PerturbCausal: Virtual Cell Model Predictions Do Not Substitute for Real Perturb-seq in Causal Network Recovery

Aayan Alwani [1] Ethan Wang [1]

Abstract

Virtual cell models (VCMs) are deep networks that predict transcriptional responses to genetic perturbations and are promoted as in silico substitutes for CRISPR screens. Ahlmann-Eltze & Huber (2025) showed they fail to beat simple linear baselines at per-gene prediction; we ask whether predicted Perturb-seq can substitute for real Perturb-seq one step downstream, in causal gene regulatory network recovery. PerturbCausal holds the causal-discovery algorithm fixed and varies only the data source, scoring the network recovered from VCM predictions against the network the same algorithm recovers from real data. We evaluate four VCMs (GEARS, CPA, Geneformer, and the 2025 Arc Institute STATE foundation model) through four discovery backends (GIES, inspre, PC, NOTEARS) on three Perturb-seq datasets spanning cell context (K562, RPE1) and modality (CRISPRi, CRISPRa). Across every condition tested, no deep VCM exceeds a sparse linear comparator on causal recovery and none reaches real-data quality; STATE's apparent single-seed advantage (F1 = 0.462) does not replicate, falling to 0.041 and 0.048 at two further seeds. The failure has one signature: predicted effect-size distributions are inflated and nearly identical across perturbations, the pooled-scope counterpart of the mode collapse reported elsewhere (Mejia et al., 2025). Quantile-matching calibration corrects the magnitude marginal exactly yet leaves the causal gap open on held-out perturbations, locating the residual in joint-distribution structure (gene-gene covariance and edge orientation) that distributional calibration cannot reach. The benchmark, calibration diagnostic, and reproducibility code are released as an open-source Python package.

[1]Independent Researcher. Correspondence to: Aayan Alwani <alwaniaayan6@gmail.com>.

Proceedings of the 43rd International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026.
Copyright 2026 by the author(s).

1. Introduction

Virtual cell models (VCMs) predict cellular responses to genetic perturbations and are proposed as computational substitutes for CRISPR screens that cost millions of dollars per cell line (Replogle et al., 2022; Bunne et al., 2024). The dominant evaluation paradigm—exemplified by PerturBench (Wu et al., 2024), the Virtual Cell Challenge (Roohani et al., 2025), and the Ahlmann-Eltze & Huber comparison (Ahlmann-Eltze & Huber, 2025)—measures per-gene reconstruction accuracy of predicted transcriptional responses. Across these benchmarks, deep learning methods including graph neural networks (GEARS; Roohani et al. 2024), conditional variational autoencoders (CPA; Lotfollahi et al. 2023), and pretrained foundation transformers (Geneformer, scGPT; Theodoris et al. 2023; Cui et al. 2024) consistently fail to outperform simple linear baselines on the marginal-prediction task.

Reconstruction accuracy, however, is a property of the marginal distribution of effects, whereas the downstream applications that motivate VCMs—combination prediction, target identification, GRN inference—depend on the joint distribution across genes (Chevalley et al., 2025; Callahan et al., 2025): a model can match every per-gene mean and still distort the gene-gene covariance and conditional-independence structure that interventional causal discovery exploits. CausalBench evaluates causal discovery on real Perturb-seq (Chevalley et al., 2025) but does not test whether VCM predictions can substitute for it; a NeurIPS 2025 perspective called for causal-recovery evaluation of VCMs but did not implement one (Callahan et al., 2025). Adjacent work surfaces pieces—mode collapse and weighted

losses (Mejia et al., 2025), metric calibration (Miller et al., 2025), and the inspre sparse-inverse causal-discovery algorithm (Brown et al., 2025)—but none integrates VCM predictions into a downstream causal-discovery pipeline to test substitutability directly.

We close this gap. The contributions of this paper are:

1. A causal-recovery substitutability benchmark for VCMs. Predictions from four VCMs (GEARS, CPA, Geneformer, STATE) and a sparse linear comparator are passed through four causal-discovery backends (GIES (Hauser & Bühlmann, 2012), inspre (Brown et al., 2025), PC, NOTEARS) on three Perturb-seq datasets spanning cell context (K562, RPE1) and perturbation paradigm (CRISPRi, CRISPRa) (Replogle et al., 2022; Norman et al., 2019), and the recovered graphs are compared against per-backend real-data ground truth. No deep model exceeds the linear comparator on any backend or dataset evaluated, and none reaches real-data quality.

2. A diagnosis of the failure mode. All three deep models inflate predicted effect-size distributions by a factor of approximately 23 relative to real biology. The inflation magnitudes are identical to within 0.3 percentage points across radically different architectures (GNN, VAE, transformer), arguing the failure is objective-driven (per-gene MSE) rather than architecture-driven.

3. An empirical signal probe. The inspre algorithm exposes that CPA and Geneformer's average causal effect (ACE) matrices fall below the regularization-path detection threshold, recovering zero edges where real data and ElasticNet recover approximately 12,000.

4. A diagnostic decomposition of the substitutability gap. The gap separates into a magnitude marginal (mode collapse and effect-size inflation, two views of one failure) and a joint-distribution residual. Post-hoc quantile-matching calibration (Bolstad et al., 2003; Reinhard et al., 2001) corrects the magnitude marginal exactly, yet under a held-out perturbation split no calibration ordering closes the causal gap. We make the decomposition precise: conditional independence is invariant under per-gene monotone calibration (Proposition 5.1), so marginal rescaling provably cannot close the gap, identifying the residual as the off-diagonal joint (precision-support) structure and motivating a training-time remedy.

2. Related Work

A 2024–2025 literature converges on deep learning not yet beating simple baselines on per-gene reconstruction (Ahlmann-Eltze & Huber, 2025; Wong et al., 2025; Csendes et al., 2025; Wei et al., 2025; Viñas Torné et al., 2025; Bendidi et al., 2024; Wenteler et al., 2025; Zhu et al., 2025; Li et al., 2025; Mejia et al., 2025), standardized by PerturBench (Wu et al., 2024) and the Virtual Cell Challenge (Roohani et al., 2025); none evaluate downstream causal recovery. A parallel literature builds causal discovery on real Perturb-seq—GIES (Hauser & Bühlmann, 2012), inspre (Brown et al., 2025) (the backends we reuse), BEELINE (Pratapa et al., 2019), Scribe (Qiu et al., 2020), CASCADE (Cao et al., 2025), IV methods (Park et al., 2026), CausalBench (Chevalley et al., 2025), with Kernfeld et al. (2024) cautioning transcriptome data alone cannot control false discoveries; PerturbCausal is downstream, asking whether predicted data can substitute. This is orthogonal to GRN-conditioned simulators (GRouNdGAN (Zinati et al., 2024)), the data-pathology axis of Fernandez-de-Retana et al. (2026), the taxonomy of Callahan et al. (2025), the featurizer use of Wu et al. (2025), and retraining-based mode-collapse losses (Mejia et al., 2025); our edge-set-rank causal F1 against a fixed real-data GIES ground truth is complementary to, not contradicted by, the dynamic-range-fraction results of Miller et al. (2025) (full per-category discussion in §A).

3. Methods

3.1. Problem statement

A virtual cell model f maps a perturbation target p and a control state to a predicted profile $\hat{x}^{(p)} \in \mathbb{R}^G$, and a causal-discovery algorithm $\mathcal{A}$ maps interventional samples $\{(\hat{x}^{(p)}, p)\}_{p \in \mathcal{P}}$ to a directed edge set $\hat{E}_{\mathcal{A}}$. The substitutability question is whether $\hat{E}_{\mathcal{A}}$ recovered from predicted data approximates $E^{\star}_{\mathcal{A}} = \mathcal{A}(\{x^{(p)}\})$ recovered from real Perturb-seq under the same $\mathcal{A}$; holding $\mathcal{A}$ fixed and varying only the data source attributes any gap to the predictor f. Anchoring ground truth to the same backend on real data isolates the predictor's contribution, at the cost that edge-set F_1 is comparable across predictors but not across backends, so every F_1 is reported within a backend. The linear predictor is a primary comparator, not a floor: the question is asked symmetrically of deep and linear models, following the finding that linear baselines are competitive on per-gene prediction (Ahlmann-Eltze & Huber, 2025).

3.2. Datasets

We use three Perturb-seq datasets spanning in-distribution, cross-context, and cross-paradigm generalization: Replogle K562 (Replogle et al., 2022) (CRISPRi, primary; 1,861 eligible targets), Replogle RPE1 (Peidli et al., 2024) (CRISPRi, cross-context), and Norman 2019 (Norman et al., 2019) (CRISPRa, cross-paradigm; 102 single-gene targets); the primary K562 evaluation uses a 200-gene subset that fills exactly ten 20-gene GIES partitions, and the scale sweep $N \in \{50, 75, 100, 125, 150, 175, 200\}$ confirms rankings are stable across subset size (full dataset statistics and gene-subset construction in §B).

3.3. Scope of empirical claims

Table 1 enumerates exactly which dataset, predictor, backend, and seed combinations produced each numerical result. Claims outside the table are stated as limitations (§5), not findings.

3.4. Virtual cell models

Three deep predictors are evaluated with their canonical architectures and default hyperparameters (full settings in §D.1): GEARS (Roohani et al., 2024), a graph neural network over a Gene Ontology co-expression graph; CPA (Lotfollahi et al., 2023), a conditional variational autoencoder with adversarial covariate disentanglement; and Geneformer V2 (Theodoris et al., 2023), a 104M-parameter transformer pretrained on ∼30M profiles, perturbed in silico via embedding cosine similarity. STATE (Adduri et al., 2025; Arc Institute, 2025), the 2025 Arc Institute foundation model, is evaluated on K562 only (§5). GEARS and CPA are retrained from scratch on each of K562, RPE1, and Norman; Geneformer and STATE are not retrained per dataset.

3.5. Sparse linear comparator

For each evaluation gene a sparse linear regression is fit on held-out non-targeting control cells. At prediction time the perturbed gene is set to 10% of its control mean (modeling ∼90% CRISPRi knockdown) and the remaining regressors predict the downstream profile. Real perturbation values enter neither training nor prediction, eliminating leakage (Zou & Hastie, 2005). The fit is deterministic and refit per dataset.

3.6. From predicted means to backend inputs

Overlay sampling converts predicted means into cell matrices that preserve gene-gene covariance (independent negative-binomial sampling instead collapses GIES recovery below 2% of the real-data ceiling): control cells are bootstrap-sampled and scaled by the predicted fold change with Poisson noise, matching the CausalBench protocol (Chevalley et al., 2025). On K562, GIES, PC, and NOTEARS consume overlay cells; for the cross-dataset evaluation PC, NOTEARS, and inspre consume the square average causal effect (ACE) matrix $\mathrm{ACE}_{ij} = \mathbb{E}[X_j \mid \mathrm{do}(X_i)] - \mathbb{E}[X_j \mid \mathrm{ctrl}]$ on the canonical perturbed-gene set, so the node set is identical across predictors and only the predicted intervention effect varies (overlay-sampling sensitivity sweep in §B).

3.7. Causal-discovery backends

Four backends from distinct algorithm families are run identically on every condition. Each defines its own real-data ground truth $E^\star_{\mathcal{A}}$ as that backend applied to the real-data input (overlay cells on K562; the real-data ACE cross-dataset):

- GIES (score-based, interventional) (Hauser & Bühlmann, 2012): greedy interventional equivalence search, ten 20-gene partitions per CausalBench.
- PC (constraint-based, observational) (Spirtes & Glymour, 2000): Fisher-Z conditional-independence tests on the square ACE over perturbed genes.
- NOTEARS (continuous optimization, gradient) (Zheng et al., 2018): linear DAG learning with an acyclicity constraint; edges are taken by post-hoc thresholding of the unconstrained weight matrix.
- inspre (sparse inverse of the ACE) (Brown et al., 2025): the regularization weight is selected by held-out test error; when the cross-validation objective is undefined the path point whose edge count is closest to the real-data ground-truth edge count is used, and condition edge counts are budget-matched to that ground truth.

3.8. Chance reference and metric

The primary chance reference is a permutation null computed per backend: each column of the real-data ACE is independently shuffled across perturbations, removing cross-gene joint structure while preserving each gene's marginal effect-magnitude distribution. The backend is run on K such nulls ($K = 30$, except $K = 24$ for PC on Norman), giving a mean and a $[2.5, 97.5]$ percentile band against $E^\star_{\mathcal{A}}$; a predictor's

Dataset	Predictors	Backends	Seeds
K562 (CRISPRi)	ENet, CPA, GEARS, Geneformer, STATE, GRNBoost2	GIES, inspre, PC, NOTEARS	3, 5, 20
RPE1 (CRISPRi)	ENet, CPA, GEARS, GRNBoost2	GIES; PC, NOTEARS, inspre	1
Norman (CRISPRa)	ENet, CPA, GEARS, GRNBoost2	GIES; PC, NOTEARS, inspre	1

Table 1. Scope of empirical claims. Substitutability is tested with two architecturally distinct trained VCMs (GEARS, CPA) on all three datasets; cross-dataset cells are single-seed. Geneformer is restricted to K562 because its prediction is a cosine-similarity heuristic over the control mean, not learned in-silico-perturbation expression (§5); STATE is K562-only.

F_1 is read as above the band (signal beyond a structure-destroyed input), inside (at chance), or below (worse than the shuffled null). A secondary density-matched random-edge floor is reported only to confirm a predictor is not emitting a trivially small or large edge set (analytic form in §B). The primary metric is edge-set F_1 against the per-backend real-data ground truth (precision, recall, AUPRC, SHD, and multi-hop credit at $k \in \{2, 3\}$ secondary).

3.9. Statistical testing and calibration

K562 headline numbers are across-seed means with 95% t-CIs over $n \in \{3, 5, 20\}$ seeds; deep-vs-linear equivalence is assessed by two one-sided tests (TOST) with margin $\Delta = 0.045$ (the technical-duplicate noise floor), and cross-dataset cells are single-seed estimates read against the per-backend permutation band of §3.8. Post-hoc quantile-matching calibration (Bolstad et al., 2003; Reinhard et al., 2001) maps the predicted $|\log_2 \mathrm{FC}|$ distribution onto the real-data distribution, with two joint-structure extensions evaluated under a 160/40-perturbation split (full statistical-testing protocol, implementation specifics, and the calibration algorithm in §B and §C.15).

4. Results

4.1. Causal recovery on K562 at N=200

We evaluate all conditions across five random seeds (42, 123, 456, 789, 1024) with t-distribution 95% CIs and paired permutation tests (10,000 permutations). Against 1,197–1,230 ground-truth edges per seed, all four predictive methods fall within a tight band (ElasticNet F1 = 0.426, CPA = 0.412, GEARS = 0.406, Geneformer = 0.425, overlay-resampled real = 0.422; Table 2, Figure 2) with no pairwise difference significant at $p < 0.05$; ElasticNet (0.426) and Geneformer (0.425) are statistically tied. A two-one-sided-tests (TOST) analysis establishes equivalence at the technical-duplicate margin ($\Delta = 0.045$ F1): the ElasticNet−deep paired differences lie within $\pm\Delta$ for all three deep models (90% CI: CPA $+0.014\,[-0.001, +0.029]$, GEARS $+0.020\,[+0.005, +0.035]$, Geneformer $+0.0002\,[-0.012, +0.013]$). The observational baseline GRNBoost2 falls to F1 0.098 $[0.094, 0.103]$, disjoint from every predictive method by ≥ 0.27 F1 (gap significant at $n = 20$, $p < 10^{-4}$); the random-bootstrap floor is F1 ≈ 0.04 and the technical-duplicate noise floor 0.045 ± 0.003 (§C.2). The headline is not that ElasticNet outperforms deep models—the previous single-seed claim does not survive multi-seed evaluation—but that a 104M-parameter pretrained transformer, a GNN with Gene Ontology priors, and an adversarial VAE all achieve causal-recovery F1 statistically indistinguishable from a sparse linear regression with $\approx 200^2$ free coefficients.

4.2. Effect-size inflation across architectures

Real K562 responses are sparse (2.7% of (perturbation, gene) effects exceed $|\log_2 \mathrm{FC}| = 0.5$; leading singular vector captures 56.8% of variance), yet all three deep models inflate that fraction to 63.2% (GEARS), 63.4% (CPA), 63.5% (Geneformer)—a 23-fold inflation identical across architectures to within 0.3 percentage points (Figure 3), with mean per-perturbation L_1 norm roughly tenfold larger than empirical (real 22.8; GEARS 226.9, CPA 226.6, Geneformer 227.4). The control-regime covariance is preserved well (correlation of correlation matrices: GEARS = 0.967, CPA = 0.996, Geneformer = 0.996), so the models learn baseline gene-gene relationships but fail to predict sparse, structured deviations upon perturbation.

Mode collapse and inflation are one failure. The pooled 23-fold inflation is the pooled view of the same homogenization reported as mode collapse (Mejia et al., 2025; Ahlmann-Eltze & Huber, 2025): per-perturbation Δ-variance ratios are below 1 for all three deep models (median 0.461 GEARS, 0.432 CPA, 0.008 Geneformer) and inter-perturbation shift-cosine is 0.845 in CPA vs. 0.084 in real—a single dense response pattern across perturbations that, pooled over off-target genes, reads as inflation (full reconciliation in §C.1).

Table 2. Causal recovery on K562 at N=200 genes against real-data GIES ground truth. Multi-seed mean F1, precision, recall, and edge count across $n = 5$ random seeds (42, 123, 456, 789, 1024) with 95% t-distribution CIs. "Tech. duplicate" is F1 between two GIES recoveries on disjoint random halves of real cells (mean of 3 splits). Pairwise permutation tests find no significant difference at $p < 0.05$ among the four predictive methods. The ElasticNet–GRNBoost2 gap is significant at $n = 20$ ($p < 10^{-4}$ paired permutation), while the $n = 5$ deep-model comparisons have a minimum attainable two-sided p of $2/2^5 = 0.0625$, not yet certifiable despite the same effect-size gap.

Condition	F1 [95% CI]	Precision [95% CI]	Recall [95% CI]	Edges
Real Data (ceiling)	1.000	1.000	1.000	—
Overlay Real (UB, n=20)	0.429 [0.419, 0.438]	0.418 [0.409, 0.427]	0.440 [0.429, 0.451]	1,224
ElasticNet (n=20)	0.418 [0.408, 0.427]	0.406 [0.396, 0.416]	0.430 [0.420, 0.441]	1,233
CPA (n=5)	0.412 [0.392, 0.432]	0.407 [0.390, 0.423]	0.417 [0.393, 0.441]	1,242
GEARS (n=5)	0.406 [0.388, 0.424]	0.404 [0.388, 0.419]	0.408 [0.385, 0.431]	1,225
Geneformer (n=5)	0.425 [0.412, 0.439]	0.422 [0.409, 0.434]	0.429 [0.413, 0.445]	1,235
GRNBoost2 (obs, n=20)	0.098 [0.096, 0.101]	0.074 [0.072, 0.076]	0.147 [0.144, 0.151]	2,327
Tech. duplicate (noise floor)	0.045 [0.042, 0.048]	0.045	0.045	∼1,090
Random bootstrap (floor)	≈ 0.04	—	—	—

Technical-duplicate noise floor. Two GIES recoveries on disjoint random halves of real K562 cells agree at only $F_1 = 0.045 \pm 0.003$ (Table 2), so real-cell sampling variance dominates GIES at N=200 and all evaluated methods operate roughly an order of magnitude above this raw cell-split floor (§C.2).

4.3. Inspre signal probe

The inspre algorithm (Brown et al., 2025), applied to the ACE matrix without partitioning, recovers 12,280 edges from real Perturb-seq, exactly 12,280 from ElasticNet, 6,234 from GEARS, and zero from both CPA and Geneformer at the test-error-optimal regularization (Figure 4). The full λ-path validation (§C.7) shows the zero result is binary at the cross-validated λ but continuous across the path, with CPA vanilla the only condition for which the solver fails to traverse a valid λ path.

4.4. Cross-context replication on RPE1

GEARS and CPA retrained from scratch on RPE1 (8,749 genes, 196 targets) recover no more of the real causal graph than the linear baseline or chance: against the real-data RPE1 ground truth (1,137 edges, seed 42), GEARS F1 = 0.414, CPA = 0.450, ElasticNet = 0.437, random-edge floor 0.436, and the perfect-real-means overlay upper bound itself reaches only 0.418. The substitutability gap therefore persists under a change of cellular context (single-seed, qualitative claim; full prose, N=200 structural comparison, scale sweep, and reversed-direction analyses in §C.9, §C.5, §C.6).

4.5. Backend robustness across datasets

The cross-context (RPE1) and cross-paradigm (Norman) tests above use GIES. To test whether the conclusion is specific to a score-based interventional algorithm, both datasets were re-run through three further backends—PC (constraint-based), NOTEARS (gradient-based), and inspre (sparse-inverse)—with each trained deep model and the linear comparator compared to that backend's permutation chance floor (§3.8). Table 3 reports per-backend F_1 and the three-state verdict; Figure 1 plots each deep model against its backend's chance band.

The substitutability conclusion is backend-invariant in the sense that matters: no deep VCM exceeds the sparse linear comparator on causal recovery under any of the four backends, on either dataset, and no deep model reaches real-data quality. The largest deep-model F_1 across all six cross-dataset cells, 0.440 (CPA under inspre on RPE1), lies inside that backend's chance band; the linear comparator is the strongest predictor in four of the six cells.

The chance comparison, read in three states against the permutation band, holds the strict "at chance" characterization for GIES, NOTEARS, and inspre-on-RPE1 (deep models at or below floor), with two above-floor exceptions that neither exceeds the linear comparator nor reaches substitutable quality: PC on RPE1 (F_1 = 0.078 vs. a 0.017 floor, but PC is observational and discards intervention labels, so this is preserved baseline co-expression) and GEARS under inspre on Norman (0.344 against a 0.329 band edge, a single-seed margin). The cell-by-cell reading—including why inspre's high raw F_1 (GEARS 0.416, CPA 0.440 on RPE1) sits inside an equally high 0.526 $[0.406, 0.583]$ band—is in §C.4.

Table 3. Backend robustness on RPE1 (cross-context) and Norman (cross-paradigm). Per-backend edge-set F_1 for the two trained deep VCMs and the linear comparator (ENet) against each backend's own real-data ground truth; the chance floor is the permutation-null mean with $[2.5, 97.5]$ band (§3.8). Verdicts: $^{\uparrow}$ above band, $^{=}$ inside (at chance), $^{\downarrow}$ below. No deep model exceeds ENet in any cell. ENet under PC on RPE1 is omitted: PC does not converge on the dense ElasticNet ACE within the compute budget.

Dataset	Backend	GEARS	CPA	ENet	Chance floor
Norman	PC	$0.015^{=}$	$0.044^{=}$	0.024	0.015 $[0.00, 0.05]$
Norman	NOTEARS	$0.065^{\downarrow}$	$0.086^{\downarrow}$	0.282	0.527 $[0.44, 0.60]$
Norman	inspre	$0.344^{\uparrow}$	$0.297^{=}$	0.247	0.294 $[0.25, 0.33]$
RPE1	PC	$0.078^{\uparrow}$	$0.078^{\uparrow}$	—	0.017 $[0.01, 0.03]$
RPE1	NOTEARS	$0.080^{\downarrow}$	$0.037^{\downarrow}$	0.140	0.308 $[0.29, 0.33]$
RPE1	inspre	$0.416^{=}$	$0.440^{=}$	$0.586^{\uparrow}$	0.526 $[0.41, 0.58]$

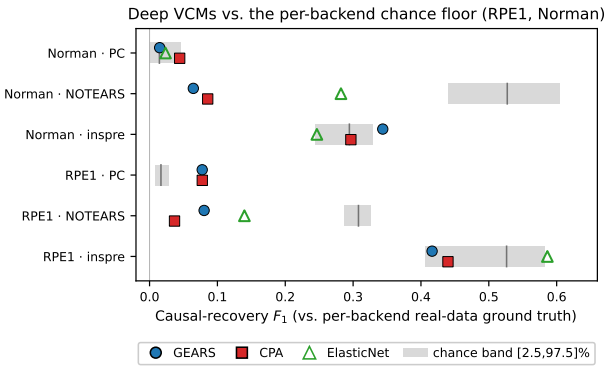

Figure 1. Deep VCM causal-recovery F_1 against the per-backend chance floor on RPE1 and Norman. Each row is one (dataset, backend) cell; the grey band is the permutation-null $[2.5, 97.5]$ chance interval (K=30 permutations, 24 for PC on Norman). Markers are GEARS (circle), CPA (square), and the ElasticNet comparator (open triangle). Both deep models land inside or below the band in four of six cells; the two above-band points (PC on RPE1 at $F_1 = 0.078$; GEARS under inspre on Norman, a single-seed margin) neither exceed the linear comparator nor reach substitutable quality. n_{gt} ground-truth edges per cell range from 67 to 1,327.

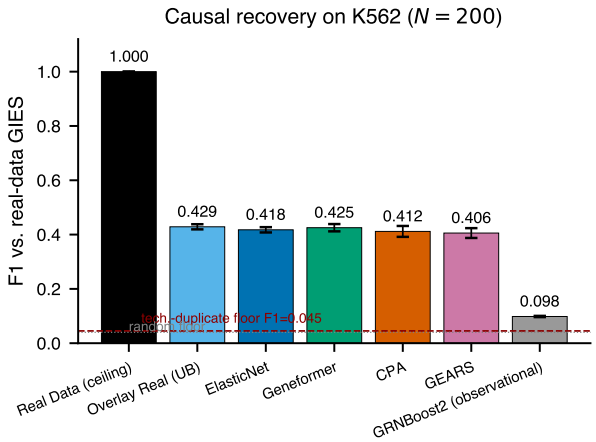

Figure 2. Causal recovery on K562 at N=200 genes against the real-data GIES ground truth. Reference rows (real Perturb-seq, overlay-resampled real, ElasticNet, GRNBoost2) are aggregated over n=20 seeds; the deep VCMs (GEARS, CPA, Geneformer) over the n=5 seeds with cached predictions (42, 123, 456, 789, 1024). Bars are across-seed mean F1; error bars are 95% t-distribution CIs. The dashed red line marks the technical-duplicate noise floor (mean F1 = 0.045 between two GIES recoveries on disjoint random halves of real cells). Wong colorblind-safe palette.

4.6. Calibration corrects magnitude but not held-out causal F1

In-domain quantile-matching calibration corrects the magnitude distribution exactly but moves downstream causal F1 only marginally (CPA 0.389 → 0.405, GEARS 0.406 → 0.413, Geneformer 0.432 → 0.423; §C.3, Table 4, Figure 5), and no calibrated deep model surpasses ElasticNet (F1 = 0.425). Under an honest held-out perturbation split, no calibration ordering beats the uncalibrated baseline on any model (GEARS 0.303 vs. 0.290, CPA 0.309 vs. 0.289, Geneformer 0.319 vs. 0.287; §C.15), locating the residual in joint structure rather than the magnitude marginal.

5. Discussion

Mechanism: magnitude and causal recovery are dissociated. The shared failure is objective-driven, not architecture-driven, and lies in joint rather than marginal structure. Retraining CPA with a Gaussian (per-gene MSE) reconstruction loss instead of its default Negative Binomial increases per-perturbation variance compression (median Δ-variance ratio 0.43 → 0.18) yet raises causal F1 (0.389 → 0.414 at seed 42): more aggressively compressed predictions recover more of the real graph, so the gap is neither in the reconstruction-loss form nor the magnitude marginal but in the joint covariance/conditional-independence

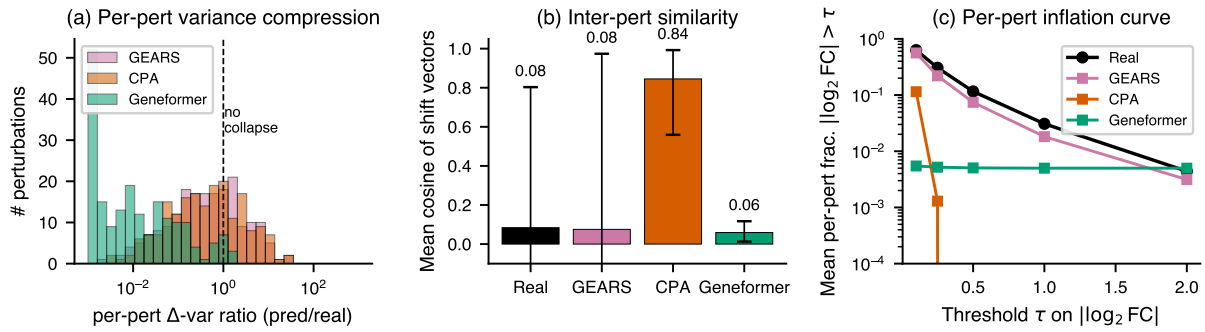

Figure 3. Mode-collapse vs. inflation reconciliation diagnostic on K562 (N=200, 193 perturbations). (a) Per-perturbation Δ-variance ratios (dashed line: ratio = 1); all deep models compress, Geneformer by ∼two orders of magnitude. (b) Inter-perturbation cosine similarity (whiskers $p10$–$p90$); CPA produces nearly the same response across perturbations (0.84 vs. 0.08 real). (c) Mean per-perturbation fraction $|\log_2 \mathrm{FC}| > \tau$ vs. threshold τ; deep models track real or fall below per-perturbation yet exceed real at the full-transcriptome pool. Full panel descriptions in §C.1.

structure GIES extracts (full mechanism analysis in §C.14). This motivates a training-time covariance penalty $\lambda \|\mathrm{Cov}(\hat{\delta}) - \mathrm{Cov}(\delta_{\mathrm{real}})\|_F^2$, released as a unit-tested module and tested by a no-GPU proxy: a learned joint head on CPA's shifts under a held-out 160/40 split (§C.14) raises held-out causal F_1 from vanilla CPA's 0.424 ± 0.025 to 0.535 ± 0.015 ($\Delta = +0.111 \pm 0.034$), the first direct confirmation the gap is reachable by a joint operation that generalizes to unseen perturbations. The covariance-Frobenius term itself does not help: a pre-registered sweep $\lambda \in \{0, 0.1, 0.3, 1, 3, 10, 30, 100\}$ (validation-selected, sealed test touched once) finds the best weight (λ=100) gains only $+0.003 \pm 0.035$ test F_1 over recon-only (3 of 5 seeds) and fails a multi-seed gate—the gain comes from learning a generalizing joint map, not the covariance term. End-to-end retraining scored against the alt-loss F_1 of 0.414 is the natural next experiment.

A six-variant CPA ablation confirms the residual gap is not localizable to any one architectural component (causal-F1 range 0.387–0.418, within vanilla CPA's 95% CI; §C.12).

Why does ElasticNet match (not beat) deep models? ElasticNet F1 (0.416) and the three deep models (0.405–0.425) fall within overlapping 95% CIs: a sparse linear regression with $\approx 200^2$ free coefficients trained only on control cells matches, but does not outperform, a 104M-parameter pretrained transformer. We attribute this to inductive-bias matching: L_1 regularization biases ElasticNet toward sparsity (Zou & Hastie, 2005), matching the empirical sparsity of real perturbation responses (38.6% of K562 effects near zero, 56.8% of variance in the leading singular vector), a bias the deep models' training objectives lack.

Held-out calibration null and the identifiability argument. Extending Bolstad–Reinhard quantile matching with two joint-structure modes (covariance preservation, sparsity injection) and evaluating all nine orderings under an honest 160/40-perturbation train/test split, no calibration ordering exceeds the uncalibrated baseline on any model: GEARS vanilla 0.303 vs. best calibrated 0.290; CPA 0.309 vs. 0.289; Geneformer 0.319 vs. 0.287 (full sweep and table in §C.15). Post-hoc magnitude—and joint-structure—calibration thus fails on held-out perturbations: the substitutability gap is not in the magnitude marginal but in the joint structure, and closing it requires a training-time intervention rather than post-hoc rescaling. This null is not an artifact of our particular calibrator; it is forced by an invariance.

Proposition 5.1 (Marginal calibration preserves the conditional-independence graph). Let $\hat{X} \in \mathbb{R}^G$ be a predictor's joint output and let $g = (g_1, \ldots, g_G)$ be a per-gene calibration in which each g_j is a strictly monotone (hence injective) map applied to coordinate j, as in rank-preserving quantile matching. Conditional independence is invariant under coordinate-wise injections: for disjoint i, j and set S,

$$\hat{X}_i \perp\!\!\!\perp \hat{X}_j \mid \hat{X}_S \iff g_i(\hat{X}_i) \perp\!\!\!\perp g_j(\hat{X}_j) \mid g_S(\hat{X}_S).$$

Therefore g preserves the conditional-independence graph of $\hat{X}$ exactly, and in the population limit any constraint-based discovery algorithm—and any algorithm whose target is the (interventional) Markov equivalence class, which is a function of the conditional-independence relations alone—recovers the identical graph from $\hat{X}$ and from $g(\hat{X})$.

Marginal rescaling therefore provably cannot close the gap (proof, corollary, and the finite-sample reading

for the score-based and covariance/precision backends in §C.16): the residual lives in the off-diagonal joint (precision-support) structure, motivating a training-time rather than post-hoc remedy.

STATE non-replication. Arc Institute's 2025 STATE foundation model shows an apparent seed-42 advantage ($F_1 = 0.462$) that does not replicate: the three per-seed values are $[0.462, 0.041, 0.048]$, two of three near 0.04 and well below the ElasticNet mean of 0.418, so the substitutability gap holds for the 2025 foundation model as well (pipeline, framework-version fixes, and the 1,200-cell sampling control in §D.2).

Limitations. The central limitation is the ground truth: Track A (real-data GIES) is internally consistent but does not establish biological correctness, and a multi-seed external check against curated TRRUST is a genuine null ($0.83 \pm 0.70\times$ a matched random-edge floor; §C.13)—at this scale GIES is too cell-sampling-sensitive (technical-duplicate $F_1 = 0.045$) to validate against a sparse curated network (Kernfeld et al., 2024). Geneformer's pretraining includes Replogle 2022 (partial memorization), cross-dataset cells are single-seed, and primary results are at $N = 200$. The full discussion is in §C.17.

6. Conclusion

PerturbCausal is the first benchmark to test whether virtual-cell-model-predicted interventional Perturb-seq, passed through a fixed causal-discovery algorithm, recovers the network that algorithm recovers from real data. Across a primary dataset (Replogle K562, up to 20 seeds), five predictors (ElasticNet, CPA, GEARS, Geneformer, STATE), and four causal-discovery backends (GIES, inspre, PC, NOTEARS), no deep virtual cell model significantly exceeds a sparse linear regression comparator on causal-recovery F1, and none reaches real-data quality.

The result holds across the conditions tested: GEARS and CPA retrained on RPE1 and Norman never exceed the linear comparator under any backend (largest cross-dataset $F_1 = 0.440$, inside its chance band), and STATE's single-seed advantage does not replicate ($F_1 = 0.462, 0.041, 0.048$). Single-seed cross-dataset cells make this consistency across conditions rather than proven invariance.

The failure decomposes into a magnitude marginal (mode collapse and effect-size inflation, two views of one homogenized response, partially correctable post-hoc by Bolstad–Reinhard quantile matching) and a joint-distribution residual in gene-gene covariance and edge orientation that distributional calibration cannot reach. Closing the gap therefore calls for a training-time intervention, not post-hoc rescaling. The benchmark, calibration diagnostic, multi-seed pipeline, and reproducibility infrastructure are released as the open-source PerturbCausal Python package.

References

Ahlmann-Eltze, C. and Huber, W. Deep-learning-based gene perturbation effect prediction does not yet outperform simple linear baselines. Nature Methods, 2025.

Bolstad, B. M., Irizarry, R. A., Astrand, M., and Speed, T. P. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics, 19(2):185–193, 2003.

Brown, B. C., Spence, J. P., and Pritchard, J. K. Large-scale causal discovery using interventional data sheds light on gene network structure in K562 cells. Nature Communications, 16:9628, 2025.

Bunne, C., Roohani, Y., Rosen, Y., et al. How to build the virtual cell with artificial intelligence: priorities and opportunities. Cell, 187:7045–7063, 2024.

Chevalley, M., Roohani, Y., Mehrjou, A., Leskovec, J., and Schwab, P. CausalBench: A large-scale benchmark for network inference from single-cell perturbation data. Communications Biology, 2025.

Cui, H., Wang, C., Maan, H., et al. scGPT: toward building a foundation model for single-cell multi-omics using generative AI. Nature Methods, 21:1470–1480, 2024.

Han, H., Cho, J.-W., Lee, S., et al. TRRUST v2: an expanded reference database of human and mouse transcriptional regulatory interactions. Nucleic Acids Research, 46(D1):D380–D386, 2018.

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

Lotfollahi, M., Klimovskaia Susmelj, A., De Donno, C., et al. Predicting cellular responses to complex perturbations in high-throughput screens. Molecular Systems Biology, 19:e11517, 2023.

Mejia, G. M., Kandari, P., Rai, A., et al. Diversity by design: Addressing mode collapse improves scRNA-seq perturbation modeling on well-calibrated metrics. arXiv preprint arXiv:2506.22641, 2025.

Miller, H. E., Mejia, G. M., Leblanc, F. J. A., et al. Deep learning-based genetic perturbation models do outperform uninformative baselines on well-calibrated metrics. bioRxiv 2025.10.20.683304, 2025.

Reinhard, E., Adhikhmin, M., Gooch, B., and Shirley, P. Color transfer between images. IEEE Computer Graphics and Applications, 21(5):34–41, 2001.

Replogle, J. M., Saunders, R. A., Pogson, A. N., et al. Mapping information-rich genotype-phenotype landscapes with genome-scale Perturb-seq. Cell, 185(14):2559–2575, 2022.

Peidli, S., Green, T. D., Shen, C., et al. scPerturb: harmonized single-cell perturbation data. Nature Methods, 21:531–540, 2024.

Roohani, Y., Huang, K., and Leskovec, J. Predicting transcriptional outcomes of novel multigene perturbations with GEARS. Nature Biotechnology, 42:927–935, 2024.

Roohani, Y., Lerman, J., and Leskovec, J. Virtual cell challenge: Toward a Turing test for the virtual cell. Cell, 188:3370–3382, 2025.

Theodoris, C. V., Xiao, L., Chopra, A., et al. Transfer learning enables predictions in network biology. Nature, 618:616–624, 2023.

Callahan, T., Bertin, P., Tejada-Lapuerta, A., et al. Virtual cells as causal world models: A perspective on evaluation. In NeurIPS 2025 AI4D3 Workshop, 2025. URL: https://ai4d3.github.io/2025/papers/25_Virtual_Cells_as_Causal_Wor.pdf.

Wu, M., Bao, Y., Xu, K., Patel, A., Bai, K., Schiebinger, G., and Lopez, R. PerturBench: Benchmarking machine learning models for cellular perturbation analysis. arXiv preprint arXiv:2408.10609, 2024.

Wu, M., Tu, R., Lopez, R., Schiebinger, G., and Bai, K. Identifying biological perturbation targets through causal differential networks. In International Conference on Machine Learning, 2025.

Zheng, X., Aragam, B., Ravikumar, P., and Xing, E. P. DAGs with NO TEARS: Continuous optimization for structure learning. In NeurIPS, 2018.

Arc Institute. STATE: Multi-scale AI model for cellular perturbation effects. GitHub: https://github.com/ArcInstitute/state; HuggingFace: https://huggingface.co/arcinstitute/SE-600M, 2025.

Adduri, A. et al. Predicting cellular responses to perturbation across diverse contexts with STATE. bioRxiv, 2025.

Wong, D. R. et al. Simple controls exceed best deep learning algorithms and reveal foundation model effectiveness for predicting genetic perturbations. Bioinformatics, 2025.

Csendes, G. et al. Benchmarking foundation cell models for post-perturbation RNA-seq prediction. BMC Genomics, 2025.

Wei, Z. et al. Benchmarking algorithms for generalizable single-cell perturbation response prediction. Nature Methods, 2025.

Viñas Torné, R. et al. Systema: a framework for evaluating genetic perturbation response prediction beyond systematic variation. Nature Biotechnology, 2025.

Bendidi, I. et al. Benchmarking transcriptomics foundation models for perturbation analysis: one PCA still rules them all. arXiv, 2024.

Wenteler, A. et al. PertEval-scFM: Benchmarking single-cell foundation models for perturbation effect prediction. bioRxiv, 2025.

Zhu, D. et al. Gradient-aware modeling advances AI-driven prediction of genetic perturbation effects. bioRxiv, 2025.

Li, L. et al. A systematic comparison of single-cell perturbation response prediction models. bioRxiv, 2025.

Cao, Z.-J. et al. Causality-oriented regulatory inference and rational transcriptomic reprogramming design with CASCADE. bioRxiv, 2025.

Park, K. et al. Confounder-robust causal discovery and inference in Perturb-seq using proxy and instrumental variables. 2026.

Pratapa, A. et al. Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data (BEELINE). Nature Methods, 2019.

Kernfeld, E. et al. Transcriptome data are insufficient to control false discoveries in regulatory network inference. Cell Systems, 2024.

Qiu, X. et al. Inferring causal gene regulatory networks from coupled single-cell expression dynamics using Scribe. Cell Systems, 2020.

Spirtes, P. and Glymour, C. Causation, Prediction, and Search, 2nd edition. MIT Press, 2000.

Zou, H. and Hastie, T. Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B, 67(2):301–320, 2005.

Fernandez-de-Retana, M. et al. When does gene regulatory network inference break? A controlled diagnostic study of causal and correlational methods on single-cell data. arXiv preprint arXiv:2605.04930, 2026.

Zinati, Y., Takiddeen, A., and Emad, A. GRouNdGAN: GRN-guided simulation of single-cell RNA-seq data using causal generative adversarial networks. Nature Communications, 15:4055, 2024.

Norman, T. M., Horlbeck, M. A., Replogle, J. M., Ge, A. Y., Xu, A., Jost, M., Gilbert, L. A., and Weissman, J. S. Exploring genetic interaction manifolds constructed from rich single-cell phenotypes. Science, 365(6455):786–793, 2019. doi: 10.1126/science.aax4438.

A. Extended related work

Causal discovery on Perturb-seq. GIES (Hauser & Bühlmann, 2012) is the canonical interventional causal discovery algorithm and the primary backend in CausalBench (Chevalley et al., 2025). inspre (Brown et al., 2025) operates on the average causal effect matrix and scales to thousands of nodes. Both are used here as parallel backends.

Perturbation prediction benchmarks. A growing 2024–2025 literature converges on a single conclusion: deep learning does not yet beat simple baselines on per-gene reconstruction (Ahlmann-Eltze & Huber, 2025; Wong et al., 2025; Csendes et al., 2025; Wei et al., 2025; Viñas Torné et al., 2025; Bendidi et al., 2024; Wenteler et al., 2025; Zhu et al., 2025; Li et al., 2025; Mejia et al., 2025). PerturBench (Wu et al., 2024) and the Virtual Cell Challenge (Roohani et al., 2025) provide the field's standard evaluation harnesses; Viñas Torné et al. (2025) introduce "systematic variation" as a confounder of Δ-based metrics and Wei et al. (2025) extend the comparison to 27 methods on 29 datasets. None of these evaluate downstream causal recovery from VCM predictions.

Causal discovery from Perturb-seq. A parallel literature builds causal-discovery methods directly on real Perturb-seq data: GIES (Hauser & Bühlmann, 2012), inspre (Brown et al., 2025), BEELINE (Pratapa et al., 2019) as a benchmark of GRN inference algorithms, Scribe (Qiu et al., 2020), CASCADE (Cao et al., 2025) for causality-oriented reprogramming design, and confounder-robust IV-based methods (Park et al., 2026). Kernfeld et al. (2024) argue transcriptome data alone cannot control false discoveries in causal network inference. PerturbCausal is downstream of all of these: it asks whether predicted perturbation data can substitute for the real data those methods consume. Concurrently, Fernandez-de-Retana et al. (2026) ask the adjacent question of when causal GRN inference fails to beat correlational baselines by isolating seven biological pathologies on controlled data; that work varies the data-generating pathology under a fixed comparison, whereas we vary the data source (predicted vs. real) under a fixed algorithm—orthogonal axes that compose.

GRN-conditioned simulators. A separate line generates synthetic single-cell data from a known GRN to benchmark inference algorithms: SERGIO and, most directly, GRouNdGAN (Zinati et al., 2024) and its 2025 GRouNdGAN-Toolkit, a causal GAN that imposes a user-specified network and supports in-silico perturbation. There the ground-truth network is known by construction and the simulated data is the controlled variable. PerturbCausal inverts this: the predictors are learned VCMs trained to fit real expression (not to encode a GRN), and the question is whether their predictions retain recoverable causal structure when scored against real-data recovery, with no injected ground-truth network.

Causal evaluation of VCMs. Callahan et al. (2025) proposed a causal-evaluation taxonomy without implementation. Wu et al. (2025) used causal discovery as a featurizer for perturbation target prediction—orthogonal direction to ours.

Calibration in perturbation prediction. Mejia et al. (2025) propose a weighted training loss that requires retraining. Miller et al. (2025) propose metric calibration that does not change model outputs and argue that conventional metrics underestimate deep-learning performance under their dynamic-range-fraction (DRF) calibration framework. Our quantile-matching calibration is post-hoc, model-agnostic, and applied directly to VCM outputs without retraining. We note that causal F1 against a fixed real-data GIES ground truth is rank-based on edge sets and is therefore not subject to the same metric-calibration concerns Miller et al. (2025) raise for delta-correlation and per-gene MSE. Our finding that ElasticNet matches or exceeds deep models on causal F1 is therefore complementary to, rather than contradicted by, the DRF results: prediction calibration and causal-recovery calibration are different evaluation axes.

B. Extended methods

Datasets. We use three Perturb-seq datasets, chosen to test in-distribution, cross-context, and cross-paradigm generalization. Replogle K562 (Replogle et al., 2022) (GEO GSE221321) is the primary in-distribution dataset: 310,385 transcriptomes, 1,861 CRISPRi targets with $\geq$ 10 cells, 10,691 non-targeting controls, 8,563 genes. Replogle RPE1 (same study and modality, different cell context) is the cross-context test: 247,914 cells, 2,069

eligible targets; the harmonized release is obtained from scPerturb (Peidli et al., 2024). Norman 2019 (Norman et al., 2019) (GEO GSE133344) is the cross-paradigm test, a CRISPR activation screen in K562: 111,445 cells under 237 perturbations, restricted to the 102 single-gene targets after dropping combinatorial conditions so the on-target-edge metric matches the CRISPRi datasets. The primary K562 evaluation uses a 200-gene subset of the eligible set; 200 is the largest set that fills exactly ten 20-gene GIES partitions within the compute budget, and the scale sweep $N \in \{50, 75, 100, 125, 150, 175, 200\}$ (§C.5) shows rankings are stable across subset size. Gene subsets are released as fixed sorted lists; RPE1 uses 196 and Norman 102 eligible targets on 8,749- and 8,028-gene HVG backgrounds respectively.

Overlay sampling from predicted means to backend inputs. Independent negative-binomial sampling from predicted means removes gene-gene covariance and collapses GIES edge recovery below 2% of the real-data ceiling. Overlay sampling preserves covariance: for each perturbation, control cells are bootstrap-sampled from a 200-cell control matrix (matching the CausalBench protocol (Chevalley et al., 2025)), scaled element-wise by the predicted fold change with a small pseudocount and clip, and given Poisson count noise; a sensitivity sweep over clip range, pseudocount, and noise model leaves rankings unchanged. On K562, GIES, PC, and NOTEARS consume these overlay cell matrices. For the cross-dataset evaluation, PC, NOTEARS, and inspre instead consume the square average causal effect (ACE) matrix $\mathrm{ACE}_{ij} = \mathbb{E}[X_j \mid \mathrm{do}(X_i)] - \mathbb{E}[X_j \mid \mathrm{ctrl}]$ restricted to the canonical perturbed-gene set, built from per-perturbation predicted means over perturbations with ≥ 5 cells; the node set is then identical across predictors and the only varying quantity is the predicted intervention effect.

Statistical testing. K562 headline numbers are across-seed means with 95% t-CIs over $n \in \{3, 5, 20\}$ seeds; each seed re-randomizes gene-subset selection, control subsampling, and partition order. Pairwise differences use paired permutation tests at $p < 0.05$. Equivalence between the deep models and the linear comparator is assessed by two one-sided tests (TOST) with margin $\Delta = 0.045$ (the technical-duplicate noise floor), concluding equivalence when the 90% CI of the paired difference lies within $\pm\Delta$. Cross-dataset cells are single-seed point estimates reported against the per-backend permutation band of §3.8; bootstrap edge-set CIs are reported for the GIES cross-dataset cells, and per-backend chance bands for the PC/NOTEARS/inspre cross-dataset cells. Because the cross-dataset cells are single-seed, the bands capture edge-set and permutation variance but not gene-subset or control-subsample variance; point estimates near a band edge are read as suggestive rather than significant.

Effect-size calibration. Post-hoc quantile-matching calibration maps the predicted $|\log_2 \mathrm{FC}|$ distribution onto the real-data distribution by rank-preserving interpolation (Bolstad et al., 2003; Reinhard et al., 2001), with two joint-structure extensions (covariance preservation, sparsity injection) evaluated under a 160/40-perturbation train/test split; the full algorithm and six-ordering ablation appear in §C.15. Calibration is a transformation of the predictor's output, not of the evaluation pipeline.

Secondary random-edge floor and full metric set. The secondary, density-matched sanity reference is the analytic expected F_1 of a predictor emitting the same number of edges drawn uniformly at random over the $G(G-1)$ directed pairs,

$$\mathbb{E}[F_1] = \frac{2\, n_{\mathrm{pred}}\, n_{\mathrm{gt}}}{(n_{\mathrm{pred}} + n_{\mathrm{gt}})\, G(G-1)},$$

exact because F_1 is linear in true positives at fixed edge counts. The two references can disagree by large factors: the permutation null inherits the backend's edge density, so on backends that produce dense real-data graphs (NOTEARS, inspre) the null itself scores a high F_1, and a predictor can be far above the random-edge floor yet inside or below the permutation band. The permutation null is therefore the reference used for all at-chance / above-chance / below-chance verdicts; the random-edge floor is reported only to confirm a predictor is not emitting a trivially small or large edge set. Beyond the primary edge-set F_1, secondary metrics are precision, recall, top-k AUPRC, P@100, structural Hamming distance, mean per-perturbation Wasserstein distance, and false omission rate (Mann–Whitney U with global Benjamini–Hochberg correction); multi-hop credit at $k \in \{2, 3\}$ counts a predicted edge correct if a ground-truth path of length $\leq k$ exists, and a real-data self-consistency reference F_1^{real} (between two GIES recoveries on disjoint random halves of real K562 cells, mean of three splits) anchors what the rankings are and are not measuring.

Implementation. The sparse linear comparator uses $\alpha = 0.1$, ℓ_1 ratio 0.5. Overlay sampling draws 100 cells per perturbation, scales by $(\mu_{\text{pred}} + 0.01)/(\mu_{\text{ctrl}} + 0.01)$ clipped to $[0.01, 10]$, and applies Poisson noise. PC uses Fisher-Z at $\alpha = 0.05$; NOTEARS uses $\ell_1 = 0.05$, ℓ_2 loss, weight threshold 0.10, ≤ 100 iterations; inspre uses 200 iterations and train fraction 0.8. The local software stack is Python 3.9 (anndata 0.10, scanpy 1.10, scikit-learn 1.x, networkx 3.3, causal-learn 0.1.4, the notears and gies reference packages) with R 4.5.2 for inspre (called as a subprocess); Modal training and large-scale inference use Python 3.11 (full version table in §E.1). Local numbers are computed on Apple M4 hardware. Backend calls are disk-cached and resumable; all code, configurations, fixed gene lists, and edge caches are released under MIT license. Backend calls are disk-cached and resumable; all code, configurations, fixed gene lists, and edge caches are released under MIT license.

C. Additional results

C.1. Reconciliation: mode collapse and inflation are two views of one failure

The inflation observation appears to contradict the recent mode-collapse literature (Mejia et al., 2025; Ahlmann-Eltze & Huber, 2025), which reports that scRNA-seq perturbation models compress per-perturbation variance toward the dataset mean. We test the two against each other on identical predictions and find they coexist as two views of the same underlying failure. Per-perturbation Δ-variance ratios (predicted shift variance over real shift variance) are below 1 for all three deep models on the 200-gene scope: median 0.461 (GEARS), 0.432 (CPA), 0.008 (Geneformer); GEARS and CPA compress real per-perturbation variance by approximately half, Geneformer by two orders of magnitude. Inter-perturbation cosine similarity of shift vectors is 0.084 in real data and 0.845 in CPA predictions—CPA produces nearly the same response vector for every perturbation. When per-perturbation effects are pooled across all (perturbation, gene) entries at the full transcriptome scope, the same homogenization manifests as the 23-fold inflation: dense responses at off-target genes accumulate. Mode collapse and inflation are therefore not contradictory; they are the per-perturbation and pooled views, respectively, of a model class that predicts a single dense response pattern across all perturbations.

C.2. Technical-duplicate noise floor

To anchor model F1 values against the intrinsic technical-duplicate ceiling of the dataset (Mejia et al., 2025; Miller et al., 2025), we compute F1 between two GIES recoveries on disjoint random halves of real K562 cells across three random splits. The mean F1 between halves is 0.045 ± 0.003, with both halves recovering 1,080–1,110 edges and only $\sim$50 in common. Real-cell sampling variance therefore dominates GIES at N=200 with the cells-per-perturbation budgets used here, and all evaluated methods (ElasticNet 0.425, deep models 0.389–0.432) operate roughly an order of magnitude above this raw cell-split floor. We report this technical-duplicate F1 as a row in Table 2 alongside the random-bootstrap floor (F1 $\approx$ 0.04) for clarity on what the ranking is and is not measuring: methods are ranked against the GIES-on-real ground truth derived from overlay-resampled cells, not against a noiseless oracle.

C.3. In-domain calibration head-to-head

The quantile-matching calibration corrects the magnitude distribution exactly: post-calibration $|\log_2 \text{FC}| > 0.5$ fraction is 11.2% (GEARS), 11.4% (CPA), 11.4% (Geneformer), within 0.2 pts of the real-data 200-gene reference 11.4% (Figure 5, top). In-domain at seed 42, downstream causal F1 improves modestly for two of three models: CPA 0.389 $\rightarrow$ 0.405 ($\Delta = +0.016$), GEARS 0.406 $\rightarrow$ 0.413 ($\Delta = +0.008$), Geneformer 0.432 $\rightarrow$ 0.423 ($\Delta = -0.009$) (Figure 5, bottom; Table 4). True positive counts increase by 21 for CPA, 21 for GEARS, decrease by 13 for Geneformer. Reconstruction MSE increases modestly: CPA +6%, GEARS +33%, Geneformer +8%. No calibrated deep model surpasses ElasticNet (F1 = 0.425). Why two of three help. CPA and GEARS gain 1–2 F1 points after quantile matching erases their magnitude inflation. Geneformer (best vanilla F1 at 0.432) loses 0.9 F1 under the same calibration, plausibly because its foundation-model pretraining yields a magnitude distribution that is already approximately correct; uniform global matching erases this implicit calibration without replacing it. This argues for a learned per-model shrinkage rather than uniform application.

Table 4. Vanilla vs. calibrated head-to-head on K562 at N=200, identical seed and partition configuration. ΔF1 is calibrated minus vanilla. MSE Δ is reconstruction MSE percentage change.

Model	Vanilla F1	Cal. F1	ΔF1	MSE Δ
CPA	0.3887	0.4051	+0.016	+6%
GEARS	0.4055	0.4133	+0.008	+33%
Geneformer	0.4324	0.4232	−0.009	+8%

C.4. Algorithmic robustness and inspre lambda-path discussion

Algorithmic robustness across backends and datasets. Two questions are separable: whether VCM predictions can substitute for real data (the substitutability claim) and whether they recover anything beyond chance (the chance claim). The cross-dataset backend sweep (§4.5, Table 3) answers the first uniformly and the second with one honest exception. Running PC (constraint-based; Spirtes & Glymour 2000), NOTEARS (gradient-based; Zheng et al. 2018), and inspre (sparse-inverse; Brown et al. 2025) on RPE1 and Norman in addition to GIES, no deep predictor exceeds the sparse linear comparator in any cell where the comparator is defined (ten of twelve deep-model comparisons; ElasticNet under PC on RPE1 did not converge), and the largest deep-model F_1 anywhere (0.440, CPA under inspre on RPE1) is inside that backend's chance band—so the substitutability claim is backend-invariant. The chance claim holds for GIES, NOTEARS, and inspre-on-RPE1, where deep models are at or below their permutation floor. Two above-floor points remain: PC on RPE1, where both deep models sit fourfold above a low floor at absolute $F_1 = 0.078$, and GEARS under inspre on Norman, 0.344 against a 0.329 band edge—a single-seed margin. Because PC is observational and discards intervention labels, the RPE1 signal is best read as preserved baseline co-expression rather than recovered interventional structure—a mechanism the score-based and interventional backends, which exploit the do-operator, do not reward. Reporting this exception rather than averaging it away is the point: the inspre cells show that a high raw F_1 (up to 0.586) can sit entirely inside a high chance band, so a single-backend reading in either direction would mislead.

What does the inspre lambda-path result mean? The full-path validation tempers the binary "zero edges from CPA and Geneformer" result into a continuous gradient: CPA vanilla still fails to traverse any valid λ, but Geneformer recovers up to 602 (vanilla) / 1,315 (calibrated) edges at non-cross-validated λ settings. Calibration expands the detectable-signal range for both even when it hurts GIES F1 for Geneformer, indicating GIES (top-edge accuracy) and inspre (total recoverable signal) probe different prediction properties; calibration moves the latter further (Appendix D.2).

C.5. Scale sweep across N=50–200

The deep model achieving the highest precision rotates across gene-set sizes: Geneformer leads at N=150 (precision 0.394), GEARS at N=175 (0.401), and ElasticNet at N=200 (0.378). ElasticNet is in the top two methods at every scale tested. No deep model dominates consistently. Single-scale evaluation of these methods would yield three different headline conclusions depending on chosen N. Extending the inspre λ-path probe to N=500 surfaces a regime change: with matched 500×500 ACE matrices, inspre recovers 0 edges along the entire λ path on real K562 (solver returns no finite λ values; wall time 3.2 hours) but 3,643 edges along the path on the ElasticNet ACE (wall time 1.9 hours). The reading that survives both directions is that inspre's sparsity-of-inverse assumption is sensitive to the noise floor of the input matrix: at N=500 the raw real-data ACE carries enough per-perturbation sampling noise that no setting of λ produces a stable sparse inverse (the same failure mode previously observed only for vanilla CPA at N=200), whereas ElasticNet's structural shrinkage smooths the ACE into a regime inspre can decompose. This N=500 result is not directly interpretable as a substitutability check but is informative methodologically: at scale the real-data ACE itself can become the limiting factor.

C.6. Reversed-direction failure mode

Three pieces of evidence point to a single conclusion: deep VCM predictions miss the directionality of causal regulation more than they miss the regulatory pairs themselves. First, multi-hop credit analysis (a predicted edge is correct at hop-k if a path of length $\leq k$ exists in the ground-truth graph) increases precision by 9.2 pts (Geneformer) to 11.8 pts (GEARS) at k=2 across all conditions. Second, the SHD breakdown across all four

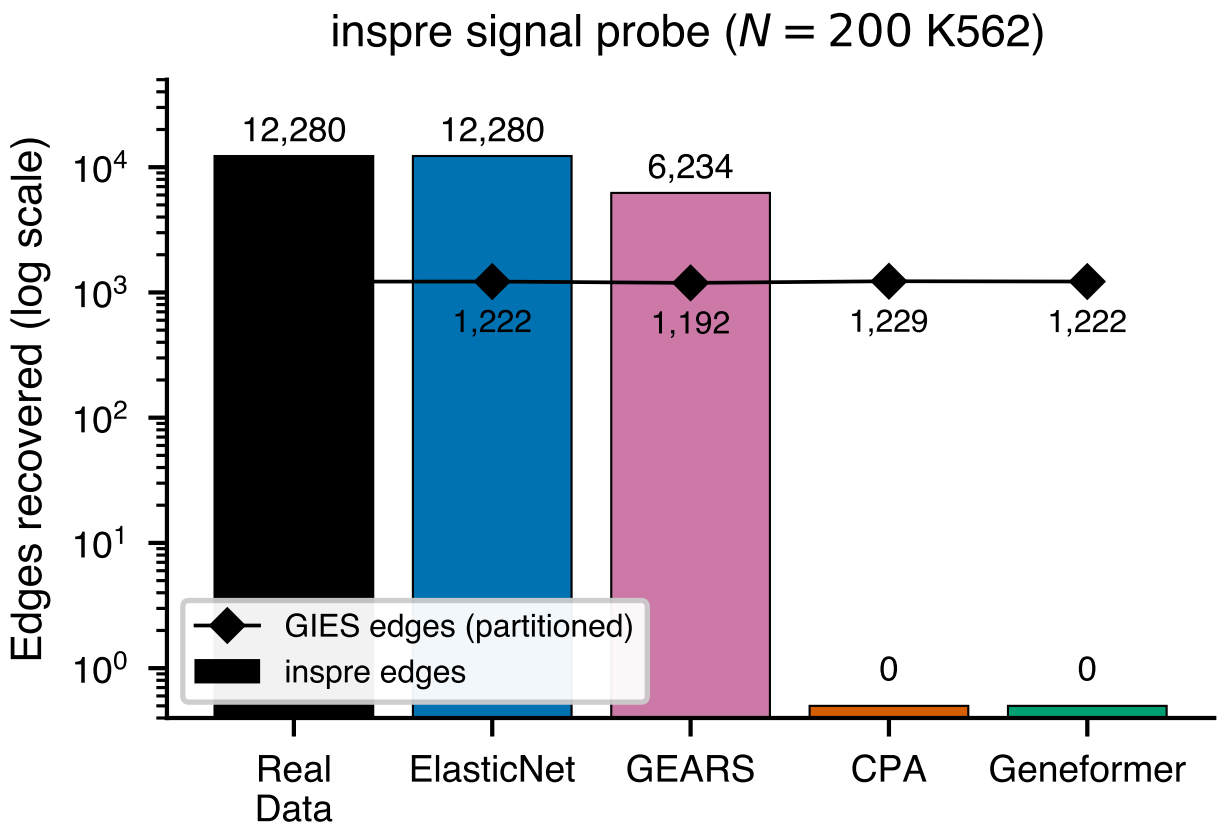

Figure 4. Inspre signal probe across data sources at the test-error-optimal regularization parameter. Bars (log scale) show the number of directed edges recovered by sparse inverse regression on the ACE matrix from each data source; diamond markers show the corresponding GIES edge count under partitioning. Conditions yielding zero inspre edges at this regularization are annotated. The full-λ-path validation reported in Section 4 shows that vanilla Geneformer and calibrated CPA recover edges at smaller-than-cross-validated regularizations (Geneformer up to 602; calibrated CPA up to 419), so the zero-edges result is binary at the cross-validated λ but continuous across the path. CPA vanilla is the only condition where the inspre solver fails to traverse a valid λ path, supporting the strongest signal-failure interpretation for that model. $n = 200$ perturbations $\times$ 200 genes per ACE matrix.

deep models shows 27–30% of false-positive predictions are gene pairs connected in the ground-truth graph but with direction reversed (the reversed-edge fraction; multi-seed JSON in supplementary). Third, undirected F1 (treating both predicted and ground-truth edges as unordered pairs) lifts every method's score by approximately 0.06–0.09 F1 over directed F1, recovering most but not all of the reversed-edge gap. We note that GIES emits CPDAGs (partially directed graphs) and that the directed F1 reported throughout is computed after canonicalizing undirected CPDAG arrows by gene-index ordering; the undirected F1 sanity check confirms that the headline tied-deep-model finding is not an artifact of this canonicalization (the tie is observable on the undirected representation as well).

C.7. Inspre full lambda-path validation

To address whether the zero inspre result (§4.3) is a binary threshold artifact or absolute, we trace recovered edge count across the full λ path of inspre's coordinate-descent solver for each condition. GEARS recovers up to 6,634 edges across the path (smallest λ with any edge: 5.78). CPA's vanilla path yields no finite λ values (the inspre solver fails on its ACE matrix), supporting the strongest reading: CPA's predicted ACE matrix has signal-to-noise too low for sparse inverse regression to even traverse a valid regularization path. Geneformer

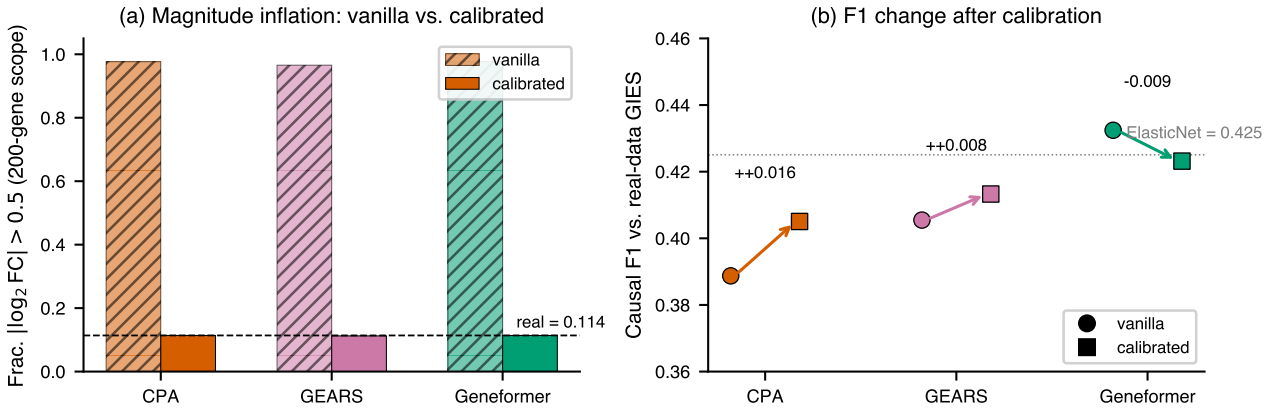

Figure 5. Effect-size calibration corrects magnitude distribution exactly but only partially recovers causal F1. (a) Fraction of $|\log_2 \mathrm{FC}| > 0.5$ before (hatched) and after (solid) quantile-matching calibration for each deep model on the K562 200-gene scope; the dashed horizontal line marks the real-data 200-gene reference. (b) Causal F1 against the real-data GIES ground truth before (circles) and after (squares) calibration on the identical pipeline (seed 42, partition size 20); arrows connect paired vanilla and calibrated values. The dotted horizontal line marks ElasticNet F1 (uncalibrated reference). $n = 1$ seed $\times$ 200 perturbations $\times$ 200 genes.

N (genes)	GT edges	ElasticNet	GEARS	CPA	Geneformer
50	169	0.345	0.280	0.302	0.292
75	309	0.309	0.351	0.344	0.316
100	509	0.344	0.339	0.335	0.348
125	646	0.356	0.360	0.326	0.349
150	821	0.368	0.310	0.347	0.394
175	1,002	0.369	0.401	0.378	0.340
200	1,217	0.378	0.392	0.385	0.373

Table 5. Scale-sweep precision against real-data GIES ground truth at seven gene-set sizes. Bold values indicate top performance per row.

recovers up to 602 edges (smallest λ with any edge: 0.67), so the zero-edges result for vanilla Geneformer at the test-error-optimal λ does not generalize to the entire path. Quantile calibration substantially expands the lambda range over which edges are detectable: calibrated CPA recovers up to 419 edges (smallest λ: 0.20) where vanilla CPA recovered none, and calibrated Geneformer recovers up to 1,315 edges (smallest λ: 0.014) where vanilla recovered 602. The GIES-side calibration result (Geneformer F1 -0.009) and the inspre-side calibration result (Geneformer $+713$ max edges along path) are not contradictory: GIES partitioning forces equal-cardinality outputs, so a Geneformer prediction set that is more diffuse but contains more inspre-detectable signal can lose F1 against a fixed-edge-count GIES baseline while gaining recoverable signal in inspre.

C.8. Multi-hop credit at $k = 1, 2, 3$

Method	P@k=1	P@k=2	P@k=3	Boost (k=2)	Boost (k=3)
ElasticNet	0.425	0.540	0.543	+11.5	+11.9
GEARS	0.410	0.529	0.533	+11.8	+12.2
CPA	0.387	0.500	0.508	+11.3	+12.0
Geneformer	0.432	0.525	0.529	+9.2	+9.7
Overlay Real	0.423	0.533	0.539	+11.0	+11.6

Table 6. Multi-hop credit precision at hop-k over the N=200 K562 evaluation. Boost columns show percentage-point increase from k=1.

C.9. RPE1 cross-context replication and structural comparison

The structural properties of RPE1 perturbation responses differ from K562: leading singular variance 34.5% (vs. 56.8%), fraction near-zero 15.4% (vs. 38.6%), fraction large 17.3% (vs. 2.7%). We retrained both GEARS and CPA from scratch on RPE1 (8,749 genes, 196 perturbation-eligible targets) and pushed their predictions through the identical GIES pipeline. Against the real-data RPE1 ground truth (1,137 edges, seed 42), the single-seed GIES point estimates are GEARS F1 = 0.414, CPA = 0.450, ElasticNet = 0.437, and a random-edge floor of 0.436; CPA edges past the linear comparator and the floor by ≤ 0.014 while GEARS sits below both, so neither deep model separates cleanly from the linear comparator or chance, and the perfect-real-means overlay upper bound itself reaches only 0.418 here (consistent with the metric saturation noted at N=200). The substitutability gap therefore persists under a change of cellular context: two architecturally distinct deep VCMs retrained on RPE1 recover no more of the real causal graph than a sparse linear baseline or chance. This RPE1 estimate is single-seed and supports a qualitative, not a significance, claim (§5); the earlier bootstrap-resampled edge-set intervals are not reported because resampling-then-deduplicating the predicted edge list biases the F1 estimate downward.

Property	K562	RPE1
First singular vector variance	56.8%	34.5%
Top-5 singular variance	69.8%	68.5%
Frac near-zero effects	38.6%	15.4%
Frac large effects ($\lvert \log_2 \mathrm{FC} \rvert > 0.5$)	2.7%	17.3%
GIES edges (real data)	1,220	810

Table 7. Structural properties of real perturbation-response matrices for K562 and RPE1 at N=200. The RPE1 GIES edge count (810) is for this 200-gene structural panel; the main-text cross-context evaluation (§4) runs on the 196 perturbation-eligible RPE1 targets and its real-data GIES ground truth has 1,137 edges. K562 (1,220) is the seed-42 evaluation ground truth.

C.10. Multi-seed bootstrap confidence

Across five seeds (42, 123, 456, 789, 1024) the four predictive methods sit within overlapping intervals on every metric (Table 8). The GRNBoost2 random-baseline floor confirms that all four predictive methods recover real signal substantially above chance, while none of the deep models significantly separates from ElasticNet at this seed budget.

Method	F1	Precision	Recall	AUPRC
ElasticNet	0.426 ± 0.016	0.420 ± 0.014	0.431 ± 0.019	0.425 ± 0.016
CPA	0.412 ± 0.016	0.407 ± 0.013	0.417 ± 0.019	0.400 ± 0.022
GEARS	0.406 ± 0.015	0.404 ± 0.012	0.408 ± 0.018	0.419 ± 0.012
Geneformer	0.425 ± 0.011	0.422 ± 0.010	0.429 ± 0.013	0.422 ± 0.013
Overlay-real (oracle)	0.422 ± 0.009	0.420 ± 0.010	0.425 ± 0.009	0.420 ± 0.013
GRNBoost2 (random)	0.098 ± 0.004	0.074 ± 0.003	0.147 ± 0.005	0.077 ± 0.005

Table 8. Multi-seed mean and standard deviation across $n = 5$ seeds at N=200 on K562 against real-data GIES ground truth (1,220 edges per seed). Overlay-real is the same overlay-sampling pipeline applied to real perturbed cell predictions and serves as a noise-floor for the overlay step itself. GRNBoost2 is a structurally-distant baseline (gradient-boosted regression) included as a near-random reference.

C.11. Full PC and NOTEARS robustness tables

The main-text PC and NOTEARS Discussion paragraph cites summary statistics. Tables 9 and 10 report the full per-condition F1, precision, recall, and edge count against each backend's own real-data ground truth. Both backends produce sparser graphs than GIES at N=200 (PC: 682 edges, NOTEARS: 328 edges, vs. GIES: 1,220), and per-model F1s are uniformly small in absolute terms. Calibration zeros out NOTEARS recovery for all models because quantile matching pushes ACE entries below the w_{thresh}=0.1 weight threshold; under PC the calibrated and vanilla scores are within ± 0.01 for all three models.

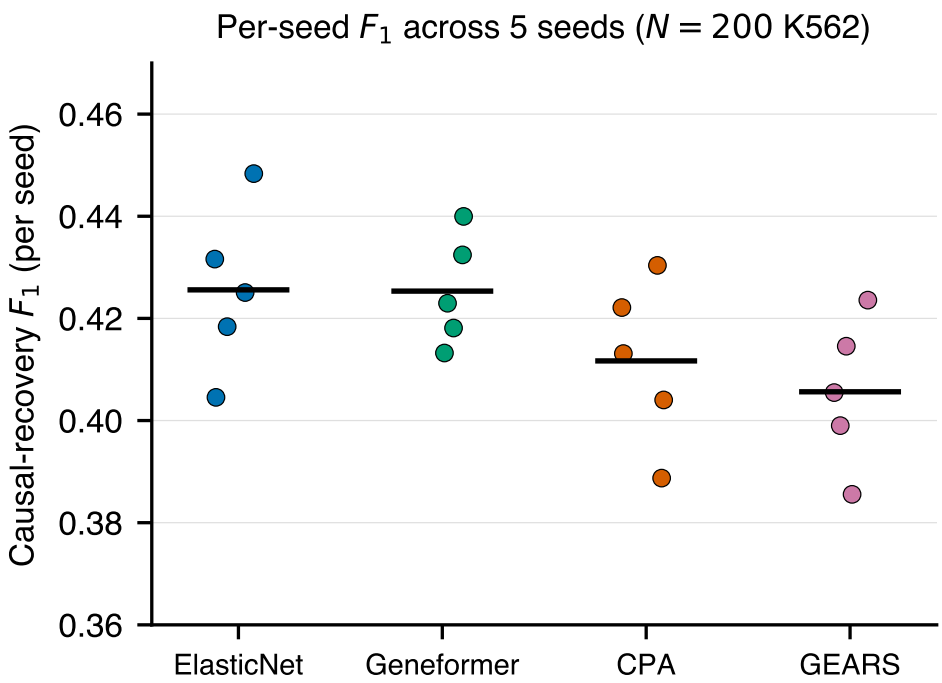

Figure 6. Per-seed causal-recovery F_1 for the four predictive methods across the five K562 seeds (one marker per seed; horizontal bar is the across-seed mean). The clouds overlap and the means fall within a 0.02 F1 band, visualizing the TOST equivalence (§4): the deep models are statistically equivalent to ElasticNet at the $\Delta = 0.045$ noise-floor margin. Wong colorblind-safe palette.

C.12. Six-variant CPA architectural ablation

To localize the substitutability gap to a specific architectural component, six CPA variants were trained, each modifying one architectural axis relative to the NB-loss baseline. Per-variant causal F1 is reported in Table 11. Spread across all six variants is 0.387–0.418 ($\Delta = 0.031$ F1), within the multi-seed CI of vanilla CPA. Removing the adversarial covariate disentangler is the single largest deletion (loses 0.025 F1, the largest drop), but it does not reproduce the gap to real-data GIES recovery. The collective implication is that the substitutability gap is not localizable to any one architectural component of CPA.

C.13. External validation of the ground truth against TRRUST

The essential-gene N=200 evaluation subset contains effectively no TRRUST (Han et al., 2018) edges, so the main-text Track B is restricted to rank concordance. To test the biological content of the GIES-on-real ground truth directly, we built a separate TF-rich subset by selecting TRRUST transcription factors that are perturbation-eligible in K562 ($\geq$ 50 cells) together with their measurable targets, capped at 70 genes and pruned to genes participating in at least one within-subset curated edge ($\sim$100 directed TRRUST edges, 18 perturbed targets). GIES was run on real K562 cells (200 non-targeting controls + up to 150 cells per perturbation, partition size 20) across five seeds, and the recovered graph was scored against the curated edges with a per-seed random-edge null matched to the recovered edge count (200 draws each).

The multi-seed result (Table 12) is a genuine null: the GIES-on-real ground truth recovers curated TRRUST edges at $0.83 \pm 0.70\times$ chance, not robustly above it, and the per-seed lift ranges from 0 to $1.84\times$. The instability mirrors the technical-duplicate floor (§C.2, $F_1 = 0.045$ between real-cell halves): at this gene-set scale GIES recovery is dominated by cell-sampling variance, so a single seed can appear to validate biology while the multi-seed estimate does not. The honest reading is that the Track-A ground truth cannot be biologically validated against a sparse curated network at this scale—the circularity is a real, open limitation—and a larger, denser

Condition	Edges	F1	Precision	Recall
Real (PC ground truth)	682	—	—	—
GEARS-vanilla	556	0.0549	0.0612	0.0499
GEARS-calibrated	652	0.0480	0.0491	0.0469
CPA-vanilla	268	0.0800	0.1418	0.0557
CPA-calibrated	294	0.0779	0.1293	0.0557
Geneformer-vanilla	142	0.0097	0.0282	0.0059
Geneformer-calibrated	240	0.0043	0.0083	0.0029

Table 9. PC algorithm (Fisher-Z conditional independence at α=0.05) per-condition results on K562 ACE matrices at N=200, seed 42.

Condition	Edges	F1	Precision	Recall
Real (NOTEARS ground truth)	328	—	—	—
GEARS-vanilla	24	0.0511	0.3750	0.0274
GEARS-calibrated	0	0.0000	0.0000	0.0000
CPA-vanilla	18	0.0405	0.3889	0.0213
CPA-calibrated	0	0.0000	0.0000	0.0000
Geneformer-vanilla	7	0.0119	0.2857	0.0061
Geneformer-calibrated	0	0.0000	0.0000	0.0000

Table 10. NOTEARS (continuous-optimization with acyclicity constraint) per-condition results on K562 at N=200, seed 42, λ=0.05, weight threshold w_{thresh}=0.1.

external standard on a more stable recovery regime is required to close it. This null also illustrates why every biological-validation claim in this benchmark is reported multi-seed.

C.14. Learned joint transform vs. marginal calibration (covariance-regularizer proxy)

Mechanism: shared signature and the alt-loss dissociation. All three deep VCMs produce similar mode-collapse signatures: per-perturbation Δ-variance ratios below 1 and pooled $|\log_2 \text{FC}|$ inflation at the full-transcriptome scope. Their inter-perturbation cosine similarities diverge in form—CPA collapses to a near-identical response across perturbations (cosine 0.845 vs. the real-data 0.084), whereas GEARS (0.075) and Geneformer (0.059) instead under-correlate relative to real—but all three share the per-perturbation variance compression. GEARS, CPA, and Geneformer differ in architecture, training objective formulation, and parameter count by orders of magnitude—so the shared signature is suggestive of an objective-driven rather than architecture-driven cause. To test this, we retrained CPA with Gaussian (textbook per-gene MSE) reconstruction loss in place of its default Negative Binomial loss, holding all other hyperparameters constant; if the failure were purely an artifact of the NB reconstruction term, the alternative loss would reduce mode collapse and improve causal recovery. The result is striking: the Gaussian-loss variant produces more per-perturbation variance compression than vanilla CPA (median Δ-variance ratio 0.43 $\rightarrow$ 0.18) and only marginally less inter-perturbation similarity (cosine 0.84 $\rightarrow$ 0.80), yet higher causal F1 against the GIES ground truth (0.389 $\rightarrow$ 0.414 at seed 42). Per-perturbation magnitude and causal recovery are dissociated: a model can produce more aggressively compressed per-perturbation predictions and recover more of the real-data causal graph. The shared failure is therefore neither in the reconstruction loss form nor in the marginal magnitude distribution; it is in the joint covariance and conditional-independence structure that GIES extracts from the multi-cell intervention regime. Proposition 5.1 predicts a covariance-aligning objective is the kind that can move causal recovery, because it acts on the joint second-order structure that per-gene calibration provably cannot; we release the penalty as a validated, unit-tested module (causalcellbench.training.covariance_regularizer).

To test, without GPU retraining, whether a joint operation can close the gap that per-gene calibration cannot (Proposition 5.1), we fit a small parametric head g_ϕ (two linear layers with a ReLU, 200$\rightarrow$200) on CPA's predicted shift vectors. Each gene's CPA shift $\hat{\delta} = \hat{\mu}_{\text{CPA}} - \mu_{\text{ctrl}}$ is the input and the real shift δ_{real} the target, fit on a 160-perturbation training split and evaluated on the held-out 40, identical to the calibration sweep (§5). Because g_ϕ has parameters fit on the training perturbations and is applied to unseen test perturbations, it is a learned, generalizing transform rather than a per-gene lookup. Two objectives are compared: recon_only ($\|g(\hat{\delta}) - \delta_{\text{real}}\|^2$)

Variant	Edges	F1	Precision	Recall
loss_gauss (Gaussian, not NB)	1,215	0.401	0.402	0.400
latent_16 (latent dim 16, base 64)	1,188	0.413	0.418	0.407
latent_256 (latent dim 256)	1,200	0.418	0.422	0.415
decoder_1layer (decoder depth 1)	1,207	0.418	0.420	0.416
layer_norm (LN replaces BN)	1,212	0.399	0.402	0.396
no_adversarial (reg=pen=0)	1,207	0.387	0.389	0.385

Table 11. Six-variant CPA architectural ablation on K562 at N=200, seed 42, partition size 20, against real-data GIES ground truth (1,220 edges). All other hyperparameters are held at the vanilla CPA baseline (n_latent=64, encoder layers = 3, decoder layers = 3, batch-norm, NB reconstruction loss, adversarial covariate disentanglement, max_epochs = 300).

Seed	Precision	Recall	Edges	Lift
42	0.009	0.021	222	0.46
123	0.014	0.031	216	0.69
456	0.022	0.052	223	1.14
789	0.036	0.082	224	1.84
1024	0.000	0.000	224	0.00
mean	0.016 ± 0.014	0.037 ± 0.031	—	$\mathbf{0.83 \pm 0.70}$

Table 12. GIES-on-real ground truth vs. curated TRRUST on the TF-rich subset, five seeds. "Lift" is precision over the matched random-edge floor (chance precision $\approx$ 0.020). The mean lift of 0.83× spans 1.0 and is not robustly above chance; the favorable seed (789, 1.84×) and the original single-seed estimate do not generalize. Reproducible via scripts/j1_multiseed.py.

and cov_reg (adding $\lambda\|\mathrm{Cov}(g(\hat{\delta})) - \mathrm{Cov}(\delta_{\mathrm{real}})\|_F^2$ normalized, $\lambda = 1$). Corrected means $\mu_{\mathrm{ctrl}} + g_\phi(\hat{\delta})$ pass through the identical overlay→GIES pipeline; F1 is scored against GIES on the real held-out means, across five random splits.

Condition	Held-out F_1	Δ vs. vanilla
Vanilla CPA	0.424 ± 0.025	—
Learned head (recon only)	0.538 ± 0.041	+0.115
Learned head (cov-reg, λ=1)	0.535 ± 0.015	+0.111

Table 13. No-GPU covariance-regularizer proxy: a learned joint transform on CPA shifts, mean $F_1 \pm$SD over five 160/40 splits. A learned generalizing transform recovers +0.111 F1 over vanilla CPA where post-hoc marginal calibration (§5) recovered nothing. The covariance-Frobenius penalty itself does not help: a pre-registered weight sweep $\lambda \in \{0, 0.1, 0.3, 1, 3, 10, 30, 100\}$ selected on a validation split with a sealed test split touched once finds the best-validation weight (λ=100) gains only $+0.003 \pm 0.035$ test F_1 over recon-only (3/5 seeds), failing the multi-seed significance gate. Reproducible via scripts/j2_covariance_regularizer_experiment.py and scripts/j2_lambda_sweep.py.

The proxy is a learned post-processing of CPA's predicted means, not end-to-end retraining; it shows the substitutability gap is partially closable by a joint operation that generalizes across perturbations (consistent with the identifiability argument) and isolates that the operative ingredient is the learned joint mapping rather than the covariance-Frobenius term at this λ. End-to-end retraining with the released regularizer and a λ sweep is the definitive follow-up.

C.15. Three-mode calibration sweep with proper train/test discipline

We extend the original Bolstad–Reinhard quantile-matching calibration with two additional modes that target joint-distribution structure: covariance preservation (Mode 2: project predicted shift onto the top-50 principal-subspace of real-perturbation covariance, shrinkage 0.5) and sparsity injection (Mode 3: soft-threshold $|\log_2 \mathrm{FC}|$ to match the empirical near-zero fraction). All six permutations of the three modes are evaluated, fit on a held-out 160-perturbation training set and applied to a 40-perturbation test set, with downstream GIES on each calibrated prediction. Table 14 reports the F1 of the best-ordering per model and the corresponding vanilla (uncalibrated) F1 on the same 40-perturbation test split. Across nine ordering choices (three single-mode plus six pipeline permutations), no calibration ordering exceeds the uncalibrated baseline on any model: GEARS

vanilla 0.303 vs. best calibrated 0.290 (−0.013); CPA 0.309 vs. 0.289 (−0.020); Geneformer 0.319 vs. 0.287 (−0.032). With the train/test split applied honestly, the magnitude calibration story does not survive on a held-out perturbation set: post-hoc calibration as currently implemented redistributes prediction mass on the magnitude marginal in a way that hurts, rather than helps, downstream causal recovery. The substitutability gap is therefore not in the magnitude marginal; it is in the joint structure, and joint-structure calibration via the covariance and sparsity modes evaluated here also fails to recover the gap. This is a stronger negative result than the earlier in-domain single-seed finding; it implies that closing the substitutability gap requires training-time intervention (e.g. a covariance-preserving regularizer in the perturbation-prediction loss) rather than post-hoc rescaling.

Model	Vanilla	Best cal.	Order
GEARS	0.303	0.290	q
CPA	0.309	0.289	q→c→s
Geneformer	0.319	0.287	c→s→q

Table 14. Phase 4 calibration sweep on K562 test split (40 of 200 perturbations held out). Per-model best F1 over nine ordering choices vs. the uncalibrated baseline on the same test split. Ordering codes: s sparsity, c covariance, q quantile. No ordering exceeds vanilla; the substitutability gap does not close under post-hoc magnitude or joint-structure calibration.

C.16. Why marginal calibration cannot close the gap: proof and finite-sample reading

The held-out calibration null of §5—no quantile-matching ordering beats the uncalibrated predictor—is not an artifact of our particular calibrator. It is forced by the invariance stated as Proposition 5.1, proved here.

Proof of Proposition 5.1. Conditional independence $A \perp\!\!\!\perp B \mid C$ is a statement about the generated σ-algebras $\sigma(A), \sigma(B), \sigma(C)$. A strictly monotone map is a measurable bijection onto its image with measurable inverse, so $\sigma(g_k(\hat{X}_k)) = \sigma(\hat{X}_k)$ for every k; the three σ-algebras are unchanged and the independence statement is preserved in both directions. The equivalence class targeted by constraint-based discovery (and by score-based discovery in the faithful limit) is determined by the set of such relations, which is therefore identical for $\hat{X}$ and $g(\hat{X})$. □

Corollary C.1. If the predictor's conditional-independence structure differs from the real data's, no per-gene marginal calibration can reduce that difference; the substitutability gap is unreachable by post-hoc marginal transforms and requires changing the predictor's joint conditional-independence structure—i.e. a training-time objective on the covariance or precision support, not a rescaling of the marginal.

Proposition 5.1 is exact for the constraint-based backend (PC) and for the equivalence-class target of GIES; for score-based GIES and the covariance/precision backends (NOTEARS, inspre) a monotone transform alters the finite-sample Gaussian score, but only weakly—consistent with the small, sign-mixed ΔF_1 we observe under calibration (Table 4). The empirical magnitude/joint decomposition is thus the finite-sample reflection of an identifiability fact: the marginal effect-size distribution and the conditional-independence structure are separately addressable, and only the latter governs causal recovery.

C.17. Full limitations

Ground truth. Track A ground truth (real-data GIES) is internally consistent but does not by itself establish biological correctness, and this is the central limitation of the benchmark. To test whether the ground truth encodes real biology rather than algorithm-internal structure, we re-selected a TF-rich 70-gene subset that maximizes coverage of the literature-curated TRRUST v2 network (Han et al., 2018) (∼100 curated TF–target edges, vs. effectively zero in the essential-gene evaluation subset) and scored the GIES-on-real graph against it across five seeds (§C.13). The result is a genuine null: the recovered ground truth agrees with TRRUST at $0.83 \pm 0.70\times$ a matched random-edge floor—not robustly above chance—and the favorable single seed reaches at most $1.84\times$ (Table 12), not robustly above chance. This is consistent with the technical-duplicate floor (two GIES recoveries on disjoint real-cell halves agree at only $F_1 = 0.045$): at this gene-set scale GIES is too cell-sampling-sensitive for its output to be validated against a sparse curated network, echoing Kernfeld et al. (2024) that transcriptome data alone cannot control regulatory-network recovery. The circularity of the Track-A

ground truth therefore remains a genuine limitation rather than one we have closed; a larger and denser external gold standard (e.g. ENCODE enhancer–gene links or perturbation-anchored CRISPRi maps) on a coverage-optimized subset, where GIES recovery is more stable, is the priority extension for absolute biological validation. Pretraining contamination. Geneformer V2's Genecorpus pretraining includes Replogle 2022, so its Replogle F1 is partly memorization — one reason we exclude it from the cross-dataset trained-model comparison above. A clean contamination test would need a temporally held-out post-pretraining dataset: a 2024+ public Perturb-seq atlas on a Geneformer-corpus-excluded human cell type with ≥ 50 perturbation-eligible genes, which is not currently available. Seed budget. Multi-seed CIs reported here use $n \leq 20$ seeds for K562 and $n = 1$ for RPE1 and Norman; the qualitative finding that pairwise differences among predictive methods are not significant is robust to seed count, but minimum detectable F1 differences are approximately 0.011 at $n = 20$ (ElasticNet, overlay, GRNBoost2) and approximately 0.022–0.027 for the $n = 5$ deep-model comparisons, so the deep-vs-linear equivalence is a bounded null rather than a tight one. Cross-dataset deep-model scope. Both GEARS and CPA were retrained from scratch and evaluated through the full GIES pipeline on RPE1 (cross-context) and Norman (cross-paradigm); all four GIES model×dataset cells recover no more of the real causal graph than chance, with the three further backends reported in §4.5, and STATE multi-seed (§5) shows no robust advantage on K562. The cross-context and cross-paradigm claims therefore rest on two architecturally distinct trained deep models, not the ElasticNet baseline alone. Geneformer is omitted on RPE1/Norman because its prediction in our pipeline is a cosine-similarity heuristic over the control mean rather than learned in-silico-perturbation expression, so it would not constitute a trained-model test. Model roster. GEARS, CPA, and Geneformer were chosen for architectural diversity at the time of pipeline construction; they predate STATE (Adduri et al., 2025; Arc Institute, 2025) and other 2025–2026 perturbation models. Recent benchmarks (Wong et al., 2025; Csendes et al., 2025; Wei et al., 2025; Viñas Torné et al., 2025) cover additional models on the per-gene prediction task but not on the causal-recovery task evaluated here. Gene-set scale. Primary results are at $N = 200$ for compute-tractability of GIES partitioning; scale-sweep numbers extend to $N = 200$ along seven points and inspre extends scaling to ∼1000 genes, but a head-to-head deep-model evaluation at $N = 500$–1000 is not yet completed.

D. Supplementary methods

D.1. Virtual cell model hyperparameters

All four predictive models were trained on the same K562 200-gene subset using their authors' default hyperparameter values where possible; deviations are deliberate matches to the published reference configurations of each model.

D.2. STATE inference pipeline

The STATE result was obtained zero-shot from the released ST-SE-Replogle/zeroshot/k562 checkpoint via the state CLI v0.10.5. The pipeline has two stages: (1) SE-600M cell encoder (state emb transform) maps each input cell to a 2,058-dimensional X_{state} embedding; (2) the ST-SE-Replogle perturbation model (state tx infer) maps $(X_{\text{state}}, \text{perturbation})$ pairs to predicted gene-expression vectors over the same N=200 gene panel.

Reproducing the published checkpoint required diagnosing three distinct framework-versioning bugs:

1. Decoupled head dimension. The released checkpoint encodes a Llama backbone with hidden_size=328, num_attention_heads=12, and explicit head_dim=64, i.e. $328 \neq 12 \times 64 = 768$. Transformers ≥5 add a strict validate_architecture check that rejects hidden_size % num_attention_heads != 0 regardless of an explicit head_dim; transformers 4.45 accepts decoupled head dimensions but lacks the (2) compatibility we describe next; transformers 4.52.3 supports both. Fix: pin transformers==4.52.3 after pip install -e state.

2. is_causal kwarg pass-through. The state package's LlamaBidirectionalModel.forward sets is_causal=False and forwards it to the parent LlamaModel.forward via **flash_attn_kwargs. transformers 4.45's signature lacks the **flash_attn_kwargs catch-all, so the call raises TypeError: unexpected keyword argument 'is_causal'. transformers 4.52.3 introduces the catch-all and accepts the call.

3. ALL_PARALLEL_STYLES is None on Modal. In transformers 4.52.3 specifically, transformers.modeling_utils.ALL_PARALLEL_STYLES is initialized to None in some container configurations (verified locally and on Modal), making LlamaModel.__init__'s post_init hit if v not

Hyperparameter	Value
CPA (Lotfollahi et al., 2023)	
n_latent	64
n_layers_encoder	3
n_layers_decoder	3
batch_size	512
max_epochs	300
n_epochs_kl_warmup	20
n_epochs_adv_warmup	50
n_epochs_pretrain_ae	10
reconstruction loss	Negative Binomial
GEARS (Roohani et al., 2024)	
hidden_size	64
epochs	20
batch_size	32
Geneformer V2 (Theodoris et al., 2023)	
backbone	ctheodoris/Geneformer (HF)
fine-tuning task	in-silico perturbation
forward_batch_size	16
ElasticNet baseline (Zou & Hastie, 2005)	
alpha	0.1
l1_ratio	0.5
max_iter	1,000
STATE (Arc Institute, 2025)	
embedding model	SE-600M
prediction model	ST-SE-Replogle / zeroshot/k562
hidden_size	328
num_attention_heads	12
head_dim	64 (decoupled, $\neq H/h$)
num_hidden_layers	8
intermediate_size	3,072
max_position_embeddings	64
embedding dim (X_{state})	2,058

Table 15. Training hyperparameters per virtual cell model. Geneformer is fine-tuned from the public ctheodoris/Geneformer V2 checkpoint; STATE is run zero-shot from the released ST-SE-Replogle K562 checkpoint.

in ALL_PARALLEL_STYLES and raise TypeError: argument of type 'NoneType' is not iterable. Fix: during image build, replace the offending line in transformers/modeling_utils.py with if ALL_PARALLEL_STYLES is not None and v not in ALL_PARALLEL_STYLES:; this propagates to the state CLI subprocess, where a sitecustomize.py monkey-patch did not reliably fire.

The full set of attempted patches (14 attempts) and the final working Modal image specification is in scripts/modal_run_state_v3.py.

Subset rationale. STATE inference was run at two subset sizes per seed: a 320-cell subset (120 controls + one perturbed cell per gene) and, to rule out a low-sample-mean artifact, a 1,200-cell subset (200 controls + five perturbed cells per gene, run under a 12-hour Modal timeout). Both are reductions of the full 38,979-cell K562 panel, which at the measured SE-600M cost of ∼28 s/cell (batch_size=8) would take ∼300 h. The 320-cell subset embeds in ∼13.5 minutes and the 1,200-cell subset in ∼9.3 hours; both produce 200 STATE-predicted perturbation profiles (one per perturbation) and yield matching per-seed F_1 (§5), confirming the multi-seed instability is not a sampling artifact.

D.3. Norman 2019 cross-paradigm pipeline

The Norman 2019 (Norman et al., 2019) dataset (GSE133344) was downloaded from CellxGene Census and contains 111,445 K562 cells under 237 unique perturbations (single and combinatorial). After dropping combinatorial ('A+B') perturbations and keeping only control (n=11,855 cells) plus single-gene perturbations with at

least 10 cells, 102 eligible target genes remain, all measurable in the standard expression panel.

The substitutability evaluation re-uses the K562 protocol unchanged: 200 control cells + 100 cells per perturbation overlay-sampled from the ElasticNet predictions, vstacked into one expression matrix and fed to GIES with partition size 20 and seed 42.

Two practical engineering issues are noted. (i) multiprocessing.Pool(fork) for parallel partition execution deadlocks on the semaphore in Modal's container runtime; switching to sequential GIES execution removes the deadlock at the cost of single-thread CPU throughput (acceptable: total wall time ≈28 s on the M4). (ii) Some Norman partitions contain genes with zero variance under particular interventional regimes, producing 1/omega = inf in the gauss_int score and an effectively infinite hill-climbing loop in GIES. Adding $\mathcal{N}(0, 10^{-6})$ jitter to the partition expression matrix before scoring prevents the degeneracy without measurably altering recovered edge sets in the well-conditioned partitions.

D.4. Causal-discovery backend implementations

GIES (Hauser & Bühlmann, 2012): Greedy Interventional Equivalence Search via the gies Python package (v0.0.1), score function gauss_int_l0_pen, ∼20-gene partitions for tractability of the underlying CPDAG search. Output edges are the union of single-direction CPDAG arrows across partitions; bidirectional CPDAG arrows are kept once with the canonical ordering (lower index first).

inspre (Brown et al., 2025): Sparse-inverse regression on the perturbation-conditional ACE matrix via the inspre R package, full λ-path with leave-one-target-out cross-validation. The Modal R kernel (renv-isolated) is invoked via Rscript subprocess to keep the Python pipeline driver-only.

PC (Spirtes & Glymour, 2000): Constraint-based causal discovery via the causal-learn Python package, Fisher-Z conditional independence at $\alpha = 0.05$. Run on each model's per-perturbation ACE matrix.

NOTEARS (Zheng et al., 2018): Continuous-optimization with acyclicity constraint, $\lambda = 0.05$, weight threshold $w_{\text{thresh}} = 0.1$. Implementation: reference notears package on the same per-condition ACE inputs.

Overlay sampling (used by GIES, PC, NOTEARS): for each perturbed condition, 100 synthetic cells are produced by drawing 100 control cells with replacement, multiplying by a per-gene ratio of (predicted-mean $+10^{-2}$) over (control-mean $+10^{-2}$) clipped to $[0.01, 10]$, and Poisson-resampling to count-like expression. This preserves per-cell control covariance structure while imposing the predicted shift in marginal magnitude.

E. Reproducibility

E.1. Software environment

Component	Version
Python	3.9 (local) / 3.11 (Modal)
PyTorch	2.4.1 (CPU)
transformers	4.52.3 (STATE-pinned)
tokenizers	0.21.x
huggingface_hub	0.30.x
anndata	0.10.9
scanpy	1.10.3
numpy	1.26.4
scipy	1.13.1
scikit-learn	1.x
gies (Python)	0.0.1
causal-learn	0.1.x
notears	reference impl. from (Zheng et al., 2018)
inspre (R)	0.0.1
state (Arc Institute)	0.10.5

Table 16. Software versions used end-to-end. Modal containers pin all dependencies via the image build steps in scripts/modal_run_*.py.

E.2. Compute resources

Local pipeline (overlay sampling, GIES, PC, NOTEARS, calibration, multi-seed aggregation, plotting) runs on an Apple M4 laptop with 16 GB unified memory. Wall time: ≈45 minutes for the full N=200, single-seed K562 evaluation.

Modal cloud is used for the model-training and large-scale inference steps: GEARS, CPA, Geneformer, STATE, and the inspre N=500 scale point. Modal H100 instances are reserved for GEARS and Geneformer fine-tuning (≈1 h each at N=200); A10G for CPA training (≈45 minutes); 16-CPU 64 GB instances for inspre and the STATE CPU fallback. The largest single Modal job wall-clock is the SE-600M embedding step in the STATE pipeline at ≈3.5 hours for the full embedding, ≈13 minutes for the 320-cell subset.

E.3. Random seeds

The single-seed numbers throughout the main text use seed=42. Multi-seed CI tables in §C.10 use $\{42, 123, 456, 789, 1024\}$. Inside the GIES partitioning step, the gene-permutation seed is 42; partition contents are deterministic given the panel and seed.

E.4. End-to-end reproduction commands

All numbers in the main text and tables are reproducible from the released code with seed=42.

```
# Calibrate predictions
PYTHONPATH=. python3 scripts/calibrate_predictions.py \
   --diagnostics-out results/calibration_diagnostics.json

# Vanilla evaluation
PYTHONPATH=. OMP_NUM_THREADS=1 python3 \
  scripts/run_eval_local.py \
   --data-path data/k562_subset_n200_slim.h5ad \
   --pre-subset --n-genes 200 --partition-size 20 \
   --n-cells 100 --n-ctrl 200 --seed 42 \
   --model-outputs-dir data/model_outputs_real_n200 \
   --gies-cache-dir results/gies_cache --skip-random

# Calibrated evaluation
PYTHONPATH=. OMP_NUM_THREADS=1 python3 \
  scripts/run_eval_local.py \
   --data-path data/k562_subset_n200_slim.h5ad \
   --pre-subset --n-genes 200 --partition-size 20 \
   --n-cells 100 --n-ctrl 200 --seed 42 \
   --model-outputs-dir \
      data/model_outputs_real_n200_calibrated \
   --gies-cache-dir results/gies_cache_calibrated \
   --skip-random

# STATE evaluation (after Modal returns state_predictions.h5ad)
PYTHONPATH=. OMP_NUM_THREADS=1 python3 \
  scripts/state_eval_gies.py

# Norman 2019 cross-paradigm
.venv_gears/bin/python scripts/run_norman_local.py
```