^\star) \ge \rho\}. \tag{7}$$

Monotone inversion gives

$$\mathbb{P}_{B_i^\star}\{B_i^\star \le \overline{B}_i^\star\} \ge 1 - \rho. \tag{8}$$

Equivalently, $\overline{B}_i^\star = U_\rho\Big(\sum_{a \in V_i^\star} Z_{a,i}^\circ; n_i^\star, s_i^\star\Big)$, with $\overline{B}_i^\star \ne \overline{B}\Big(Y_i^{(a^\star)}\Big)$: randomness comes from verification, while graph errors may be adversarial.

5.2 FROM THE CERTIFICATE TO A TEST-PAIR-SPECIFIC CONFORMAL QUANTILE

Let $E_i^\star = \{B_i^\star \le \overline{B}_i^\star\}$ and define the largest conditional error budget compatible with total miscoverage α as

$$\beta = \frac{\alpha - \rho}{1 - \rho}, \qquad 0 < \rho < \alpha < 1. \tag{9}$$

For the fixed test pair, set

$$k_{\mathrm{cert},i}^\star = \overline{B}_i^\star + \left\lceil (1 - \beta)(n_i^\star - \overline{B}_i^\star + 1) \right\rceil, \tag{10}$$

Its effective miscoverage level is $\alpha_{\mathrm{cert},i}^\star = \beta \frac{n_i^\star - \overline{B}_i^\star + 1}{n_i^\star + 1}$, so $k_{\mathrm{cert},i}^\star = \lceil (n_i^\star + 1)(1 - \alpha_{\mathrm{cert},i}^\star) \rceil$.

The CGLCP threshold for test pair $(a^\star, i)$ is therefore

$$\widehat{q}_i(a^\star; V_i^\star) = R_{i,[k_{\mathrm{cert},i}^\star]}^\star = \mathrm{Quantile}_{1 - \alpha_{\mathrm{cert},i}^\star}^{\mathrm{conf}}(\mathcal{R}_i^\star)\,.$$

If $k_{\mathrm{cert},i}^\star = n_i^\star + 1$, the sentinel in Eq. (3) gives $\widehat{q}_i(a^\star; V_i^\star) = +\infty$.

Proposition 1 (Minimal certificate-valid rank) *Conditional on a realized certificate satisfying $B_i^\star \le \overline{B}_i^\star$, $k_{\mathrm{cert},i}^\star$ is the smallest certificate-measurable deterministic order-statistic rank that guarantees coverage at least $1 - \beta$ for every allocation and distribution of the unsafe scores. Any smaller rank fails for $B_i^\star = \overline{B}_i^\star$ when continuous safe scores are used and all unsafe scores occupy the lowest ranks.*

Operational procedure. We detail the complete procedure in Appendix C. Queries share the global calibration data but use their own pools, verification certificates, ranks, and intervals. Empty pools and sentinel ranks return infinite intervals, reported as non-finite outputs.

5.3 PREDICTION BY CANDIDATE-SCORE INVERSION

For a candidate outcome y, define $r_i^\star(y) = \left| y - \widehat{f}_i(a^\star) \right|$. Inverting the score yields

$$C_i(a^\star) = \{y : r_i^\star(y) \le \widehat{q}_i(a^\star; V_i^\star)\} = \left[\widehat{f}_i(a^\star) - \widehat{q}_i(a^\star; V_i^\star),\ \widehat{f}_i(a^\star) + \widehat{q}_i(a^\star; V_i^\star) \right]. \tag{11}$$

Coverage is equivalent to $R_i^\star \le \widehat{q}_i(a^\star; V_i^\star)$.

Theorem 2 (End-to-end query-wise coverage) *Under Assumptions 1–3 and* $0 < \rho < \alpha < 1$*, for every fixed test pair* $(a^\star, i)$ *with* $Z^\circ_{a^\star,i} = 0$*, the interval in Eq.* (11) *satisfies*

$$\mathbb{P}\Big\{Y_i^{(a^\star)} \in C_i(a^\star)\Big\} \geq 1 - \alpha.$$

The probability is joint over the random verification subset and the calibration/test scores, conditional on the fixed non-score construction information. The guarantee is per prespecified query; simultaneous coverage over an adaptively selected collection of queries requires an additional multiple-query procedure.

5.4 Quantifying certificate and rank error

Equation (8) controls under-certification. On the valid event, let $\Delta^\star_{B,i} = \overline{B}^\star_i - B^\star_i \geq 0$. Let $k^\star_{\rho,i}(B^\star_i)$ denote Eq. (10) with the true $B^\star_i$ inserted but the same certificate budget β. The rank overhead caused only by estimating $B^\star_i$ is

$$k^\star_{\mathrm{cert},i} - k^\star_{\rho,i}(B^\star_i) = \lfloor \beta(M^\star_i + 1) \rfloor - \lfloor \beta(M^\star_i - \Delta^\star_{B,i} + 1) \rfloor \leq \lceil \beta \Delta^\star_{B,i} \rceil.$$

Relative to an oracle that knows $B^\star_i$ and spends no certificate budget, the effective level loss decomposes as

$$\alpha^\star_{\mathrm{orc},i} - \alpha^\star_{\mathrm{cert},i} = \frac{\rho(1-\alpha)}{1-\rho} \frac{M^\star_i + 1}{n^\star_i + 1} + \frac{\beta \Delta^\star_{B,i}}{n^\star_i + 1}, \tag{12}$$

where $\alpha^\star_{\mathrm{orc},i} = \alpha(M^\star_i + 1)/(n^\star_i + 1)$. The first term is the price of certificate validity; the second is the price of over-certifying the local unsafe count.

6 Experiments

We evaluate whether local certification protects coverage as graph errors concentrate near a query, and what this protection costs in interval width, finite availability, and verification. Simulations isolate these mechanisms; K562 Perturb-seq assesses their practical consequences with assay-based reference labels. We compare CGLCP with Pooled CP, GraphLCP, PCS Corrected (partial-causal-structure correction), and Localized PCS Corrected, using a global-rate plug-in and a local oracle as diagnostic controls; we specify these methods and supplementary comparator sets in Appendix D. Unless stated otherwise, $\alpha = 0.10$ and $\rho = 0.02$.

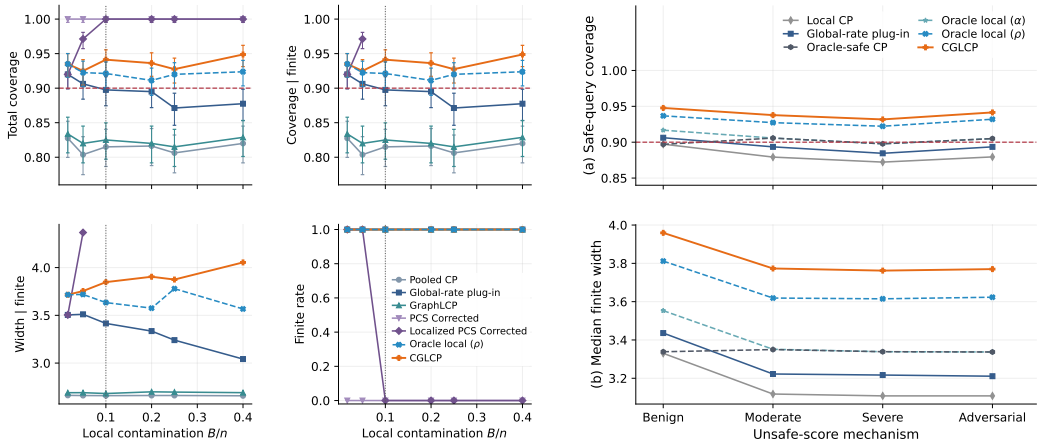
We report marginal coverage, finite-output rate, coverage conditional on finite output, finite-only interval width, and reference-label use, together with false alarms and detection sensitivity. Infinite intervals count as covered, so coverage must be read with finite availability. K562 widths average within-target finite-width medians; simulations use replicate medians unless a graph-seed average is specified. We detail data generation, processing, and experimental settings in Appendices E.1 and F.1.

6.1 Simulation

Data. We generate directed modular graphs with a fixed global false-safe fraction and vary the concentration of unsafe members in a local pool. Safe calibration and test scores share a half-normal law, while low-rank unsafe scores create adverse contamination. Repeated score draws and uniform verification isolate the effect of membership error from safe-score heterogeneity; we specify the graph, scores, and budgets in Appendix E.1.

Results. In Figure 1(a), as local contamination rises from $1/60$ to $24/60$ at a fixed global rate, global-rate plug-in coverage falls from 0.920 to 0.878, and Pooled CP and GraphLCP also under-cover. CGLCP maintains empirical coverage of 0.925–0.949 with fully finite output, although its intervals are wider than the local oracle's. The global PCS correction returns only infinite intervals, while its localized counterpart loses finite output from $6/60$ contamination onward. In this design, local certification improves the coverage–availability trade-off without uniformly reducing width.

In Figure 1(b), low-rank contaminants approach the sharp bound in Theorem 1, whereas higher-rank contaminants make the correction conservative. Exact enumeration over 90 configurations respects the certificate failure budget; increasing verification reduces excess certificate size, rank, and width



(a) Fixed global contamination, varying local risk. (b) Unsafe-score placement at the high-risk setting.

Figure 1: **Synthetic-data evaluation of local contamination effects and theoretical coverage bounds.** (a) Comparison with baseline methods as query-local contamination increases at a fixed global contamination rate, measured by coverage, finite-output rate, finite-conditional coverage, and interval width. (b) Comparison of CGLCP's empirical coverage with theoretical bounds under four unsafe-score placements in the high-risk setting.

relative to Oracle local (ρ), and full verification eliminates the quantile-rank gap between CGLCP and Oracle local (ρ), as shown in Figure 2. These results agree with the coverage and efficiency analysis under its assumptions. However, to probe practical failure boundaries, we also test safe-score mismatch and noisy verification in Appendices E.4, E.6, and E.5; Theorem 3, Corollary 1, and Proposition 2 quantify sensitivity and explain why membership certification alone cannot ensure coverage under unrestricted score mismatch.

The fixed intervals also provide negative-control alarms. At local contamination 0.30, verification fraction 0.20, and outward shift $\kappa = 2$ (in query-specific scale units), CGLCP combines a false-alarm rate of 0.0530 with recall 0.9105 and fully finite output; the global-rate plug-in and GraphLCP have higher recall but false-alarm rates above 0.10, while both PCS corrections abstain, as shown in Figure 14. Detection varies with width, verification effort, and abstention, as shown in Figure 15.

Additional experiments. To explain the trade-offs of our method, we vary verification fraction, ρ, and radius in Appendix E.2, and test graph seeds and added/deleted edges in Appendix E.3. We assess score mismatch and verification noise separately in Appendices E.4 and E.5, and examine within-radius weighting in Appendix E.6. We further report anomaly-strength and verification sensitivity in Appendix E.7.

6.2 K562 PERTURB-SEQ

Data. We use a prespecified regulator subset of the scPerturb-harmonized day-8 K562 screen (Replogle et al., 2022) and an outcome-independent TF (transcription factor) ChIP-seq (chromatin immunoprecipitation sequencing)/promoter graph. Intervention splits and separate batch folds isolate fitting, calibration, reference adjudication, tuning, and evaluation. The primary cohort contains 5,994 working-graph-predicted-safe, reference-safe test pairs, with $\tau = 3$ and planned verification fraction 0.30. Rounding up the verification count within each local calibration pool yields the mean realized fraction of 0.328 reported in Table 1. We detail selection, the frozen predictor, standardized scores, and assay-based reference labels in Appendix F.1; these labels are not exact causal memberships.

Results. In Figure 3(a), Pooled CP and GraphLCP have mean coverage 0.909 and 0.911, above $1 - \alpha = 0.90$, but their bootstrap confidence intervals extend below the target, leaving undercoverage compatible with sampling uncertainty. Their false-alarm rates reach about 0.13 in the highest-risk stratum in Figure 3(b). CGLCP attains coverage 0.978 and finite availability 0.613, compared with 0.996–0.997 and 0.101–0.109 for PCS corrections, as reported in Table 1. Although PCS corrections

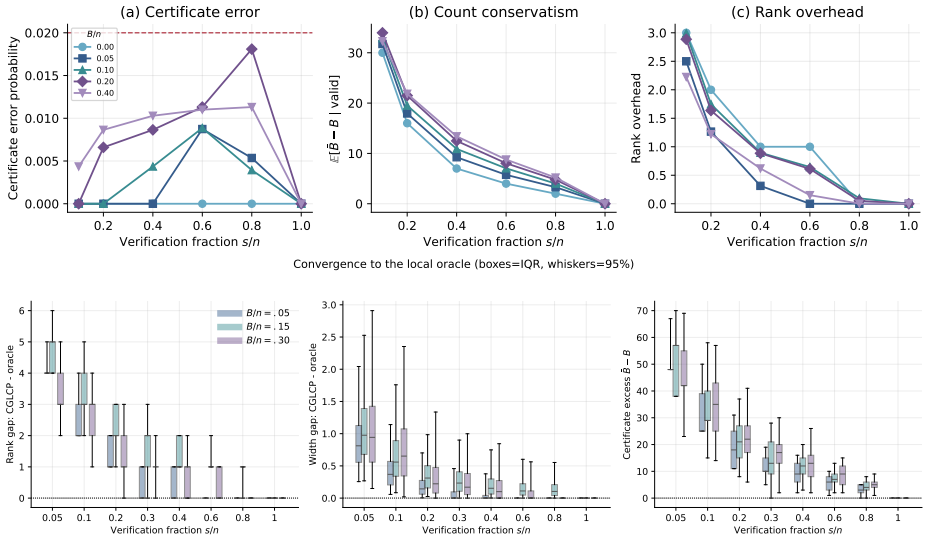


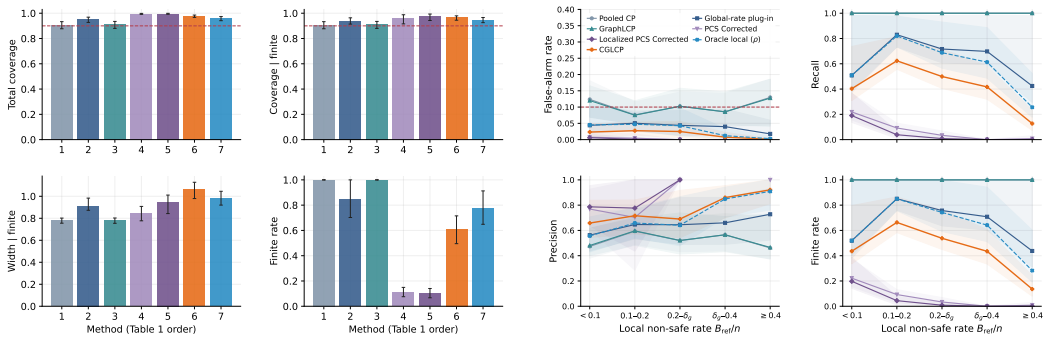
Figure 2: **Synthetic-data evaluation of certificate validity and convergence to the local oracle as verification increases.** Top: certificate failure probability, count conservatism, and rank overhead computed by exact hypergeometric enumeration, without Monte Carlo uncertainty. Bottom: paired certificate, rank, and width excess relative to Oracle local (ρ). Boxes indicate interquartile ranges, center lines indicate medians, and whiskers indicate empirical 2.5th–97.5th percentiles across replicates, rather than confidence intervals for a mean. Width comparisons use jointly finite pairs. Full exact verification eliminates certificate and rank excess.

and the global-rate plug-in yield smaller finite-only widths, PCS has lower finite availability than CGLCP; the plug-in attains 0.950 coverage and 0.845 availability while accessing all global working-safe reference labels.

The K562 results are qualitatively consistent with simulation: stronger local protection comes with wider intervals and reduced availability. As shown in Figure 17, CGLCP widens intervals or invokes the sentinel as certified risk increases. Compared with Oracle local (ρ), CGLCP verifies an average of 32.8% of its local pool, reducing label-acquisition requirements and potential assay costs relative to verifying the entire pool. Its larger width (1.061 versus 0.978) and lower finite-output rate (0.613 versus 0.777) reflect the conservatism introduced by this limited-verification certificate, illustrating the trade-off between verification cost and predictive efficiency in Table 1.

For detection, we evaluate frozen intervals after outward shifts of two query-specific scales, treating infinite intervals as abstention. In Figure 3(b), CGLCP trades lower recall for low false-alarm rates and relatively high precision at prevalence 0.10; this contrast with Pooled CP and GraphLCP is more pronounced at high local contamination. In the highest-risk stratum, its false-alarm rate is 0.001 and precision 0.921, with finite availability 0.136. This pattern is consistent with the method's emphasis on safe-pair coverage, but abstention contributes to low recall and false alarms; coverage alone guarantees neither precision nor detection power. We further compare false-positive-rate (FPR)-matched detection and 32 definitively reference-unsafe pairs in Appendix F.3; GraphLCP can have greater ranking power, as shown in Figure 19.

Additional experiments. Following the simulation diagnostics, we examine local-risk strata and repeated verification in Appendix F.2, and extend the shift and reference-adjudicated detection analyses in Appendix F.3. We assess graph, radius, target-universe, and baseline sensitivity in Appendix F.4. Finally, we evaluate certificate-based prioritization for independent follow-up in Appendix F.5, without claiming validated edge repair.



(a) Coverage–finiteness–cost comparison. (b) Negative-control sentinel operating metrics.

Figure 3: **Prediction and detection performance on the K562 reference-safe cohort.** (a) Comparison with baseline methods in coverage, finite-conditional coverage, interval width, and finite-output rate. (b) Negative-control detection performance across local contamination strata, using outward shifts of $\kappa = 2$ and precision evaluated at prevalence 0.10. Error bars and shaded bands indicate pointwise 95% bootstrap confidence intervals, resampling query interventions and target genes. Width is the target-balanced mean of within-target finite-width medians. Infinite intervals count as covered for total coverage and as abstentions for detection.

Table 1: **Overall predictive performance and reference-label requirements on the K562 reference-safe cohort.** The same target-balanced point estimates are plotted with confidence intervals in Figure 3(a). The categories in the final column indicate the scope and proportion of ground-truth label information used by each method, corresponding to different label-verification requirements and associated experimental costs in practical applications.

Method	Coverage	Finite-output rate	Coverage given finite output	Interval width	Reference-label use
Pooled CP	.909	1.000	.909	.780	None
Global-rate plug-in	.950	.845	.941	.914	All global
GraphLCP	.911	1.000	.911	.780	None
PCS Corrected	.996	.109	.961	.842	All selected
Localized PCS Corrected	.997	.101	.976	.949	All local
CGLCP	.978	.613	.967	1.061	.328 local
Oracle local (ρ)	.959	.777	.948	.978	Oracle

7 Discussion

CGLCP couples finite-population verification with a sharp conformal rank to protect prespecified safe queries against local membership contamination. Its guarantee depends on exact independent verification and safe-score exchangeability; the experiments assess the resulting coverage–width–availability trade-off and illustrate failure cases when these assumptions break. A promising direction of the future work is to reduce interval width through more efficient label acquisition, drawing on active statistical inference and robust sampling (Zrnic & Candes, 2024; Li et al., 2025); nonuniform verification would require a new certificate rather than direct reuse of the hypergeometric law. Budgeted reference cleaning for conformal outlier detection (Bashari et al., 2025) also motivates studying detection power while preserving false-alarm control under explicit noise assumptions. Besides, recent interventional network discovery in K562 (Brown et al., 2025) suggests a route from certificate-based triage to candidate graph revisions, with independent confirmation and renewed calibration after each update.

ETHICS STATEMENT

The K562 analysis uses previously released, de-identified cell-line data and public regulatory annotations. Reported alarms are research diagnostics and must not be interpreted as clinical decisions or established causal mechanisms.

REPRODUCIBILITY STATEMENT

We provide proofs and sensitivity results in Appendices A–B, and specify the algorithm, experimental settings, data splits, baselines, uncertainty procedures, and robustness analyses in Appendices C–F. The code to reproduce our results is available in the anonymous CGLCP repository.

REFERENCES

Daron Acemoglu, Vasco M. Carvalho, Asuman Ozdaglar, and Alireza Tahbaz-Salehi. The network origins of aggregate fluctuations. *Econometrica*, 80(5):1977–2016, 2012.

Daron Acemoglu, Ufuk Akcigit, and William Kerr. Networks and the macroeconomy: An empirical exploration. *NBER Macroeconomics Annual*, 30:273–335, 2016.

Anastasios N. Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *Annual Review of Statistics and Its Application*, 10: 1–25, 2023.

Amir Asiaee, Kaveh Aryan, and James P. Long. Partial causal structure learning for valid selective conformal inference under interventions. In *Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence*, volume 337 of *Proceedings of Machine Learning Research*, pp. 161–184. PMLR, 2026.

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.

Jay Bartroff et al. Exact confidence bounds for the hypergeometric distribution. *The American Statistician*, 2023.

Mahdi Bashari et al. Robust conformal outlier detection with limited clean labels. *arXiv preprint*, 2025.

Brielin C. Brown, Alex Tokolyi, John A. Morris, Tuuli Lappalainen, and David A. Knowles. Large-scale causal discovery using interventional data sheds light on gene network structure in K562 cells. *Nature Communications*, 16:9628, 2025. doi: 10.1038/s41467-025-64353-7. URL https://www.nature.com/articles/s41467-025-64353-7.

Jase Clarkson et al. Conformal prediction under contaminated calibration data. *arXiv preprint*, 2024.

Atray Dixit, Oren Parnas, Biyu Li, et al. Perturb-seq: Dissecting molecular circuits with scalable single-cell rna profiling of pooled genetic screens. *Cell*, 167(7):1853–1866.e17, 2016.

Shai Feldman and Yaniv Romano. Robust conformal prediction using privileged information. *Advances in Neural Information Processing Systems*, 37:117813–117852, 2024.

Leying Guan. Localized conformal prediction: A generalized inference framework for conformal prediction. *Biometrika*, 110(1):33–50, 2023. doi: 10.1093/biomet/asac040.

Emre Guney, Jörg Menche, Marc Vidal, and Albert-László Barabási. Network-based in silico drug efficacy screening. *Nature Communications*, 7:10331, 2016. doi: 10.1038/ncomms10331.

Rohan Hore and Rina Foygel Barber. Conformal prediction with local weights: Randomization enables local guarantees. *Journal of the Royal Statistical Society, Series B*, 87(2):549–578, 2025.

Puheng Li, Tijana Zrnic, and Emmanuel Candes. Robust sampling for active statistical inference. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-2312. URL https://papers.nips.cc/paper_files/paper/2025/hash/6389470564214983604d1ac81631c2c5-Abstract-Conference.html.

Marc Lipsitch, Eric Tchetgen Tchetgen, and Ted Cohen. Negative controls: A tool for detecting confounding and bias in observational studies. *Epidemiology*, 21(3):383–388, 2010.

Roberto I. Oliveira, Paulo Orenstein, Thiago Ramos, and Yaniv Romano. Split conformal prediction and non-exchangeable data. *Journal of Machine Learning Research*, 25:1–38, 2024.

Joseph M. Replogle, Reuben A. Saunders, Angela N. Pogson, et al. Mapping information-rich genotype-phenotype landscapes with genome-scale perturb-seq. *Cell*, 185(14):2559–2575.e28, 2022.

Maresa Schröder, Dennis Frauen, Jonas Schweisthal, Konstantin Heß, Valentyn Melnychuk, and Stefan Feuerriegel. Conformal prediction for causal effects of continuous treatments. *Advances in Neural Information Processing Systems*, 38:1500–1539, 2026.

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

Sanjay R. Srivatsan, José L. McFaline-Figueroa, Vijay Ramani, et al. Massively multiplex chemical transcriptomics at single-cell resolution. *Science*, 367(6473):45–51, 2020.

Aravind Subramanian, Rajiv Narayan, Steven M. Corsello, et al. A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. *Cell*, 171(6):1437–1452.e17, 2017.

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

Weizhen Wang. Exact confidence intervals for a hypergeometric parameter. *Journal of Statistical Planning and Inference*, 2015.

Tijana Zrnic and Emmanuel Candes. Active statistical inference. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 62993–63010, 2024. URL `https://proceedings.mlr.press/v235/zrnic24a.html`.

A PROOFS

A.1 GLOBAL-TO-LOCAL ENVELOPE

A local subset cannot contain more than all B_{all} global contaminants or more than its own n elements, yielding the upper endpoint in Eq. (4). Its complement holds at most $N - n$ contaminants, so the subset contains at least $B_{\mathrm{all}} - (N - n)$, truncated at zero. Filling the subset or complement first attains both endpoints.

A.2 PROOF OF THEOREM 1

Fix test pair $(a^\star, i)$ and place the $B_i^\star$ arbitrary scores in the lowest ranks, which minimizes $R^\star_{i,[k]}$. If $k \le B_i^\star$, the distribution-free lower bound is zero. Otherwise $R^\star_{i,[k]}$ is no smaller than the $(k - B_i^\star)$th ordered safe calibration score. Exchangeability makes the test rank among the $M_i^\star + 1$ safe calibration/test scores uniform after randomized tie breaking, giving $[k - B_i^\star]_+/(M_i^\star + 1)$. With continuous safe scores, placing every unsafe score strictly below every safe score attains equality. The $k = n_i^\star + 1$ case follows from the infinity sentinel.

For $k^\star_{\alpha,i} = \lceil (n_i^\star + 1)(1 - \alpha) \rceil$,

$$\begin{aligned}\frac{k^\star_{\alpha,i} - B_i^\star}{M_i^\star + 1} &\ge \frac{(n_i^\star + 1)(1 - \alpha) - B_i^\star}{M_i^\star + 1} \\ &= 1 - \alpha - \frac{\alpha B_i^\star}{M_i^\star + 1},\end{aligned}$$

which proves Eq. (6) when positive.

A.3 VALIDITY OF THE QUERY-SPECIFIC FINITE-POPULATION CERTIFICATE

For a fixed true count $B_i^\star$, $H_{n_i^\star, b, s_i^\star}(x)$ is nonincreasing in b. Therefore $U_\rho(X_i^\star; n_i^\star, s_i^\star) < B_i^\star$ implies $H_{n_i^\star, B_i^\star, s_i^\star}(X_i^\star) < \rho$. If x_0 is the largest integer satisfying $H_{n_i^\star, B_i^\star, s_i^\star}(x_0) < \rho$, then

$$\mathbb{P}_{B_i^\star}\{H_{n_i^\star, B_i^\star, s_i^\star}(X_i^\star) < \rho\} = H_{n_i^\star, B_i^\star, s_i^\star}(x_0) < \rho.$$

If no such integer exists, the probability is zero. Hence $\mathbb{P}_{B_i^\star}(B_i^\star > \overline{B}_i^\star) \le \rho$.

A.4 PROOF OF PROPOSITION 1

On $B_i^\star \le \overline{B}_i^\star$, Theorem 1 is minimized at $B_i^\star = \overline{B}_i^\star$ because $(k - B)/(n_i^\star - B + 1)$ is nonincreasing in B for $k \le n_i^\star + 1$. Coverage at least $1 - \beta$ therefore requires

$$k \ge \overline{B}_i^\star + (1 - \beta)(n_i^\star - \overline{B}_i^\star + 1).$$

The smallest integer satisfying this inequality is Eq. (10). Any smaller rank fails when $B_i^\star = \overline{B}_i^\star$ and all unsafe scores lie below continuous safe scores.

The effective-level representation follows from

$$\begin{aligned}k^\star_{\mathrm{cert},i} &= n_i^\star + 1 - \left\lfloor \beta(n_i^\star - \overline{B}_i^\star + 1) \right\rfloor \\ &= \left\lceil (n_i^\star + 1)\left\{1 - \beta\frac{n_i^\star - \overline{B}_i^\star + 1}{n_i^\star + 1}\right\}\right\rceil.\end{aligned}$$

A.5 PROOF OF THEOREM 2

Fix the complete latent membership vector $(Z^\circ_{a,i} : a \in \mathcal{S}_i^\star)$ and all non-score construction information, and let $\mathcal{H}^\star_{\mathrm{ver},i}$ contain $V_i^\star$ and its exact queried labels. The hypergeometric certificate is valid uniformly for every fixed membership vector. Let $E_i^\star = \{B_i^\star \le \overline{B}_i^\star\}$ and $p_i^\star = \mathbb{P}\big((E_i^\star)^c\big) \le \rho$. For

every realization of $\mathcal{H}^\star_{\mathrm{ver},i}$ on which $E^\star_i$ holds, Proposition 1 and Assumption 2 give conditional miscoverage at most β. Integrating over the verification record gives

$$\begin{aligned}\mathbb{P}(\mathrm{miss}) &\le \mathbb{P}(\mathrm{miss}, E^\star_i) + \mathbb{P}\big((E^\star_i)^c\big)\\ &\le \beta\mathbb{P}(E^\star_i) + p^\star_i = \beta(1-p^\star_i) + p^\star_i\\ &\le \beta + (1-\beta)\rho = \alpha.\end{aligned}$$

Because the inequality holds uniformly for every fixed latent membership vector, integrating over that vector preserves the same bound.

A.6 Efficiency-error identities

Write $\Delta^\star_{B,i} = \overline{B}^\star_i - B^\star_i$ and $M^\star_i = n^\star_i - B^\star_i$. From the floor representation,

$$k^\star_{\mathrm{cert},i} - k^\star_{\rho,i}(B^\star_i) = \lfloor \beta(M^\star_i + 1)\rfloor - \lfloor \beta(M^\star_i - \Delta^\star_{B,i} + 1)\rfloor,$$

which is nonnegative and at most $\lceil \beta\Delta^\star_{B,i}\rceil$. The ideal oracle and certified effective levels are

$$\alpha^\star_{\mathrm{orc},i} = \alpha\frac{M^\star_i + 1}{n^\star_i + 1}, \qquad \alpha^\star_{\mathrm{cert},i} = \beta\frac{M^\star_i - \Delta^\star_{B,i} + 1}{n^\star_i + 1}.$$

Using $\alpha - \beta = \rho(1-\alpha)/(1-\rho)$ gives Eq. (12).

B Sensitivity Extensions of the Main Coverage Results

B.1 Scope and conditional information

This appendix supplements Theorem 1 and Theorem 2; it does not introduce a second estimation algorithm. The first extension adds a safe-score nonexchangeability term to the contaminated-rank bound. Its corollary propagates that term through the existing CGLCP certificate. Membership contamination and safe-score mismatch are distinct error sources, although they need not be statistically independent. Exact membership verification controls the former and does not estimate the latter. The swap argument below specializes the total-variation approach to nonexchangeable conformal prediction (Barber et al., 2023) to the truly safe subvector of a query-local contaminated pool.

Fix a truly safe query $(a^\star, i)$ and a nonempty pool, $n^\star_i \ge 1$. Retain Assumptions 1 and 3, while replacing Assumption 2 by the sensitivity conditions stated below. Let $\mathcal{F}^\star_i$ contain the fitted predictor, working graph, split, selected pool, complete true membership vector, and all other fixed non-score construction information. Let

$$\mathcal{H}^\star_i = \sigma(\mathcal{F}^\star_i, V^\star_i, \{Z^\circ_{a,i} : a \in V^\star_i\})$$

also contain the verification record. Here $\sigma(\cdot)$ denotes the generated sigma-algebra. Neither information set contains the realized final calibration scores or test outcome. All probabilities in this appendix are conditional on $\mathcal{F}^\star_i$ unless conditioning on $\mathcal{H}^\star_i$ is displayed explicitly. Thus $n^\star_i$, $B^\star_i$, and $M^\star_i$ are fixed; the verification record and scores remain random. Assumption 3 preserves the certificate validity under arbitrary safe-score heterogeneity.

Write the $M^\star_i$ truly safe calibration scores in a fixed order as $T_1, \dots, T_{M^\star_i}$ and put $T_0 = R^\star_i$. Let $P^\star_i$ be the conditional joint law of $\mathbf{T} = (T_0, \dots, T_{M^\star_i})$ given $\mathcal{H}^\star_i$. For $j = 0, \dots, M^\star_i$, let π_j swap coordinates 0 and j (π_0 is the identity), and let $P^{\star,(j)}_i = P^\star_i \circ \pi_j^{-1}$. With $d_{\mathrm{TV}}(P,Q) = \sup_A |P(A) - Q(A)|$, define

$$\varepsilon^\star_i := \frac{1}{M^\star_i + 1}\sum_{j=0}^{M^\star_i} d_{\mathrm{TV}}(P^\star_i, P^{\star,(j)}_i) \in \left[0, \frac{M^\star_i}{M^\star_i + 1}\right]. \tag{13}$$

This conditional-law quantity is not an estimate computed from observed scores. A bound $\bar{\varepsilon}^\star_i$ below always means an $\mathcal{F}^\star_i$-fixed, nonnegative upper bound satisfying $\varepsilon^\star_i \le \bar{\varepsilon}^\star_i$ almost surely over verification records. It must not be confused with a realized sample diagnostic. Under exact conditional exchangeability $\varepsilon^\star_i = 0$. For $M^\star_i = 0$, the definition gives $\varepsilon^\star_i = 0$; this degenerate case supplies no finite-rank information.

B.2 Extending the contaminated-rank bound

Lemma 1 (Approximate safe-rank bound) *Suppose $M_i^\star \geq 1$, and let $T_{[r]}^{\mathrm{safe}}$ be the ordinary rth order statistic of $T_1, \ldots, T_{M_i^\star}$. For $1 \leq r \leq M_i^\star$,*

$$\mathbb{P}\{T_0 \leq T_{[r]}^{\mathrm{safe}} \mid \mathcal{H}_i^\star\} \geq \left[\frac{r}{M_i^\star + 1} - \varepsilon_i^\star\right]_+ . \tag{14}$$

The inequality allows ties without randomizing the implemented interval. For the sentinel $T_{[M_i^\star+1]}^{\mathrm{safe}} = +\infty$, coverage is one.

Theorem 3 (Sensitivity extension of Theorem 1) *For every $\mathcal{H}_i^\star$-measurable rank $k \in \{1, \ldots, n_i^\star\}$ and arbitrary unsafe calibration scores,*

$$\mathbb{P}\{R_i^\star \leq R_{i,[k]}^\star \mid \mathcal{H}_i^\star\} \geq \left[\frac{[k - B_i^\star]_+}{M_i^\star + 1} - \varepsilon_i^\star\right]_+ . \tag{15}$$

For $k = n_i^\star + 1$, coverage is one. When $M_i^\star = 0$, the finite-rank lower bound is zero. Setting $\varepsilon_i^\star = 0$ recovers the rank inequality of Theorem 1.

For the uncorrected rank $k_{\alpha,i}^\star = \lceil (n_i^\star + 1)(1-\alpha) \rceil$, the same algebra as in Eq. (6) gives

$$\begin{aligned} &\mathbb{P}\{R_i^\star \leq R_{i,[k_{\alpha,i}^\star]}^\star \mid \mathcal{H}_i^\star\} \\ &\qquad \geq \left[1 - \alpha - \frac{\alpha B_i^\star}{M_i^\star + 1} - \varepsilon_i^\star\right]_+ . \end{aligned} \tag{16}$$

The contamination term and the safe-score mismatch term now appear in the same lower bound. This is a sensitivity extension of the sharp exchangeable result; no claim is made that the new total-variation bound is itself sharp for every law.

B.3 Extending the end-to-end CGLCP guarantee

Corollary 1 (Sensitivity extension of Theorem 2) *Use the original CGLCP rank in Eq.* (10)*, with $0 < \rho < \alpha < 1$ and $\beta = (\alpha - \rho)/(1 - \rho)$. If $\varepsilon_i^\star \leq \bar{\varepsilon}_i^\star$ almost surely and $0 \leq \bar{\varepsilon}_i^\star \leq 1$, then*

$$\begin{aligned} &\mathbb{P}\{Y_i^{(a^\star)} \in C_i(a^\star)\} \\ &\qquad \geq [1 - \alpha - (1 - \rho)\bar{\varepsilon}_i^\star]_+ . \end{aligned} \tag{17}$$

In particular, the shortfall below the nominal target satisfies

$$\left[(1 - \alpha) - \mathbb{P}\{Y_i^{(a^\star)} \in C_i(a^\star)\}\right]_+ \leq (1 - \rho)\bar{\varepsilon}_i^\star .$$

Corollary 1 concerns the original algorithm without an additional correction. It bounds possible undercoverage relative to the nominal target; it does not assert that the actual coverage decreases by exactly this amount, or that the difference from an exchangeable reference experiment equals this bound. Increasing the verification budget improves membership certification but need not reduce the safe-score defect.

Optional correction with an externally justified bound. If $0 \leq \bar{\varepsilon}_i^\star < \beta$ is a valid bound fixed as specified above, define

$$\begin{aligned} \beta_{\mathrm{het},i}^\star &:= \beta - \bar{\varepsilon}_i^\star , \\ k_{\mathrm{cert,het},i}^\star &:= \overline{B}_i^\star + \left\lceil (1 - \beta_{\mathrm{het},i}^\star)(n_i^\star - \overline{B}_i^\star + 1) \right\rceil . \end{aligned} \tag{18}$$

Use the same observed score sequence, but the new rank, to define

$$\begin{aligned} \widehat{q}_{\mathrm{het},i}(a^\star; V_i^\star) &:= R_{i,[k_{\mathrm{cert,het},i}^\star]}^\star , \\ C_{\mathrm{het},i}(a^\star) &:= \{y : |y - \widehat{f}_i(a^\star)| \leq \widehat{q}_{\mathrm{het},i}(a^\star; V_i^\star)\}. \end{aligned}$$

Then $\mathbb{P}\{Y_i^{(a^\star)} \in C_{\mathrm{het},i}(a^\star)\} \geq 1-\alpha$. The rank lies in $\{1,\ldots,n_i^\star+1\}$, with the usual infinity sentinel. If $\bar{\varepsilon}_i^\star \geq \beta$, explicitly set the rank to $n_i^\star+1$ instead of applying the displayed formula outside its stated range. At equality the formula already gives that sentinel.

The optional correction is a sensitivity calculation, not a new claim that CGLCP estimates $\bar{\varepsilon}_i^\star$. Fixing a numerical value in advance is insufficient: it must bound the relevant conditional laws. If an independent pilot procedure supplies only a high-probability bound, its failure probability requires an additional error allocation; the guarantee above cannot be applied unchanged. Also, the corrected rank can be conservative: we claim sufficiency, not minimality under arbitrary nonexchangeability.

B.4 Heavy tails, context shifts, and the limits of safety labels

A common heavy-tailed distribution does not in itself invalidate conformal coverage. For example, suppose, *conditional on the same $\mathcal{H}_i^\star$ used above*,

$$T_j = \sigma_j|\xi_j|, \qquad (\sigma_j,\xi_j) \overset{\mathrm{iid}}{\sim} Q, \quad j=0,\ldots,M_i^\star,$$

where Q may be heavy tailed. Here $\sigma_j \geq 0$ is a random scale, ξ_j is base noise, and iid means independent and identically distributed. The conditional scores are exchangeable and $\varepsilon_i^\star = 0$. Merely having the same unconditional mixture law is insufficient for a query-conditional guarantee: fixing a test module or selecting calibration members by distance can change the conditional mixture weights.

For a contrasting example, suppose the scores are independent *conditional on $\mathcal{H}_i^\star$*, with conditional laws F_0 for the test score and F_j for safe calibration score j. Then

$$\varepsilon_i^\star \leq \min\left\{\frac{M_i^\star}{M_i^\star+1}, \frac{2}{M_i^\star+1}\sum_{j=1}^{M_i^\star} d_{\mathrm{TV}}(F_j,F_0)\right\}. \tag{19}$$

Systematic distance, module, or batch mismatch can therefore affect the sensitivity bound even with no membership errors. Graph distance helps with this additional error source when it also improves score-distribution matching; topology alone does not guarantee a small defect. Reducing τ may improve matching while reducing the number of available calibration scores.

Proposition 2 (No finite-rank guarantee from safety labels alone) *Even when $B_i^\star = 0$ is known exactly, no rank $k \leq n_i^\star$ guarantees positive coverage uniformly over unrestricted safe-score mismatch.*

This statement permits arbitrary score laws across truly safe pairs. It does not contradict invariance in a homogeneous causal model with a common measurement and sampling regime. Such additional assumptions constrain these laws; causal membership alone does not impose them.

B.5 Additional proofs

Proof of Lemma 1. Introduce iid continuous tie breakers $U_0,\ldots,U_{M_i^\star}$ independent of the scores and of $\mathcal{H}_i^\star$, for this proof only. Order the augmented pairs (T_j,U_j) lexicographically. For any realized augmented vector, exactly r of the $M_i^\star+1$ swaps put coordinate zero among the lowest r pairs. If A_r is this event and $\widetilde{P}_i^\star$ denotes the augmented law, then

$$\frac{1}{M_i^\star+1}\sum_{j=0}^{M_i^\star} \widetilde{P}_i^{\star,(j)}(A_r) = \frac{r}{M_i^\star+1}.$$

The iid tie-breaker law is invariant to swaps and independent of the scores, so

$$d_{\mathrm{TV}}(\widetilde{P}_i^\star, \widetilde{P}_i^{\star,(j)}) = d_{\mathrm{TV}}(P_i^\star, P_i^{\star,(j)}).$$

Apply the defining total-variation inequality to A_r for each j and average. This gives $\widetilde{P}_i^\star(A_r) \geq r/(M_i^\star+1) - \varepsilon_i^\star$. Finally, A_r implies $T_0 \leq T_{[r]}^{\mathrm{safe}}$ for the ordinary, unrandomized score threshold. Truncating at zero proves the lemma. ∎

Proof of Theorem 3. The conditioning fixes k. If $k \le B_i^\star$, the claimed lower bound is zero. Otherwise at most $B_i^\star$ of the first k mixed scores are unsafe, so

$$R^\star_{i,[k]} \ge T^{\mathrm{safe}}_{[k-B_i^\star]}.$$

For a finite rank, $1 \le k - B_i^\star \le M_i^\star$; apply Lemma 1. If $M_i^\star = 0$, all finite ranks satisfy $k \le B_i^\star$ and the bound is zero. The sentinel covers surely. For Eq. (16), substitute $k^\star_{\alpha,i} \ge (n_i^\star + 1)(1-\alpha)$ in the finite-rank bound and simplify; the sentinel case also satisfies the displayed inequality. ∎

Proof of Corollary 1. Let $E_i^\star = \{B_i^\star \le \overline{B}_i^\star\}$ and $p_i^\star = \mathbb{P}((E_i^\star)^c) \le \rho$. The certificate event is measurable with respect to $\mathcal{H}_i^\star$. On this event, write $\Delta^\star_{B,i} = \overline{B}_i^\star - B_i^\star \ge 0$. For the original CGLCP rank,

$$\begin{aligned} k^\star_{\mathrm{cert},i} - B_i^\star &\ge \Delta^\star_{B,i} + (1-\beta)(M_i^\star - \Delta^\star_{B,i} + 1) \\ &= (1-\beta)(M_i^\star + 1) + \beta\Delta^\star_{B,i} \\ &\ge (1-\beta)(M_i^\star + 1). \end{aligned}$$

If the rank is finite, Theorem 3 gives conditional coverage at least $[1 - \beta - \bar{\varepsilon}_i^\star]_+$. If it is the sentinel, coverage one is at least the same bound. This also handles $M_i^\star = 0$: on $E_i^\star$, $\overline{B}_i^\star = B_i^\star = n_i^\star$, so the algorithm uses the sentinel. Integrating over the verification record, without requiring independence of the certificate event and the coverage event, gives

$$\begin{aligned} \mathbb{P}(\mathrm{cover}) &\ge \mathbb{P}(\mathrm{cover}, E_i^\star) \\ &\ge (1-p_i^\star)[1-\beta-\bar{\varepsilon}_i^\star]_+ \\ &\ge (1-\rho)[1-\beta-\bar{\varepsilon}_i^\star]_+ \\ &= [1-\alpha-(1-\rho)\bar{\varepsilon}_i^\star]_+. \end{aligned}$$

The identity uses $(1-\rho)(1-\beta) = 1-\alpha$. The shortfall inequality follows; if its right side exceeds $1-\alpha$, it holds trivially. ∎

Justification of the optional correction. Put $c_i^\star = 1 - \beta^\star_{\mathrm{het},i} = 1 - \beta + \bar{\varepsilon}_i^\star \in (0,1)$. On $E_i^\star$, let $\Delta^\star_{B,i} = \overline{B}_i^\star - B_i^\star \ge 0$. The ceiling in Eq. (18) yields

$$\begin{aligned} k^\star_{\mathrm{cert,het},i} - B_i^\star &\ge \Delta^\star_{B,i} + c_i^\star(M_i^\star - \Delta^\star_{B,i} + 1) \\ &= c_i^\star(M_i^\star + 1) + (1-c_i^\star)\Delta^\star_{B,i} \\ &\ge c_i^\star(M_i^\star + 1). \end{aligned}$$

For a finite rank, Theorem 3 and the assumed defect bound therefore give

$$\begin{aligned} &\mathbb{P}\{Y_i^{(a^\star)} \in C_{\mathrm{het},i}(a^\star) \mid \mathcal{H}_i^\star\} \\ &\quad \ge c_i^\star - \varepsilon_i^\star \\ &\quad = 1 - \beta + \bar{\varepsilon}_i^\star - \varepsilon_i^\star \\ &\quad \ge 1 - \beta. \end{aligned}$$

The sentinel case gives coverage one. Thus the same lower bound holds for every valid verification record. Integrating and discarding the invalid-certificate event gives

$$\begin{aligned} \mathbb{P}\{Y_i^{(a^\star)} \in C_{\mathrm{het},i}(a^\star)\} &\ge (1-p_i^\star)(1-\beta) \\ &\ge (1-\rho)(1-\beta) = 1-\alpha. \end{aligned}$$

This is an error-budget calculation: the larger rank supplies an additional $\bar{\varepsilon}_i^\star$ in the safe-rank lower bound, which absorbs the possible loss $\varepsilon_i^\star$ before the certificate failure allowance is applied. ∎

Proof of Eq. (19). Conditional independence makes the joint law a product. A swap changes only the two factors F_0 and F_j. Inserting $F_0 \otimes F_0$ and applying the triangle inequality gives

$$d_{\mathrm{TV}}(F_0 \otimes F_j, F_j \otimes F_0) \le 2 d_{\mathrm{TV}}(F_j, F_0).$$

Take the average in Eq. (13) and use its upper bound $M_i^\star/(M_i^\star + 1)$. ∎

Proof of Proposition 2. Let every safe calibration score be zero and the safe test score be one almost surely. All memberships are safe, but $R^\star_{i,[k]} = 0 < R_i^\star = 1$ for each $k \le n_i^\star$. Every finite-rank interval under consideration has zero coverage. ∎

C COMPLETE CGLCP PROCEDURE

Algorithm 1 applies the same query-wise certificate to a frozen score specification. Fit the predictor on training interventions and freeze the graph, score scales, radius, and verification budgets before inspecting final calibration or test scores. For absolute residuals, candidate-score inversion gives the symmetric interval in Eq. (11); the K562 standardized-score inversion is specified in Appendix F.1.

Algorithm 1 Certified Graph-Local Conformal Prediction

Require: Prespecified queries $\mathcal{T}_{\mathrm{test}}$; frozen predictor, graph, distance, and score rule; calibration data, radius τ, valid budgets $1 \le s_i^\star \le n_i^\star$ for nonempty pools, exact oracle, and $0 < \rho < \alpha < 1$.

Ensure: Intervals $C_i(a^\star)$ and query-wise verification records.

Set $\beta \leftarrow (\alpha - \rho)/(1 - \rho)$.
for each $(a^\star, i) \in \mathcal{T}_{\mathrm{test}}$ **do**
  Form $\mathcal{S}_i^\star \leftarrow \{a \in \mathcal{A}_{\mathrm{cal}} : \widehat{Z}_{a,i} = 0,\ \widehat{d}(a, a^\star) \le \tau\}$; set $n_i^\star \leftarrow |\mathcal{S}_i^\star|$.
  if $n_i^\star = 0$ **then**
    Set $C_i(a^\star) \leftarrow \mathbb{R}$ and record a non-finite output.
  else
    Draw $V_i^\star$ uniformly without replacement, independently of scores and cached label values.
    Obtain $s_i^\star$ exact labels, reusing cached labels only for identical pairs and contexts.
    $X_i^\star \leftarrow \sum_{a \in V_i^\star} Z_{a,i}^\circ$; $\overline{B}_i^\star \leftarrow U_\rho(X_i^\star; n_i^\star, s_i^\star)$.
    Compute and order $\mathcal{R}_i^\star = \{R_i^{(a)} : a \in \mathcal{S}_i^\star\}$; set $R^\star_{i,[n_i^\star+1]} \leftarrow +\infty$.
    $k^\star_{\mathrm{cert},i} \leftarrow \overline{B}_i^\star + \lceil (1-\beta)(n_i^\star - \overline{B}_i^\star + 1) \rceil$.
    $\widehat{q}_i \leftarrow R^\star_{i,[k^\star_{\mathrm{cert},i}]}$; $C_i(a^\star) \leftarrow \{y : r_i^\star(y) \le \widehat{q}_i\}$.
    Store $V_i^\star$, its labels, $\overline{B}_i^\star$, $k^\star_{\mathrm{cert},i}$, and the finite-output indicator.
  end if
end for
return all intervals and verification records, retaining non-finite outputs.

C.1 PER-QUERY COMPUTATIONAL COMPLEXITY

After graph preprocessing, let $N = |\mathcal{A}_{\mathrm{cal}}|$, $T = |\mathcal{A}_{\mathrm{train}}|$, $n = n_i^\star$, and $s = s_i^\star$. A sort-based implementation costs $O(T + N + n \log n + \log(n+1) C_H)$ time and $O(n + s)$ working space: the graph-kernel center costs $O(T)$, pool construction costs $O(N)$, and monotone inversion uses $O(\log(n+1))$ hypergeometric CDF calls, each costing C_H. Direct support summation gives $C_H = O(s)$ under unit-cost arithmetic. The experimental partition-based order statistic typically replaces sorting by $O(n)$ selection. Online graph traversal adds $O(|V_r| + |E_r|)$ for visited nodes and edges; up to s new label acquisitions are a separate assay cost.

D COMPARISON METHODS AND EXPERIMENTAL ROADMAP

Within this appendix, $n, B, s, \overline{B}$ abbreviate the current query's pool size, unsafe count, verification size, and certificate; δ_g is the global unsafe fraction. All methods use the same frozen predictor and nonconformity scores within a comparison. A method name specifies both its calibration population and its correction; changing either is not merely renaming a baseline. We use the names below throughout. Reference labels in K562 are assay-based classifications, whereas simulation labels are obtained from the reference graph. None of the reference-informed comparators is a label-free competitor.

Prediction methods.

(i) **Pooled CP** applies the ordinary rank $\lceil (n+1)(1-\alpha) \rceil$ to the pooled calibration scores, without membership correction. In the controlled simulations, the available global pool consists of working-graph-predicted-safe interventions; in K562 it uses all calibration interventions for the target. **Local CP** applies this same uncorrected rule to the query-local predicted-safe pool. These are different baselines.

(ii) **GraphLCP** uses the graph kernel $\exp\{-d^{\mathrm{und}}_{\widehat{G}}(a,b)/h\}$ ($d^{\mathrm{und}}_{\widehat{G}}$ is undirected shortest-path distance and $h > 0$ is bandwidth) and the localized conformal acceptance-level adjustment of Guan (2023), including the test-point mass. It uses no reference membership labels. Its calibration population is the global simulation pool or all K562 calibration interventions. It is not simply an unadjusted weighted empirical quantile and does not certify false-safe contamination.

(iii) **PCS Corrected** uses a selected predicted-safe pool and the additive correction $\alpha' = \alpha - B/(n - B + 1)$, followed by the ordinary rank at α'. It returns an infinite interval when $\alpha' \leq 0$ or the rank is $n + 1$. **Localized PCS Corrected** first restricts that selected pool to the query neighborhood and applies the same correction using its own n and B. Both receive reference membership for all selected pairs in these benchmarks. In simulations the selector is held fixed to the working-graph-safe population to isolate the correction; in K562 the partial-structure selector is estimated from differential-expression/intersection information on training and calibration interventions. The K562 selector is therefore distinct from the ENCODE selector used by CGLCP. Historical plot labels "PCL-900" or "PCS-SC" denote these PCS implementations, not additional methods.

(iv) **Global-rate plug-in** uses the local scores but substitutes $\lceil n\delta_g \rceil$ for B in the sharp rank at level α. It diagnoses the consequences of replacing local information by a global fraction; there is no local upper-bound certificate. In simulations δ_g is known; in K562 it uses reference labels over all global working-safe calibration pairs.

(v) **CGLCP** uses the hard-selected local pool, a uniform verification sample of size s, the hypergeometric upper bound $\overline{B}$, and rank $k_\rho(\overline{B}) = \overline{B} + \lceil (1-\beta)(n - \overline{B} + 1) \rceil$, where $\beta = (\alpha - \rho)/(1 - \rho)$. Its reference-label requirement is only the sampled local members. The earlier **additive certified** diagnostic instead substitutes the same $\overline{B}$ into $\alpha' = \beta - \overline{B}/(n - \overline{B} + 1)$; it is not the reference-informed PCS comparator.

Verification versus prediction oracles. The *verification oracle* returns a membership label for a sampled pair; it is an input to CGLCP, not a competing predictor. Three infeasible prediction benchmarks expose different information costs. **Oracle local (ρ)** knows B and uses $k_\rho(B)$ with the same β as CGLCP, isolating the cost of estimating B. **Oracle local (α)** knows B but uses $B + \lceil (1-\alpha)(n - B + 1) \rceil$, avoiding the certificate-failure budget. **Oracle-safe CP** knows every membership, removes all unsafe scores, and uses ordinary CP on the remaining $n - B$ scores. Full exact verification makes CGLCP equal Oracle local (ρ), not necessarily either stronger oracle. "Oracle certificate budget", "same-ρ oracle", and "rho-budget oracle" refer to Oracle local (ρ). In K562 these benchmarks know the operational reference classifications, not unobserved causal truth.

Within-radius weighted variants. The two additional methods in Appendix E.6 keep CGLCP's hard-selected pool and uniform verification. **Weighted plug-in** replaces the certified order statistic with an exponential-distance-weighted quantile at the same rank fraction, including test-point mass; it is an empirical extension without the original unweighted-rank guarantee. **Weighted envelope** takes the larger of that threshold and the CGLCP threshold. Its interval contains CGLCP's and therefore preserves its coverage whenever the original assumptions hold. The exact weighting formula and the uniform-weight limit are given with the experiment.

Which comparisons answer which question? The main simulation, Simulation B, and main K562 comparison use seven methods: Pooled CP, Global-rate plug-in, GraphLCP, PCS Corrected, Localized PCS Corrected, CGLCP, and Oracle local (ρ). Simulation B includes a paired Pooled CP extension using the original score draws. The sharpness and multi-graph diagnostics instead compare Local CP, Global-rate plug-in, CGLCP, and the three prediction oracles: their purpose is to separate contamination correction, graph variation, and information costs. Verification-budget plots need only CGLCP and its same-budget oracle; exact hypergeometric enumeration does not require unrelated prediction methods. Full sentinel curves retain the six graph/correction methods, with the Pooled CP sentinel operating point supplied in the main K562 comparison. The small reference-unsafe endpoint uses the frozen Local CP and oracle diagnostics. These intentionally different comparator sets are not claims that every experiment evaluates every method.

The appendix follows the main paper's evidence chain: construction and score laws establish the controlled setting; verification, ρ, and radius studies explain the validity–availability trade-off; graph-seed and score-mismatch studies separate structural robustness from failures of exchangeability; oracle-noise and within-radius weighting studies examine the two proposed extensions; sentinel experiments assess an auxiliary use of the fixed intervals. The K562 supplement mirrors the local-risk, verification, and sentinel questions, then checks analysis-design choices. Graph-risk triage is exploratory prioritization for independent follow-up, not demonstrated causal-edge repair.

D.1 Computational environment

The hardware and software used for the experiments are as follows.

- CPU: Intel(R) Xeon(R) Platinum 8352V CPU @ 2.10GHz
- GPU: NVIDIA GeForce RTX 4090
- System: Ubuntu 22.04
- Python: 3.10.18
- NumPy: 2.2.6
- SciPy: 1.15.3

E Simulation supplement

E.1 Data generating process and experiment setup

We construct directed modular graphs in which the working graph predicts 240 calibration interventions to be safe, including a query-local pool of 60. Each setting contains 24 false-safe interventions globally, so the global false-safe fraction remains 0.10, while the number in the local pool varies over $1, 3, 6, 12, 15$, and 24. We generate 800 score and audit replicates per setting. Reference-safe calibration scores and the safe test score follow $|N(0, 1)|$; unsafe calibration scores follow $|N(0, 0.1^2)|$. The latter deliberately places unsafe scores at low ranks, where contamination can cause undercoverage. CGLCP audits 30 local members. Pooled CP and GraphLCP use all calibration interventions, whereas the PCS comparisons use fixed working-graph-selected pools with reference membership available for their corrections.

Unless stated otherwise, $\alpha = 0.10$ and $\rho = 0.02$. Pooled CP ignores graph information; GraphLCP localizes by working-graph distance without verifying safety labels. PCS Corrected and its radius-restricted variant use reference labels for their corrections, while the global-rate plug-in and the local oracle diagnose the cost of global rather than local information and estimated rather than known contamination. We report total coverage, finite-output rate, coverage conditional on finite output, finite-only width, and reference-label use. Infinite intervals count as covered but not finite.

Simulation A uses a modular graph to isolate the global-to-local contamination mechanism. Simulation B uses eight directed regulatory modules, query-adjacent regulation, paths of different depths converging on a target, deleted ancestor-to-hub edges that create false-safe members, and added working-graph edges that exclude true-safe members. Prediction-result panels distinguish total coverage from finite availability; width is computed only among finite outputs. Repeated-score uncertainty uses Wilson intervals for Bernoulli endpoints and empirical intervals for width. The multi-graph study averages within graph seed before bootstrapping seeds. As illustrated in Figure 4,

we allocate the same number of omitted ancestor-to-hub edges across modules, so the distant, spread, and concentrated settings change local risk without changing the global omission count.

Score-generating laws and pool designs. These are graph-indexed score simulations, not simulations from a fitted gene-expression structural equation model. For a response interpretation one may set the fixed predictor to zero, the scale to one, and $Y = \varepsilon R$ with an independent random sign $\varepsilon \in \{-1, 1\}$; all reported coverage calculations depend only on $R = |Y|$. Graph reachability supplies membership labels, while the score law is specified separately. In the original sharpness/budget suite, safe calibration and test scores are independent $|N(0,1)|$; unsafe scores are $|N(0, \sigma_u^2)|$ with $\sigma_u = 1, 0.5, 0.1$ for benign, moderate, and severe contamination, respectively, and are identically zero for the adversarial condition. Thus no intervention-to-expression equation beyond the stated score experiment is assumed; the graph controls which scores may be contaminated, not a fitted causal response magnitude.

The 800-replicate main comparison uses $n = 60$, 240 global predicted-safe interventions, 24 global false-safe members, and $s = 30$. The older mechanism scans use their own pool sizes and budgets (including $n = 100$), not this main-panel configuration. Their nearest-K pools use deterministic distance tie breaking. A fixed-capacity neighborhood is not generally the full ball at a radius. The original layered radius control, reused for the new weighting experiment, instead uses a fixed graph with distance shells $d = 2, 3, 4, 5, 6$, shell sizes $(25, 25, 50, 100, 100)$, and unsafe shell counts $(10, 5, 5, 5, 5)$; cumulative (n, B) are $(25, 10), (50, 15), (100, 20), (200, 25), (300, 30)$. All of these are fixed discrete populations before verification.

For Simulation B, let d_a be undirected working-graph distance to the query, $m_a \in \{0, \dots, 7\}$ the module index, and b_a the sum of the node identifier's UTF-8 bytes modulo four. With independent $L_a \sim N(-0.3^2/2, 0.3^2)$ and $T_a \sim t_5/\sqrt{5/3}$, the safe calibration law is

$$R_a = \exp\{L_a - 0.040 I_D d_a - 0.020 I_M \min(m_a, 7) + I_B c_{b_a}\}|T_a|,$$
$$(c_0, c_1, c_2, c_3) = (-0.06, -0.02, 0.02, 0.06).$$

The safe query follows $R^\star = \exp\{L_\star + 0.10 I_B\}|T_\star|$, with independent variables from the same base laws. Distance, Module, and Batch set respectively only I_D, I_M, or I_B to one; Combined sets all three to one. Exchangeable pair-heavy scores set all three to zero. Exchangeable Gaussian scores additionally set $L = 0$ and replace T by standard Gaussian noise. Unsafe scores always follow $|N(0, 0.1^2)|$. In the global-pool implementation, infinite graph distances are replaced by one plus the maximum finite distance before computing the scale shift. These mismatch regimes intentionally violate safe-score exchangeability; heavy tails alone do not. Simulation B uses 800 replicates per cell, $s/n = 0.40$, requested local unsafe counts $0, 3, 6, 12, 18, 24$, and zero or 24 added working edges. Added edges connect reference-safe interventions to the target, while remote safe buffer nodes preserve the global predicted-safe size; actual local membership is recomputed from the resulting graphs.

Verification-noise design. We retain the modular graph with $n = 60$ local members and $B \in \{6, 24\}$ true unsafe members. Safe scores and the safe query follow independent $|N(0,1)|$; unsafe scores follow $|N(0, 0.1^2)|$. For each of 10,000 replicates per setting (master seed 20260922), scores, uniform verification permutations, and label-flip uniforms are shared across noise levels. We use $s = 30$ or 60, $\alpha = 0.10$, and $\rho = 0.02$. With unsafe as positive, an audited safe label flips to unsafe with probability p_{FP} and an unsafe label flips to safe with probability p_{FN}, independently of the scores. FP-only, FN-only, and equal-FP/FN paths use probabilities $0, 0.05, 0.10, 0.20, 0.40, 0.60$. The unmodified CGLCP calculation receives the noisy count $\widetilde{X}$; this deliberately measures the consequence of violating exact verification, rather than assuming noise correction.

Within-radius weighting design. We use 10,000 paired replicates (seed 20261022), $\alpha = 0.10$, and $\rho = 0.02$. In the exchangeable condition, safe scores and the test score follow $|N(0,1)|$. In a separate prespecified distance-mismatch stress test, safe calibration scores are multiplied by $\exp(-0.12 d_a)$ while the safe test score is unchanged. Unsafe scores follow $|N(0, 0.1^2)|$ in both conditions. This stronger distance shift is specific to this new stress test and is not the 0.040 shift in Simulation B. All methods share the scores and verification sample at every bandwidth. In the weighting diagnostics, R_a and Z_a denote the selected calibration score and true unsafe indicator for member a.

Graph-seed and baseline alignment. MG-HET (local-contamination heterogeneity) and MG-EDGE (deleted/added-edge factorial) use 12 graph seeds each; MG-TAU (radius sensitivity) uses 16, with

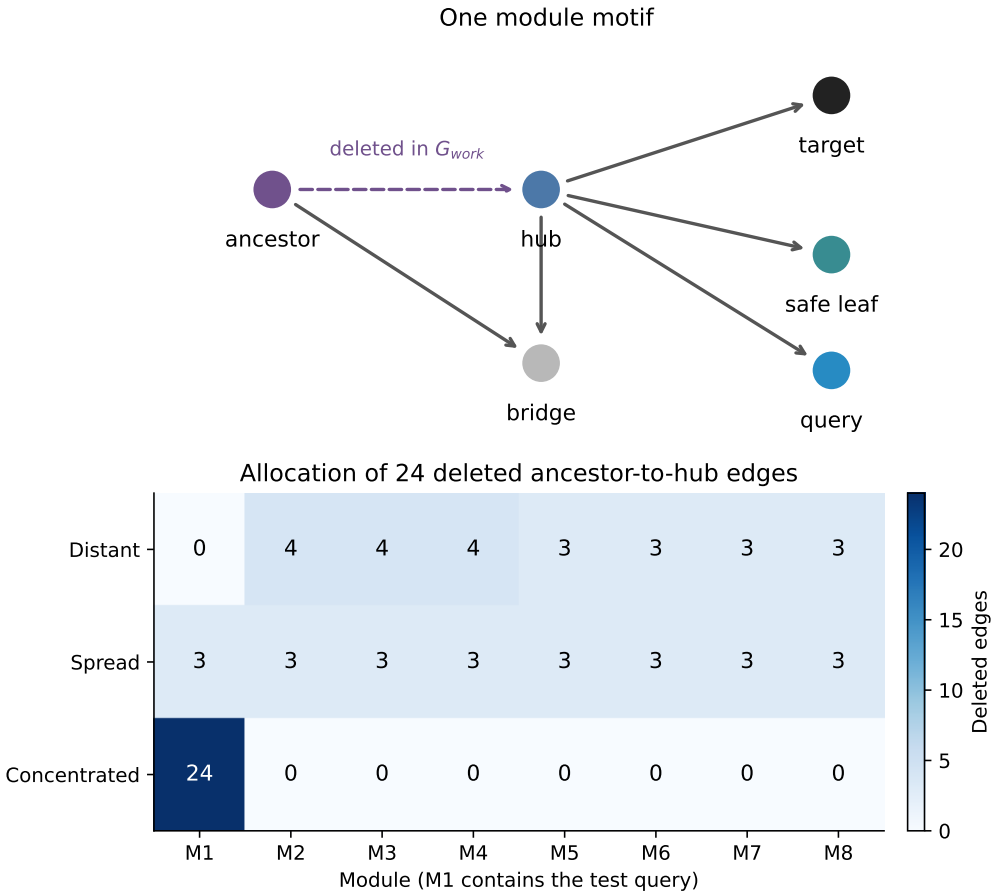


Figure 4: Simulation B construction schematic: reference/working graph motif and allocations of 24 deleted ancestor-to-hub edges. Counts are fixed graph-design quantities, not statistical estimates. The enlarged edge-allocation table preserves the distant, spread, and concentrated settings.

500 inner replicates per graph. Graph seeds randomize edge identities, module roles, path depths, and distance ties. Simulation B includes seven methods; Pooled CP was added by replaying the original random-number stream, after checking exact agreement of all 57,600 original CGLCP outcomes and certificates.

Radius/budget panel design. For the sensitivity analysis in Figure 6, we use the pair-heavy extension with 24 added working edges and 1,000 replicates per setting. Radius-1 through radius-6 pools have $(n, B) = (53, 3), (75, 3), (84, 12), (99, 12), (230, 14), (240, 24)$, distinct from the layered design in Appendix E.6.

E.2 Finiteness, verification fraction, ρ, and radius

We vary verification and contamination to assess finite-output behavior, with results in Figure 5. More verification reduces certificate excess and narrows the gap to Oracle local (ρ), whereas increasing contamination can force the sentinel. Coverage near one can therefore reflect abstention rather than useful finite prediction; the paired oracle comparison isolates the additional cost of estimating the unsafe count.

Because ρ affects both certificate confidence and the conformal error allocation, width and rank need not vary monotonically, as shown in Figure 6. In the pair-heavy radius study, CGLCP covers 0.920–0.946 with fully finite output, while median width ranges from 3.6608 to 4.1649. Radius changes pool size and unsafe composition together, explaining why a larger neighborhood is not uniformly

more efficient. These curves support prespecified sensitivity analysis rather than outcome-based radius selection.

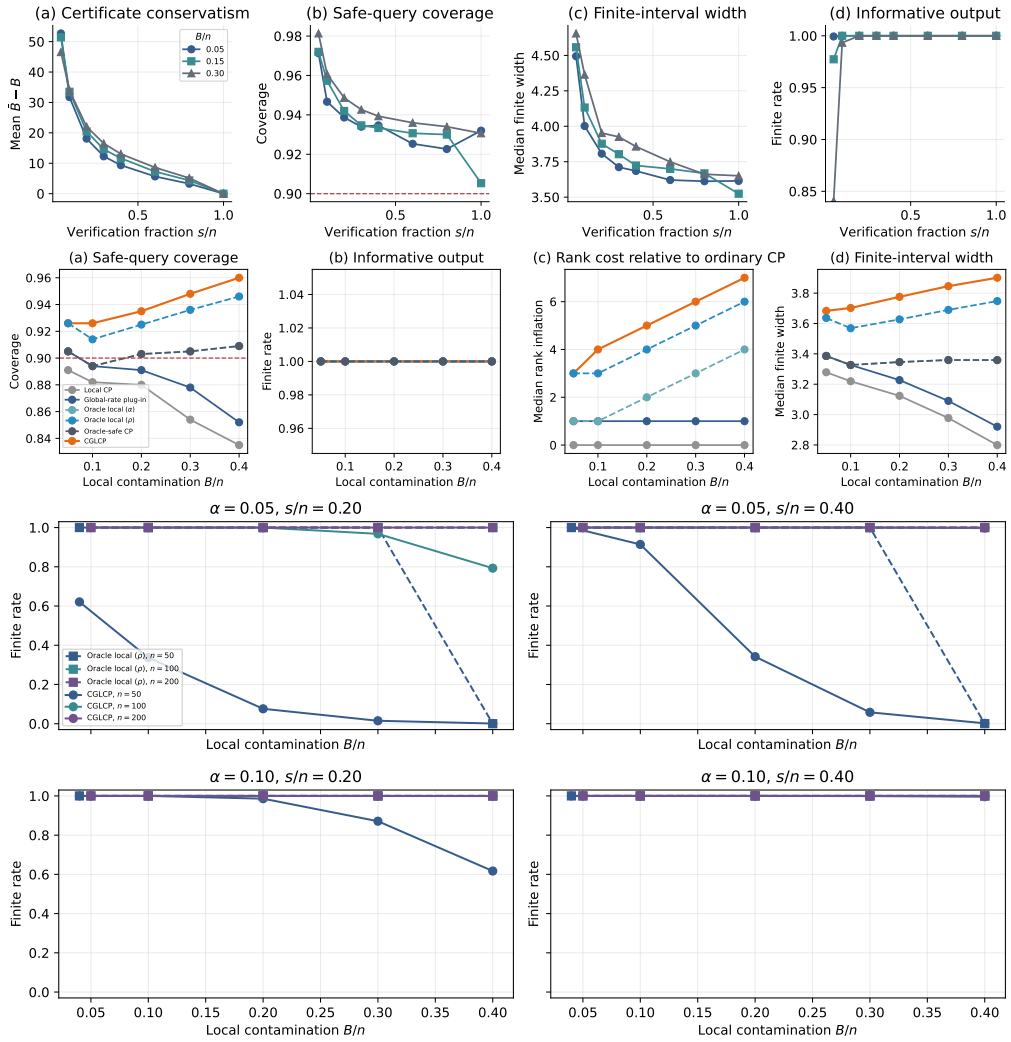


Figure 5: Mechanism diagnostics: verification budget (top), contamination–efficiency trade-off (middle), and finite-output boundary (bottom). Curves are simulation point estimates without confidence bands. Width summaries condition on finite output. The bottom panel compares CGLCP with Oracle local (ρ) across pool sizes and error budgets; their separation isolates certificate-estimation cost.

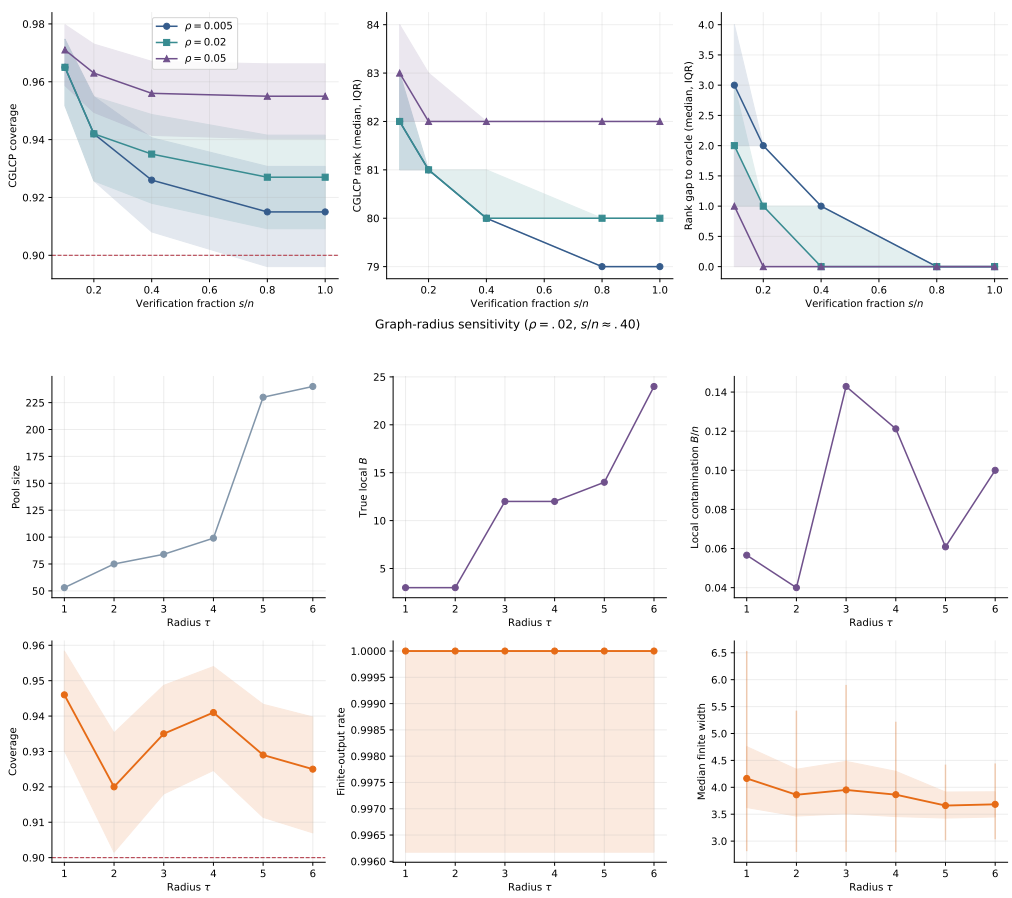


Figure 6: Sensitivity to certificate budget ρ (top) and graph radius τ (bottom). Coverage and finite-rate bands are 95% Wilson confidence intervals. Rank and paired rank-gap shading in the top panel shows interquartile ranges. Finite-width shading in the bottom panel shows the interquartile range and vertical lines the empirical 2.5th–97.5th percentiles. These distributions are not confidence intervals for mean width. Pool size and true contamination are fixed design quantities.

E.3 Graph-seed robustness and added/deleted edges

The graph-seed experiments test whether local certification depends on one network realization. At requested local contamination $24/60$ under exchangeable pair-heavy scores, CGLCP has mean graph-level coverage 0.9595 (95% graph-bootstrap CI $[0.9547, 0.9643]$), versus 0.8647 for the global-rate plug-in and 0.8388 for Local CP, with finite rate 0.9998, as shown in Figure 7. Its mean within-graph median width is 4.6753, compared with 3.6516 for Oracle local (ρ): protection persists across graphs with a measurable efficiency cost, rather than through universal abstention.

The deleted/added-edge factorial distinguishes false-safe contamination caused by deleted edges from loss of safe candidates caused by added edges, as shown in Figure 7. Radius trajectories vary across graphs because both pool size and composition change, while the aggregate results retain the same coverage–availability trade-off, as shown in Figure 8. These checks support structural robustness within the simulated graph family, not invariance to arbitrary graph or score shifts.

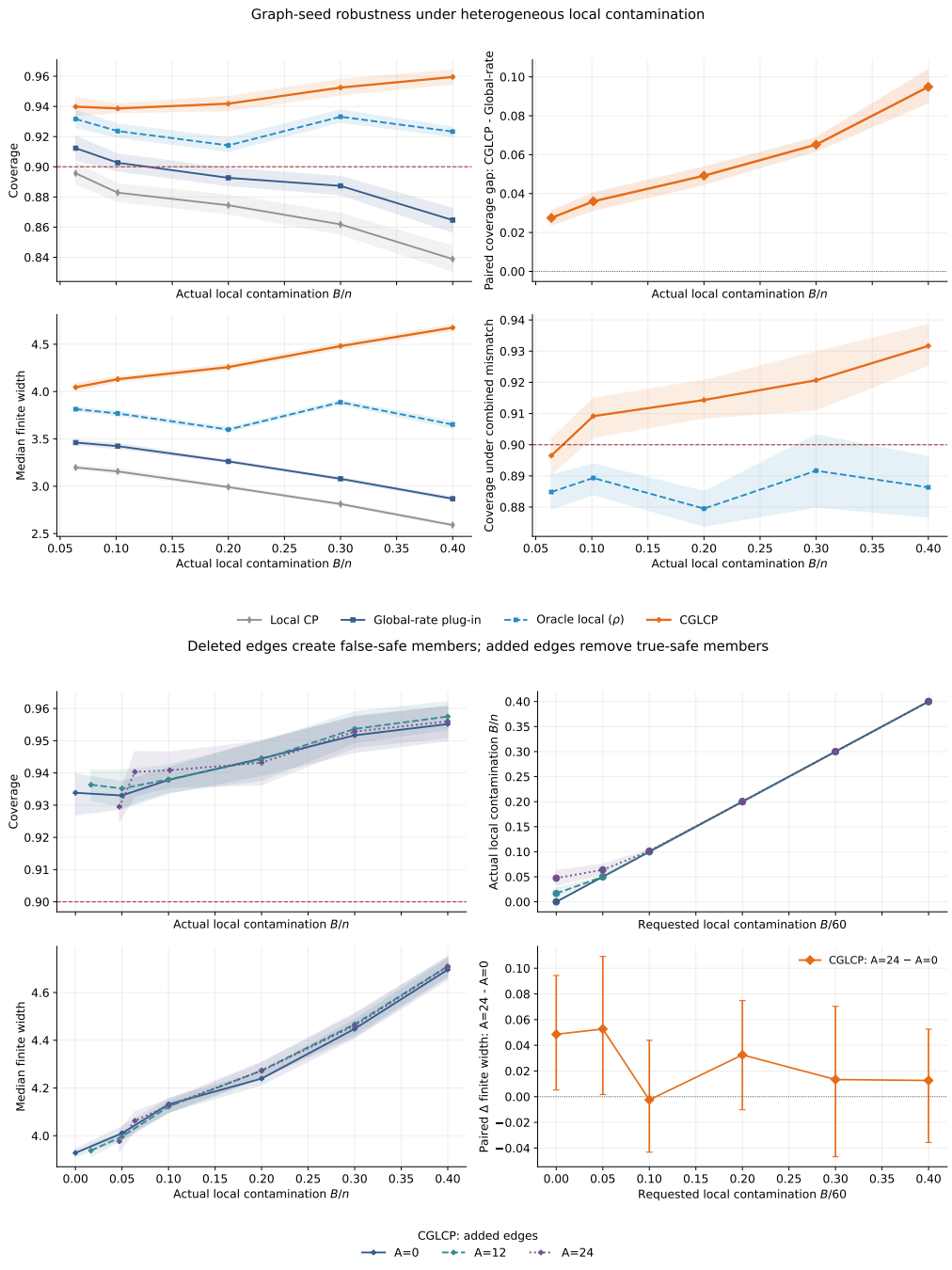


Figure 7: Multi-graph local heterogeneity (top) and deleted-edge × added-edge factorial (bottom). Shaded bands/error bars are pointwise 95% percentile bootstrap confidence intervals from 2,000 resamples of graph seeds after averaging within each seed; 12 graph seeds and 500 inner replicates per seed are used. Finite-width summaries average within-seed medians over seeds with finite output. Paired contrasts preserve the graph-seed pairing.

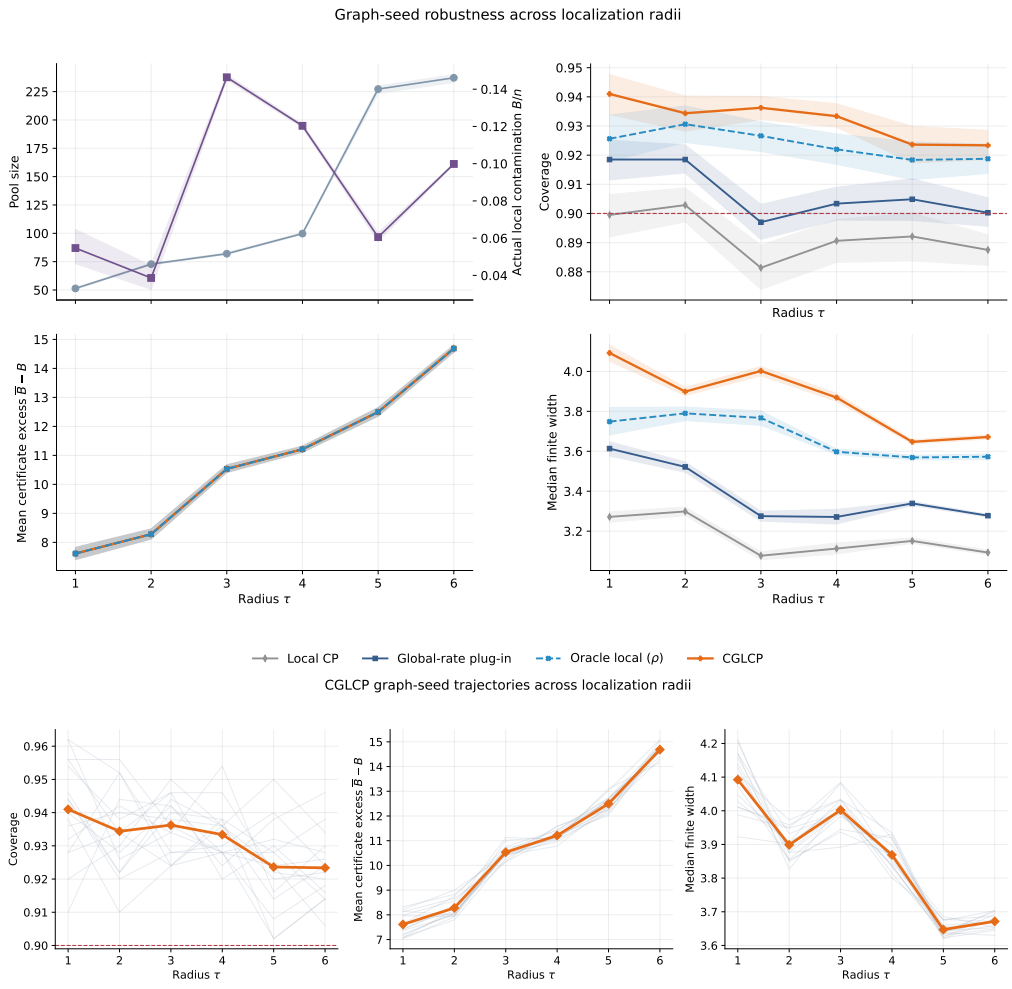


Figure 8: Multi-graph radius sensitivity. Top: 95% graph-seed percentile bootstrap confidence bands (2,000 resamples; 16 graphs, 500 inner replicates each). Bottom: individual graph trajectories, not confidence intervals. Radius changes sample size and contamination simultaneously; width conditions on finite output.

E.4 SAFE-SCORE HETEROGENEITY: ROBUSTNESS AND BOUNDARY

At local contamination 0.40 with 24 added working edges, CGLCP covers 0.9625 under exchangeable Gaussian scores and 0.9575 under exchangeable pair-heavy scores, versus 0.8963/0.8850 for Pooled CP and approximately 0.870/0.850 for the global-rate plug-in, as shown in Figures 9 and 10. CGLCP and Pooled CP remain fully finite, whereas both PCS corrections have zero finite availability. Thus heavy tails alone do not remove the benefit of local correction when safe scores remain exchangeable.

Under combined distance, module, and batch mismatch, coverage is 0.934 for CGLCP, 0.893 for Oracle local (ρ), and 0.7888 for Pooled CP, as shown in Figure 10. The higher CGLCP value is setting-specific conservatism, not an extension of its theorem: even exact membership information cannot repair score nonexchangeability. We further assess coverage across local risk and graph radius in Figures 9 and 11.

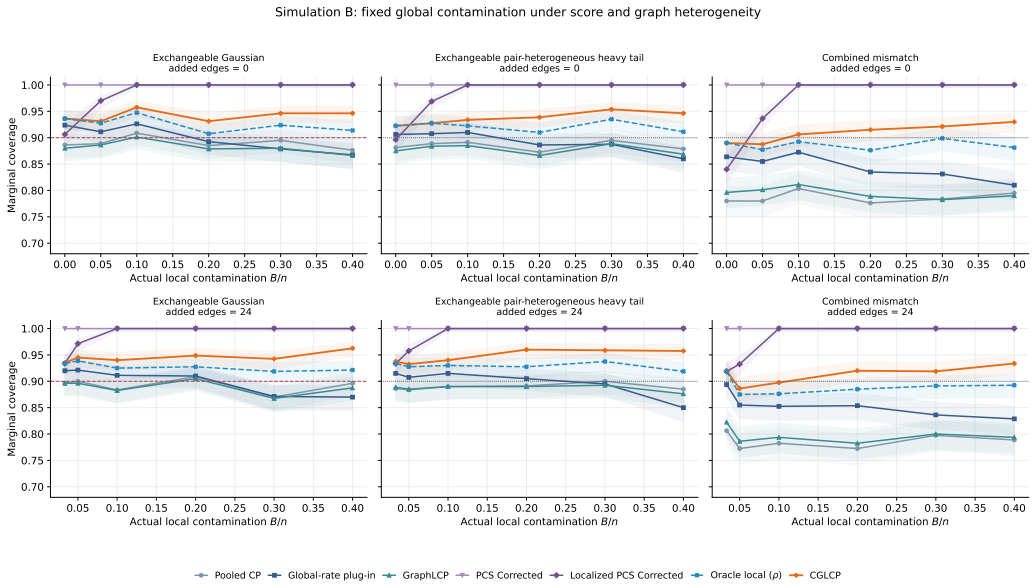


Figure 9: Simulation B local-risk scan: seven methods with 800 shared replicates per cell. Rows vary added working edges and columns vary score regimes. Shaded bands are 95% Wilson coverage confidence intervals. Pooled CP uses the same replayed global score draws as the six previously evaluated methods.

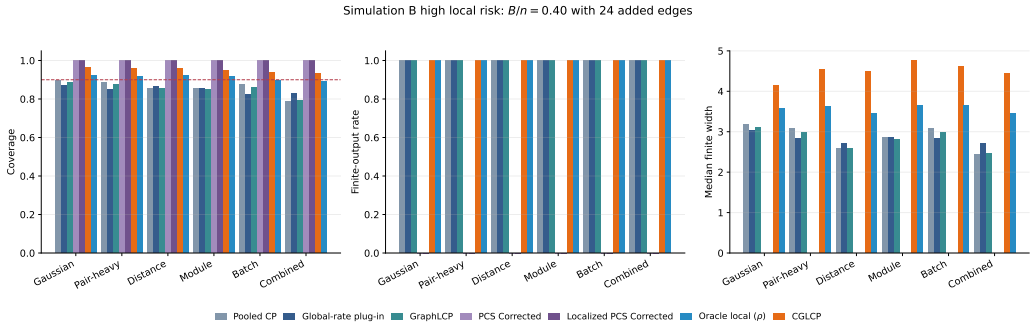


Figure 10: High-risk Simulation B operating point: coverage, finite rate, and median finite width across all six score regimes. These panels show point estimates without error bars. PCS coverage one is accompanied by zero finite availability. The combined-mismatch setting is outside the exchangeability theorem.

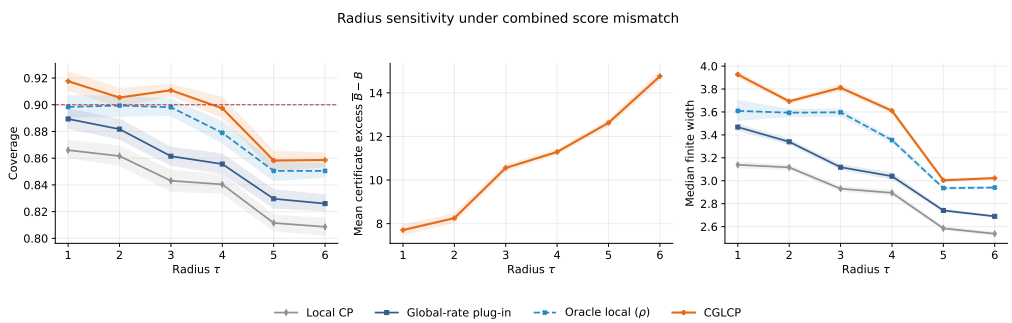


Figure 11: Radius sensitivity under combined score mismatch. Bands are 95% graph-seed percentile bootstrap confidence intervals from 2,000 resamples (16 graphs, 500 inner replicates each). This is an assumption-boundary experiment; membership certification alone does not restore score exchangeability.

E.5 Propagation of verification-oracle errors

This experiment isolates verification noise while keeping scores and sampling designs paired; its settings are given in Appendix E.1. We assess the resulting changes in certificate validity, coverage, width, and finite availability in Figure 12 and Table 2.

Propagation. The noisy audited count has expectation $s\{(1 - p_{\mathrm{FN}})B/n + p_{\mathrm{FP}}(1 - B/n)\}$. It is not generally hypergeometric with the true unsafe count B. For FP-only noise, $\widetilde{X} \geq X$ replicate by replicate: monotonicity of hypergeometric inversion implies $\overline{B}(\widetilde{X}) \geq \overline{B}(X)$, so the rank and interval cannot shrink. For FN-only noise the inequalities reverse. For mixed noise, net count bias depends on both error probabilities and the safe/unsafe composition; equal FP and FN rates do not imply cancellation. If $p_{\mathrm{fail}} = \Pr\{\overline{B}(\widetilde{X}) < B\}$, the end-to-end lower bound becomes $(1 - p_{\mathrm{fail}})(1 - \beta)$, provided the conditional safe-score assumption still holds. Thus a failed ρ-level certificate need not immediately produce empirical undercoverage, but its nominal justification has been lost.

Results. At $B/n = 0.40$ and $s/n = 0.50$, exact verification gives coverage 0.9467, certificate failure 0.0075, and fully finite output. Raising FP-only noise to 0.40 raises coverage to 0.9797 but reduces finite output to 0.7050; at 0.60, finite output falls to 0.1487. This illustrates conservative protection with a substantial availability cost. At FN probability 0.40, certificate failure rises to 0.5711 while coverage remains 0.9249: coverage alone would conceal the broken certificate. With full verification and FN probability 0.60, coverage falls to 0.8928, whereas the noiseless full-verification comparator covers 0.9181. More verification cannot repair a systematically incorrect label oracle. These results support CGLCP under accurate or one-sided conservative labels while identifying false negatives as a distinct threat to its guarantee. We report full noise paths in Figure 12 and selected operating points, including lower contamination and finite availability, in Table 2; the full grid and finite-width summaries are retained in the machine-readable experiment outputs.

A noise-aware deployment would need independently justified error bounds and a certificate that incorporates them. As a simple sufficient construction, if an upper bound U_{FN} on the number of missed unsafe labels in the audited sample fails with probability at most η, then applying the hypergeometric inversion to $\min(s, \widetilde{X} + U_{\mathrm{FN}})$ at failure budget ρ_0 gives total certificate failure at most $\rho_0 + \eta$ by a union bound. The conformal budget must then reserve $\rho_0 + \eta < \alpha$, and the conditional safe-score assumption must hold given the enlarged verification information. No such independent error bound is available for the current K562 assay labels, so this experiment does not upgrade them to exact causal labels.

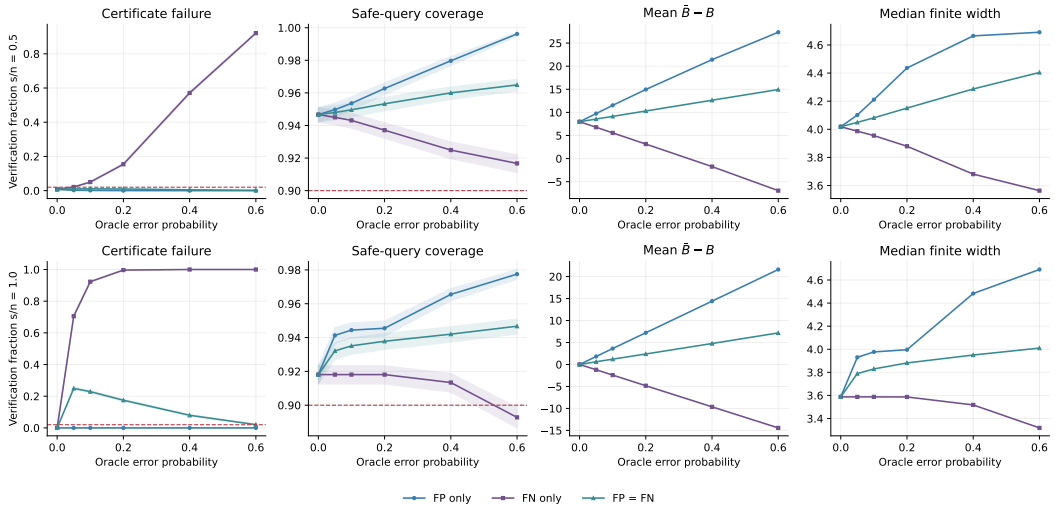


Figure 12: Verification-label noise at $B/n = 0.40$. Rows use verification fractions $s/n = 0.50$ and 1.00; each setting has 10,000 paired replicates. Shaded probability bands are pointwise 95% Wilson confidence intervals. Certificate bias and finite-width medians are point estimates without confidence bands. FP denotes a safe pair incorrectly labeled unsafe; FN denotes an unsafe pair incorrectly labeled safe.

Table 2: Selected verification-noise operating points, each based on 10,000 replicates. Entries are empirical point estimates; probability uncertainty for the high-contamination setting is reported in Figure 12. Certificate failure is $\Pr(\overline{B} < B)$, Cov. includes infinite intervals, and Finite counts only finite outputs.

B/n	s/n	Verification noise	Cert. failure	Cov.	Finite
0.10	0.50	Exact labels	0.0123	0.9391	1.0000
0.10	0.50	FP = 0.40	0.0000	0.9711	0.9943
0.10	0.50	FN = 0.60	0.2596	0.9283	1.0000
0.10	1.00	Exact labels	0.0000	0.9268	1.0000
0.10	1.00	FP = 0.40	0.0000	0.9610	1.0000
0.10	1.00	FN = 0.60	0.9953	0.9268	1.0000
0.40	0.50	Exact labels	0.0075	0.9467	1.0000
0.40	0.50	FP = 0.40	0.0000	0.9797	0.7050
0.40	0.50	FN = 0.60	0.9210	0.9167	1.0000
0.40	1.00	Exact labels	0.0000	0.9181	1.0000
0.40	1.00	FP = 0.40	0.0000	0.9655	0.9998
0.40	1.00	FN = 0.60	1.0000	0.8928	1.0000

E.6 Distance weighting within a hard-selected population

Design and method. We fix the complete radius-6 ball of the layered graph: $n = 300$, $B = 30$, and uniform verification size $s = 120$. Membership and the verification design do not depend on score values or weights. Inside this population set $w_a(h) = \exp(-d_a/h)$, where d_a is distance to the query and h is bandwidth, with $h \in \{0.5, 1, 2, 4, 8, \infty\}$; set the test-point mass to one because its self-distance is zero. Let $k = k_\rho(\overline{B})$ and define the diagnostic weighted threshold

$$\widehat{F}_h(r) = \frac{\sum_a w_a(h)\mathbf{1}\{R_a \le r\} + \mathbf{1}\{r = +\infty\}}{1 + \sum_a w_a(h)},$$

$$q_w(h) = \inf\left\{ r \in \{R_{[1]}, \ldots, R_{[n]}, +\infty\} : \widehat{F}_h(r) \ge \frac{k}{n+1} \right\}.$$

The **Weighted plug-in** uses q_w; it exactly reduces to CGLCP when $h = \infty$. The **Weighted envelope** uses $\max(q_{\mathrm{CGLCP}}, q_w)$ with the same center and scale, where q_{CGLCP} is the original CGLCP threshold. The envelope contains the original CGLCP interval pointwise, so it inherits that interval's coverage guarantee whenever the original assumptions hold. It cannot improve width or finite availability over CGLCP. The direct weighted plug-in has no coverage guarantee from our unweighted contaminated-rank theorem: a certificate for the number of unsafe members does not itself certify their weighted mass or establish weighted safe-score invariance. This diagnostic is also distinct from the adjusted GraphLCP algorithm.

Results. In Figure 13, under exchangeability, CGLCP attains coverage 0.9222, finite rate 1.0000, and median finite width 3.5713. At $h = 2$, the weighted plug-in attains coverage 0.9467 with median finite width 3.9109 and fully finite output; it provides no efficiency improvement in this design. At $h \leq 1$, the test-point mass dominates the strongly decayed calibration mass and both weighted variants return only infinite intervals. This makes the availability cost visible rather than interpreting coverage one as a gain. Under distance mismatch, weighting at $h = 2$ improves empirical coverage from 0.6939 to 0.7860, still below 0.90; even the envelope cannot restore a guarantee whose exchangeability assumption has failed. Hard selection therefore remains compatible with internal weighting, but weighting is neither a replacement for contamination certification nor an automatic remedy for score mismatch.

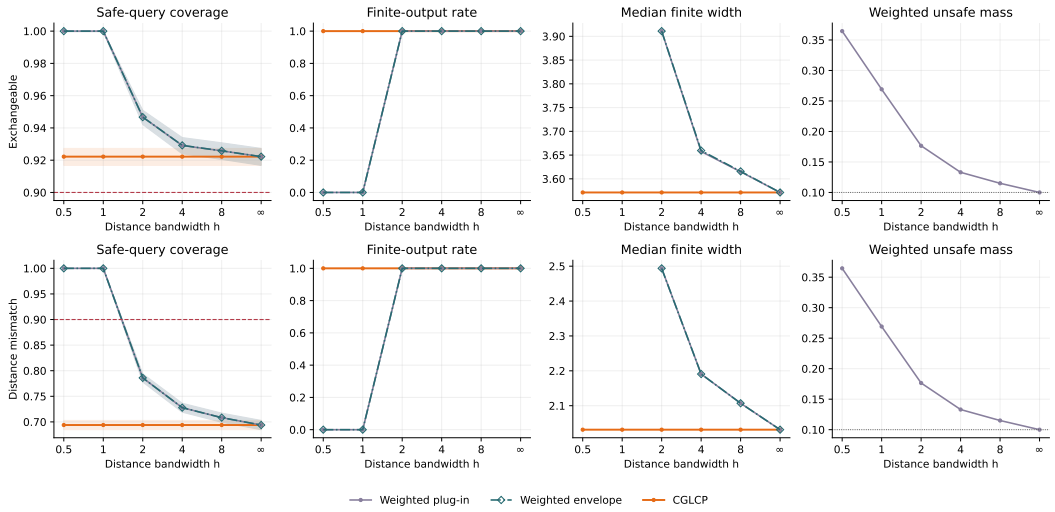


Figure 13: Distance weighting after fixed hard-radius selection. Rows separate exchangeable scores and the distance-mismatch stress test; 10,000 common replicates per condition. Coverage and finite-rate bands are 95% Wilson confidence intervals; widths are finite-only medians without confidence bands. Missing widths indicate zero finite output. Weighted unsafe mass $\sum_a w_a Z_a / \sum_a w_a$ is a deterministic graph diagnostic, with the unweighted fraction $B/n = 0.10$ shown for reference. At infinite bandwidth all three methods coincide.

E.7 Simulation sentinel extension

We evaluate the same frozen intervals as negative-control alarms in Figure 14: zero shift measures false alarms, and increasing outward shifts measure sensitivity. At local contamination 0.30, verification fraction 0.20, and $\kappa = 2$, CGLCP has false-alarm rate 0.0530, recall 0.9105, precision 0.6562 at prevalence 0.10, and fully finite output. The global-rate plug-in and GraphLCP have recall 0.9995 and 1.0000 but false-alarm rates 0.1275 and 0.1165; both PCS corrections abstain. In this setting, CGLCP combines low false-alarm rates with finite availability, at the cost of lower recall than GraphLCP and the global-rate plug-in.

We decompose unconditional detection into finite-output availability and performance conditional on issuing an interval, and assess width and verification sensitivity, with results in Figure 15. This separates low false alarms caused by abstention from those achieved with finite predictions, and

connects the cost of certification to detection performance. Precision is undefined when no alarms occur and is not replaced by zero.

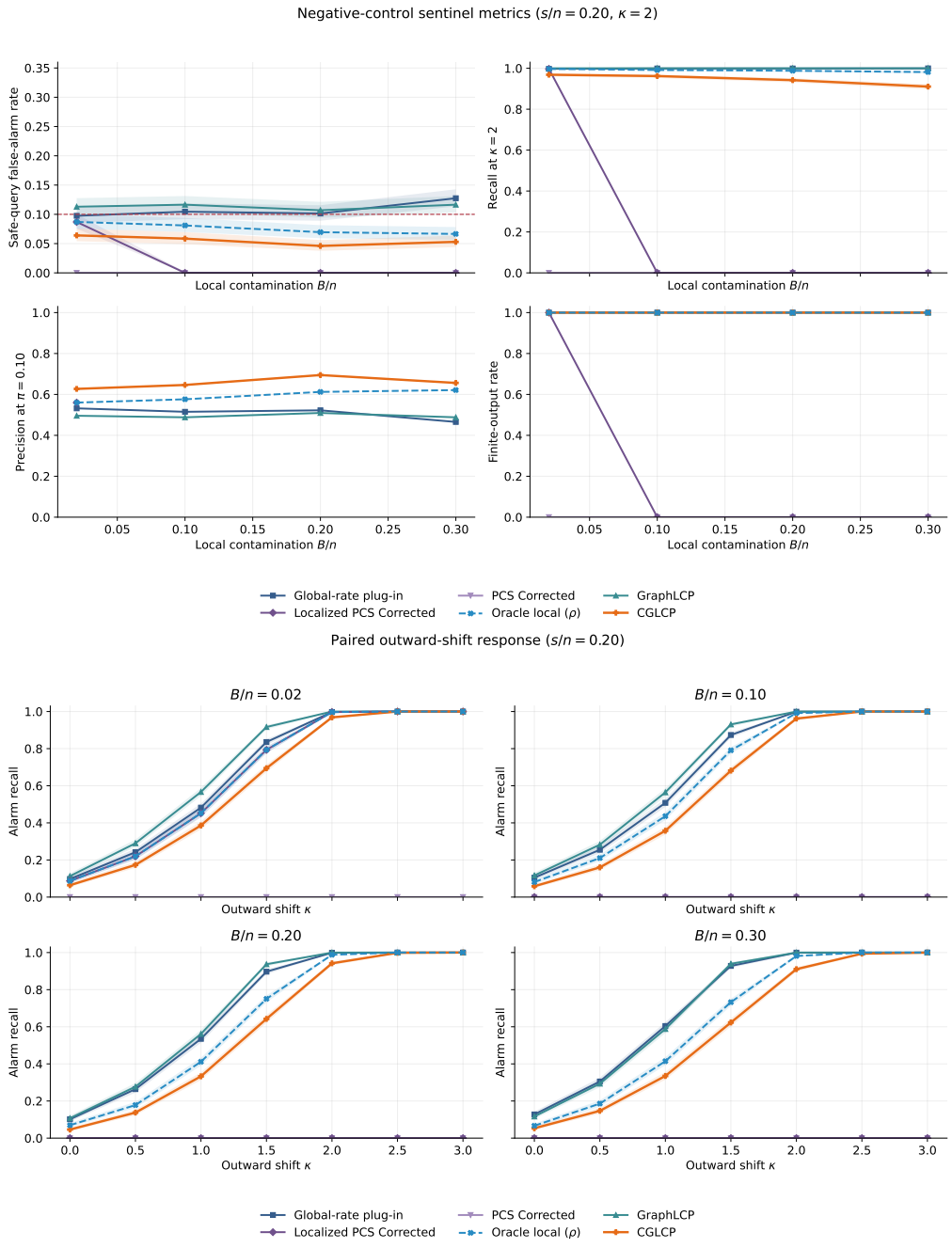


Figure 14: Simulation sentinel operating points and shift curves. Shaded bands for false-alarm probability, recall, and finite rate are 95% Wilson confidence intervals over independent simulation replicates; precision at prevalence 0.10 is a plug-in point estimate without a confidence band. Intervals are frozen before the outward shifts; infinite output is abstention.

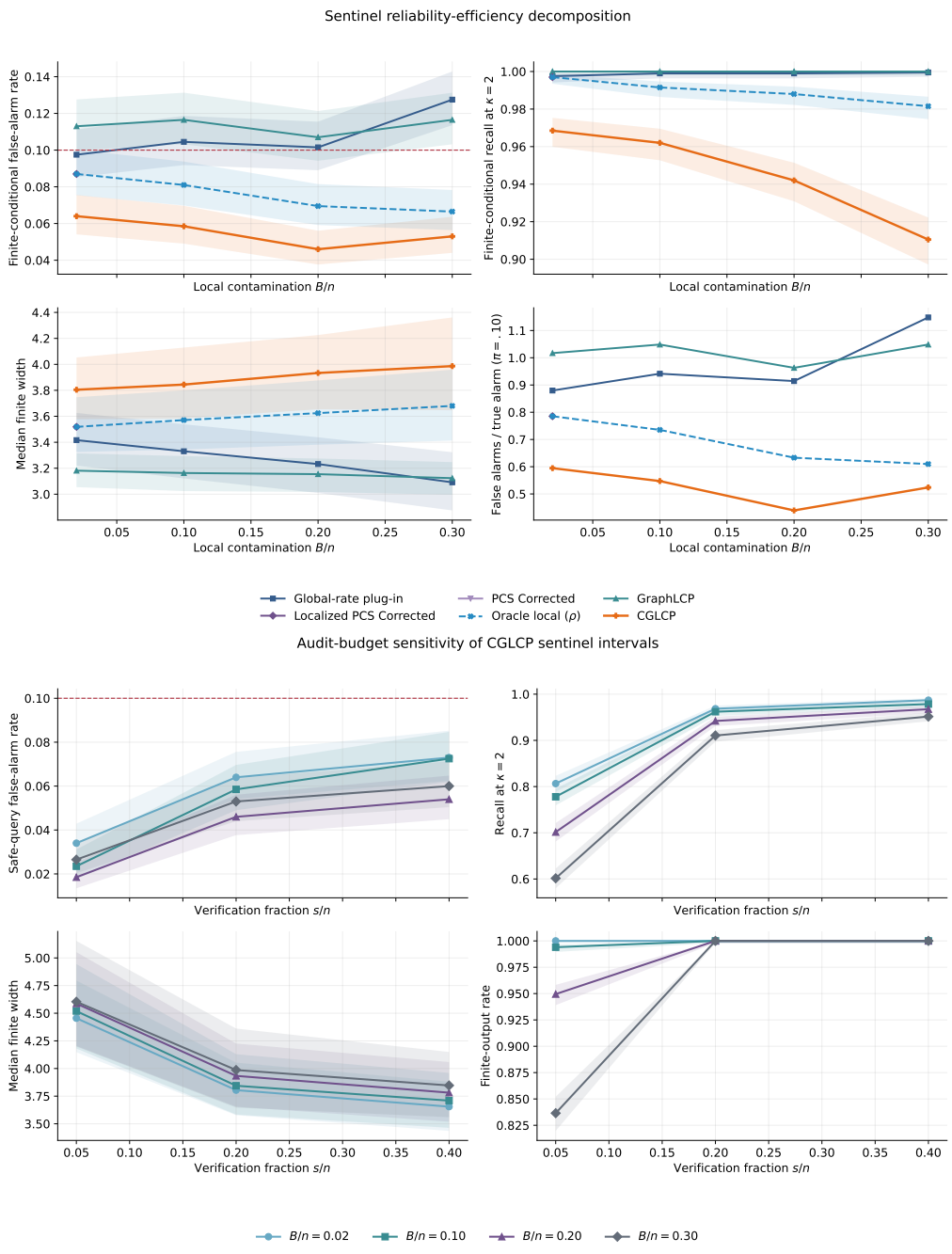


Figure 15: Simulation sentinel mechanism and verification sensitivity. Probability bands are 95% Wilson confidence intervals, using finite replicates as the denominator for finite-conditional metrics. Finite-width bands show the empirical interquartile range, not 95% confidence intervals. False/true-alarm ratios are plug-in point estimates.

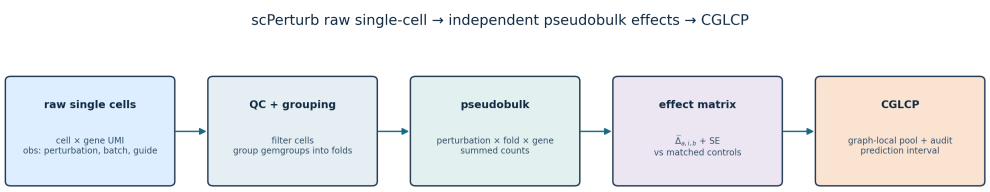


Figure 16: K562 processing schematic. Six nonoverlapping batch folds implement five roles because reference adjudication uses folds 2 and 3. The diagram reports design counts, not confidence intervals.

F K562 PERTURB-SEQ SUPPLEMENT

F.1 DETAILS OF DATA PROCESSING

We analyze a prespecified regulator subset of the day-8 K562 genome-wide Perturb-seq screen, rather than the full intervention library. Before examining intervention–target effect estimates or sealed evaluation outcomes, we retained single-perturbation interventions whose genes were regulators with at least one edge in the pruned ENCODE K562 TF ChIP-seq/promoter working graph. We further required at least 100 quality-controlled cells and representation in at least 20 batches per intervention. These outcome-independent criteria yielded 231 interventions, which were assigned by a fixed-seed random split to predictor training (69), calibration (115), and testing (47). Six nonoverlapping batch folds separate predictor fitting (fold 0), calibration-score construction (fold 1), reference adjudication (folds 2–3), auxiliary tuning (fold 4), and sealed outcome evaluation (fold 5).

The frozen response predictor is a graph-distance kernel smoother of fold-0 pseudobulk log2 rate-ratio estimates from the 69 training interventions. Its bandwidth is selected from $\{0.5, 1, 2, 4\}$ by leave-one-intervention-out prediction error on the independent auxiliary fold 4; the selected value is 4. All compared methods share this predictor and the same fold-1 calibration scores. The predictor, bandwidth, and score scales are fixed before the sealed fold-5 outcomes are accessed.

For each intervention–target pair, reference folds 2–3 provide an inverse-variance-weighted log2 fold-change estimate $\widehat{\theta}^{\mathrm{ref}}$ and standard error $\widehat{\sigma}^{\mathrm{ref}}$. We form the approximate 95% interval $[L, U] = [\widehat{\theta}^{\mathrm{ref}} - 1.96\widehat{\sigma}^{\mathrm{ref}}, \widehat{\theta}^{\mathrm{ref}} + 1.96\widehat{\sigma}^{\mathrm{ref}}]$ and use a prespecified equivalence margin $\Delta = 0.50$ log2FC. A pair is *reference-safe* if $[L, U] \subseteq [-\Delta, \Delta]$, *reference-unsafe* if $L > \Delta$ or $U < -\Delta$, and *unresolved* otherwise. For contamination counting and verification, safe maps to 0, while unsafe and unresolved are conservatively combined as *reference non-safe* and map to 1. The primary coverage analysis uses the same 5,994 test pairs that are predicted safe by the working graph and classified reference-safe by the independent reference folds; neither unsafe nor unresolved pairs are treated as safe test outcomes. We fix $\tau = 3$ and a planned verification fraction of 0.30. These operational reference classifications are not exact causal labels.

The day-6 essential screen contains 310,385 cells, 8,563 genes, 2,058 interventions, and 48 batches; the day-8 genome-wide screen contains 1,989,578 cells, 8,248 genes, 9,867 interventions, and 267 batches. QC-passing cells are aggregated within intervention and batch fold into pseudobulk log-rate-ratio outcomes. Cells are observational units, pseudobulks are effect-estimation units, intervention–target pairs are conformal-score units, and held-out fold effects are evaluation outcomes. The working graph contains 5,280 nodes and 24,075 edges, obtained from K562 TF ChIP-seq peak/promoter overlaps without reading Perturb-seq pair outcomes. As illustrated in Figure 16, we separate cell-level input, pseudobulk effect estimation, intervention-level splitting, graph-local calibration, and sealed evaluation.

Pseudobulk outcome definition. Let $C_{ui}^{(b)}$ and $L_u^{(b)}$ denote the aggregated target count and library size for intervention u in fold b; subscript 0 denotes the matched control pseudobulk. The implemented estimator is

$$\widehat{\theta}_{ui}^{(b)} = \log_2 \left\{ \frac{(C_{ui}^{(b)} + 0.5)/(L_u^{(b)} + 1)}{(C_{0i}^{(b)} + 0.5)/(L_0^{(b)} + 1)} \right\}.$$

The count pseudocount is 0.5 and the library-size offset is 1. The approximate variance is $\{1/(C_{ui}^{(b)}+0.5)+1/(C_{0i}^{(b)}+0.5)\}/(\log 2)^2+2\widehat{\phi}_i$, where $\widehat{\phi}_i$ is the shrinkage-estimated excess log2-effect variance learned using training interventions and non-evaluation folds. This is a log-rate-ratio/delta-variance estimator, not a fitted negative-binomial GLM. The six-fold mapping above assigns two independent folds to the single adjudication role.

Frozen predictor and nonconformity scores. Let $\widehat{\theta}_{ui}^{(b)}$ and $\widehat{\mathrm{se}}_{ui}^{(b)}$ denote the pseudobulk log2 rate-ratio estimate and its standard error for intervention u, target i, and batch fold b. These quantities are obtained from pseudobulk counts using a log-rate-ratio estimator with shrinkage-estimated overdispersion; we do not fit a neural network. Let $d_{\widehat{G}}^{\mathrm{und}}$ be shortest-path distance in the undirected version of the outcome-independent working graph, and set

$$w_\lambda(a,u)=\exp\Big\{-d_{\widehat{G}}^{\mathrm{und}}(a,u)/\lambda\Big\}\,\mathbf{1}\Big\{d_{\widehat{G}}^{\mathrm{und}}(a,u)\le 8\Big\}.$$

For a calibration or test intervention a, the prediction center is

$$\widehat{f}_i(a)=\begin{cases}\dfrac{\sum_{u\in\mathcal{A}_{\mathrm{train}}} w_{\widehat{\lambda}}(a,u)\widehat{\theta}_{ui}^{(0)}}{\sum_{u\in\mathcal{A}_{\mathrm{train}}} w_{\widehat{\lambda}}(a,u)}, & \sum_u w_{\widehat{\lambda}}(a,u)>0,\\[2ex] \dfrac{1}{|\mathcal{A}_{\mathrm{train}}|}\displaystyle\sum_{u\in\mathcal{A}_{\mathrm{train}}}\widehat{\theta}_{ui}^{(0)}, & \text{otherwise.}\end{cases}$$

We select $\widehat{\lambda}$ from $\{0.5,1,2,4\}$ by minimizing the root-mean-squared discrepancy between fold-4 estimates and leave-one-intervention-out predictions formed from fold-0 training estimates. This selects $\widehat{\lambda}=4$.

For each target, let $\widehat{v}_i$ be the root-mean-squared fold-4 leave-one-out prediction error, floored at 10^{-4}. The fold-1 calibration score is

$$R_i^{(a)}=\frac{\left|\widehat{\theta}_{ai}^{(1)}-\widehat{f}_i(a)\right|}{\sqrt{\left(\widehat{\mathrm{se}}_{ai}^{(1)}\right)^2+\widehat{v}_i^2}}.$$

For a test intervention, the corresponding scale is fixed before evaluation as

$$c_i^\star=\sqrt{\widehat{v}_i^2+\widetilde{\mathrm{se}}_i^{\,2}},$$

where $\widetilde{\mathrm{se}}_i$ is the median standard error across training interventions and folds 0 and 4. A selected score quantile $q_i^\star$ therefore gives the response-scale interval $[\widehat{f}_i(a^\star)-c_i^\star q_i^\star,\ \widehat{f}_i(a^\star)+c_i^\star q_i^\star]$. The validity assumption concerns exchangeability of these resulting safe scores; standardization alone does not establish it.

The global reference non-safe rate $\delta_g=0.254$ is $4896/19271=0.25406$, computed over working-graph-predicted-safe *calibration* intervention–target pairs, pooling targets. Its numerator includes both definitively unsafe and unresolved reference classifications. It remains fixed while test queries select different local subsets of those calibration pairs. The separate detection endpoint contains 5,994 reference-safe and 32 definitively reference-unsafe *test* pairs after unresolved pairs are excluded: $32/(5994+32)=0.00531$. These fractions have different populations and different positive-label definitions; they are not two estimates of the same prevalence, and their difference is not evidence that calibration was intentionally enriched.

F.2 Local risk and repeated verification

In the upper panels of Figure 17, local query strata have different reference non-safe fractions despite the fixed global calibration rate 0.254. CGLCP responds through its certificate and rank, whereas a global-rate substitution cannot represent this heterogeneity. The calibration non-safe fraction and the separate test-set unsafe prevalence have different populations and label definitions, as detailed in Appendix F.1; their difference is not a measure of distribution shift by itself.

We isolate audit randomness by repeating verification 100 times on fixed scores, graph, pools, and outcomes, with results in the lower panels of Figure 17. Larger verification budgets tighten the certificate and raise finite availability, linking the K562 efficiency loss to the same finite-population estimation mechanism as in simulation. These repetitions assess stability of verification, not replication across biological samples or cell types.

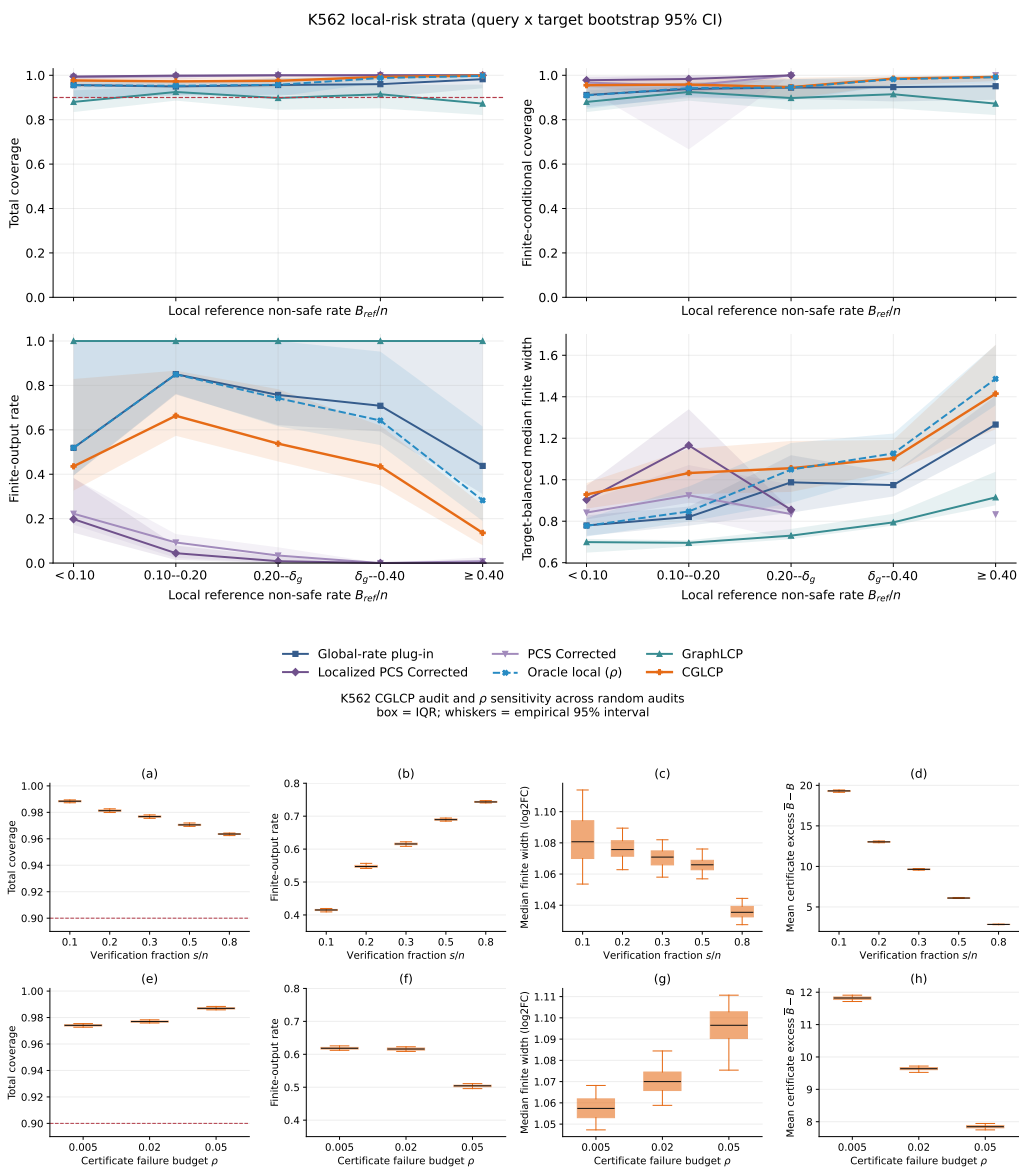


Figure 17: K562 local-risk strata (top): shaded bands are 95% two-way query-intervention × target-gene bootstrap confidence intervals. Verification/ρ sensitivity (bottom): 100 verification seeds on fixed scores, graph, and outcomes; boxes show the interquartile range, center lines the median, and whiskers the empirical 2.5th–97.5th percentiles. The latter summarize verification randomness, not biological sampling confidence.

F.3 Complete sentinel and positive-endpoint analyses

We assess how CGLCP's low false-alarm rates relate to finite availability and sensitivity conditional on finite output across shift sizes, as shown in Figure 18. Because intervals remain frozen during the shifts, we evaluate the consequences of the original calibrated decision rather than a refitted detector.

At matched false-alarm rates, GraphLCP can have greater ranking power, limiting the claim of CGLCP's advantage to calibration of its original interval decision, as shown in Figure 19. Thresholds are tuned on half the query interventions and evaluated on the other half over 200 repeated splits. The separate reference-adjudicated endpoint contains only 32 definitively unsafe pairs after excluding

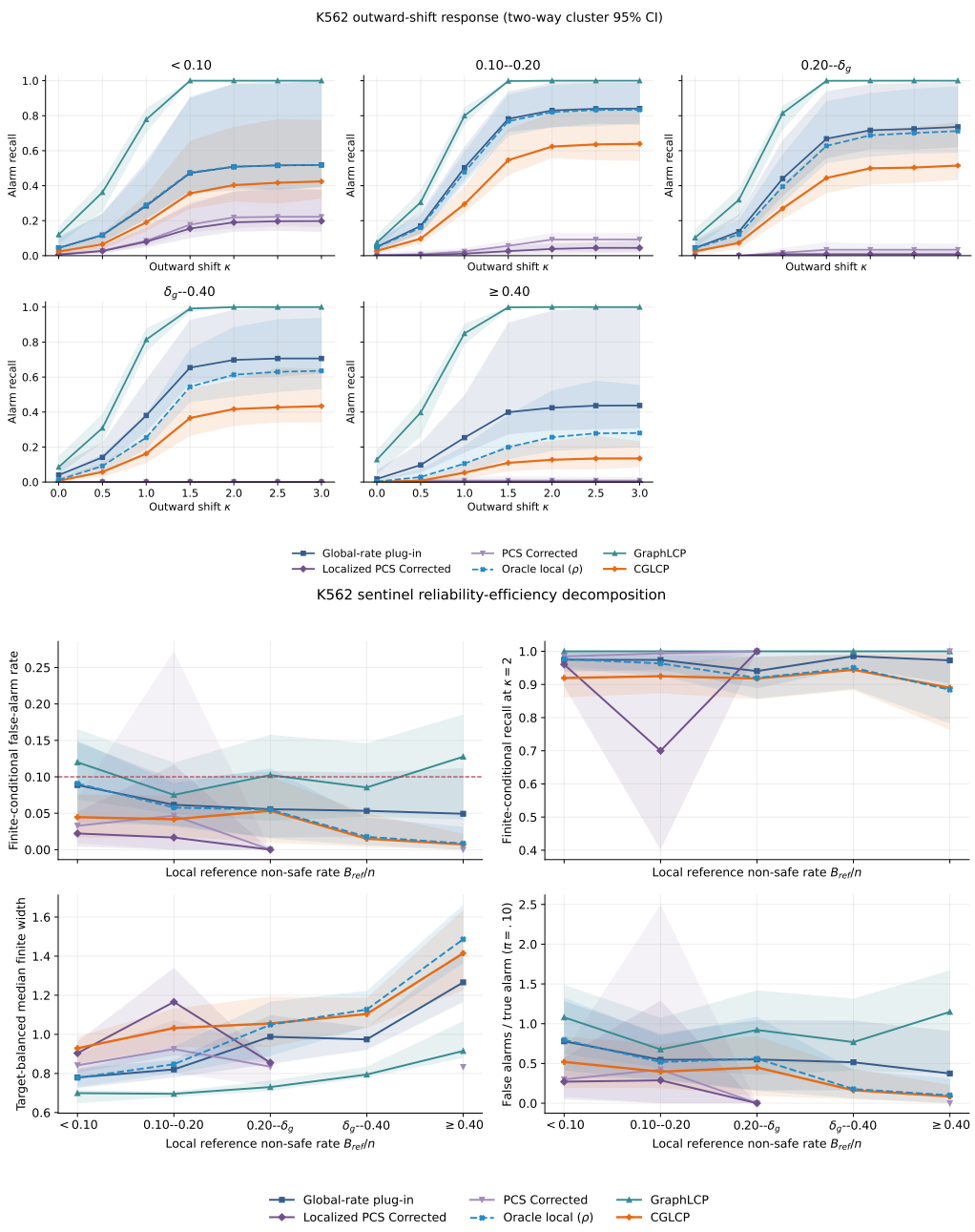


Figure 18: K562 outward-shift response and finite-conditional mechanism decomposition. Shaded regions are pointwise 95% two-way cluster bootstrap confidence intervals over query interventions and target genes, including bootstrap-transformed alarm ratios. Predictions and intervals remain frozen throughout; widths condition on finite output.

unresolved pairs; its wide recall intervals support an exploratory assessment rather than a precise ranking of biological detection power, as shown in Figure 19.

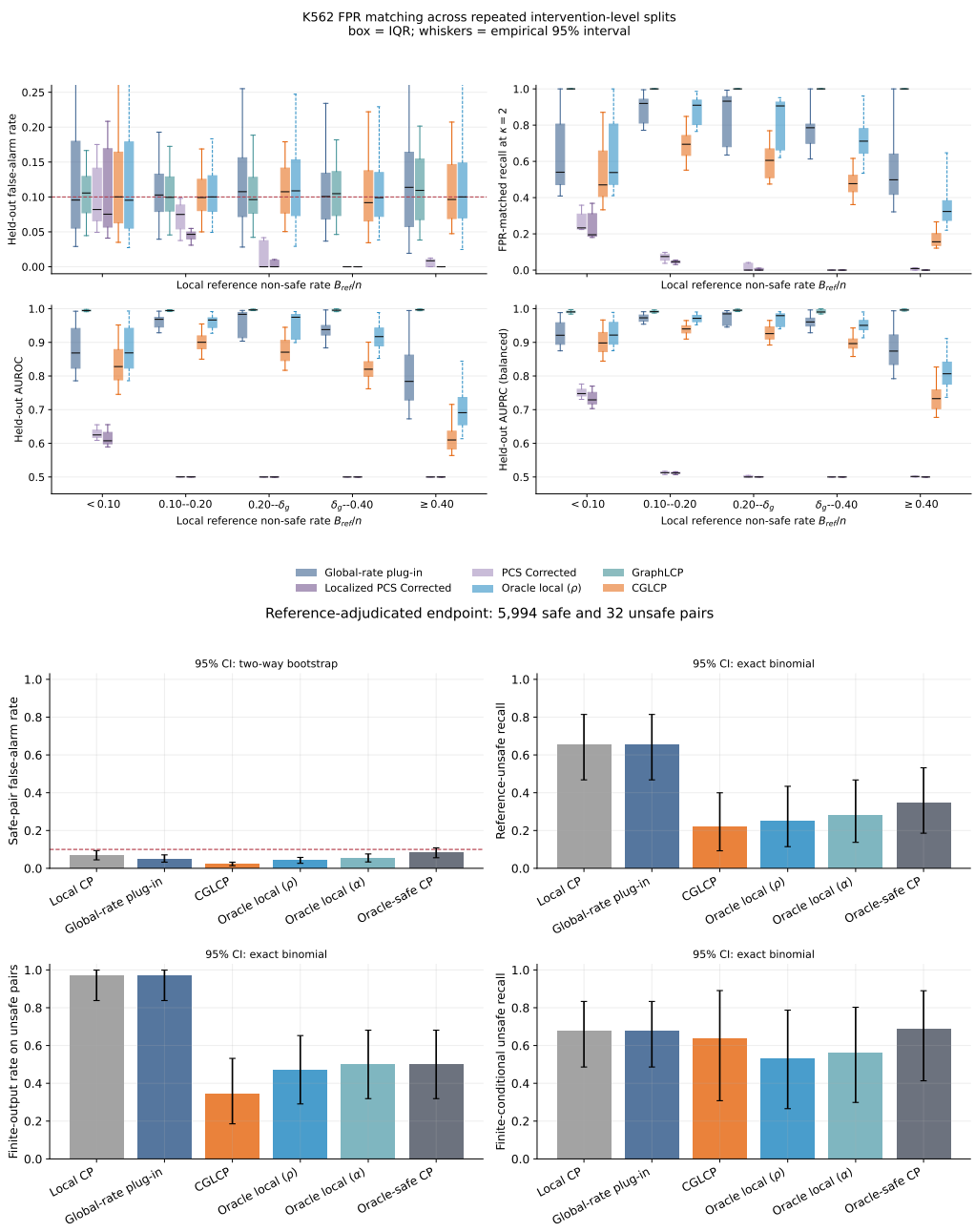


Figure 19: Top: FPR-matched detection over 200 intervention-level tuning/evaluation splits; boxes show medians and interquartile ranges, with empirical 2.5th–97.5th percentile whiskers, not confidence intervals for a mean. Bottom: 5,994 reference-safe and 32 definitively reference-unsafe test pairs, excluding unresolved pairs. Safe FPR error bars are 95% two-way cluster bootstrap confidence intervals; unsafe recall/finite-rate bars are 95% Clopper–Pearson binomial intervals. The binomial intervals do not adjust for shared intervention/target dependence and are descriptive for this small endpoint.

F.4 Design and baseline robustness

We vary radius, graph sparsity, and the target universe, then check GraphLCP bandwidth and PCS thresholds on common cohorts, with results in Figure 20. The primary radius avoids the nearly empty pools at $\tau = 1$; GraphLCP remains finite across its bandwidth scan, while PCS retains high coverage

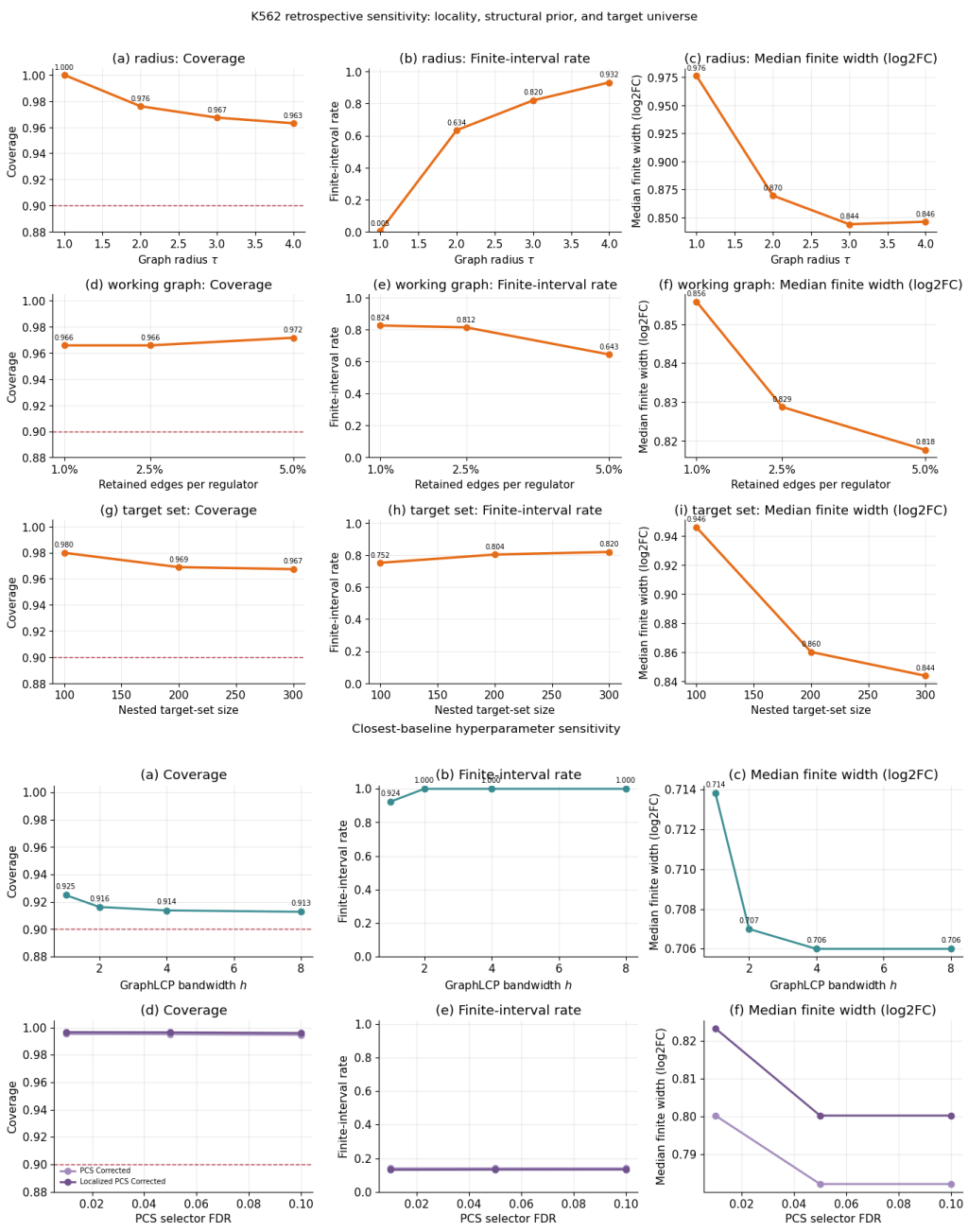


Figure 20: K562 design and baseline sensitivity: radius, working-graph sparsity, target universe, GraphLCP bandwidth, and PCS selection threshold. These panels show point estimates without confidence intervals. Widths are finite-conditional and must be interpreted with finite availability; retrospective sensitivities do not constitute new sealed external validation.

with low availability. We use these separate sensitivity analyses to assess how the observed trade-off depends on design and baseline choices, without interpreting them as a joint factorial comparison. The pair-averaged coverage/availability and pooled finite-width medians used here differ from the main target-balanced summaries and should not be compared numerically as identical estimands.

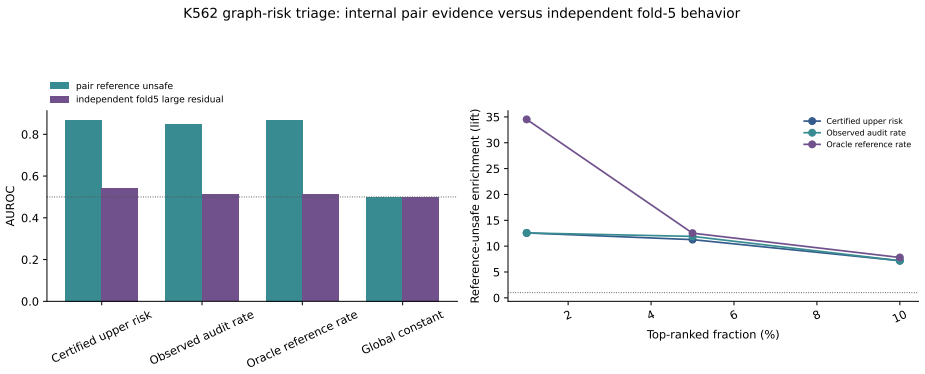


Figure 21: Exploratory certificate-based graph-risk prioritization on K562. Curves show point estimates of reference-risk enrichment and selected fractions, without confidence intervals. They evaluate prioritization of independent follow-up, not causal-edge recovery or experimentally validated graph repair.

F.5 Exploratory graph-risk triage

We evaluate certificate-based prioritization through reference-risk enrichment and the fraction selected for follow-up, as shown in Figure 21. This assesses whether CGLCP can identify neighborhoods that merit additional checking, rather than infer which edge is missing. Without a third confirmation source, the enrichment remains an exploratory diagnostic and does not demonstrate causal-edge recovery or improved coverage after graph repair.

G The Use of Large Language Models (LLMs)

Generative AI tools assisted manuscript polishing and code refactoring. All mathematical claims, experimental choices, code outputs, and citations were reviewed by the authors, who take responsibility for the final content.