BEYOND PERTURBATION PREDICTION: LEARNING DIRECTED GENE STRUCTURE WITH RISCAL

Anonymous authors
Paper under double-blind review

ABSTRACT

Predicting genetic perturbation responses and recovering their regulatory mechanisms are central to understanding gene function and therapeutic target prioritization. Existing methods often emphasize response prediction or structure recovery, while assuming perfect perturbations or modeling imperfect ones as unrestricted mechanism changes. We introduce RISCAL (Residual-activity Interventions for Structural Causal Learning), a nonlinear gene-level structural causal model that jointly learns directed dependencies and predicts responses under residual-activity interventions. Its intervention model captures imperfect loss-of-function by separating gene-specific residual functional activity from measured transcript abundance. RISCAL accurately recovers synthetic ground-truth DAGs and matches the strongest predictive baselines on LINCS L1000 profiles from ten cell lines. Its predicted graph cores are enriched for CollecTRI/SIGNOR interactions and form coherent, partially conserved functional programs. RISCAL can also incorporate unseen regulators and predict responses consistent with experimental evidence.

1 INTRODUCTION

Large-scale perturbational resources such as the Connectivity Map (CMap) and Library of Integrated Network-based Cellular Signatures (LINCS) measure how chemical and genetic interventions reshape gene-expression programs (Subramanian et al., 2017; Keenan et al., 2018). These data support two complementary tasks: predicting an intervention's expression response from a baseline profile, and identifying the directed regulatory relationships that mediate it. Most methods emphasize either one of the tasks. Perturbation-response models map baseline profiles to treated expression, sometimes using protein-protein, gene-regulatory networks, or biological pathways as prior structure (Gonzalez et al., 2026; Roohani et al., 2024; de la Fuente et al., 2025). Their interaction structure is typically fixed or not interpreted as a learned directed gene-level graph. On the other hand, causal-discovery methods estimate directed acyclic graphs (DAGs) from observational and interventional data (Zheng et al., 2018; Brouillard et al., 2020), but generally explain distributions across intervention regimes rather than predict intervened profiles matched to baselines. Solving these tasks jointly could connect predictive accuracy with biological insight.

To generate such perturbations, CRISPR loss of function often reduces rather than eliminates target activity due to incomplete editing, in-frame repair, transcriptional adaptation, or variable guide efficacy (Smits et al., 2019; Wang et al., 2024; Ito et al., 2024; Jost et al., 2020). So setting target activity to zero (i.e., a perfect intervention) may misrepresent downstream effects. Additionally, measured target mRNA may not reflect residual function, as transcripts can persist after protein function is lost. These observations motivate partial attenuation and a distinction between measured target transcript and the effective activity propagated downstream.

We introduce RISCAL (Residual-activity Interventions for Structural Causal Learning), a nonlinear causal model that uses a gene-specific residual-activity parameter ρ_k to attenuate the target's effective activity propagated downstream, while a separate readout predicts its measured transcript. Further, nonlinear child-specific mechanisms predict responses and define directed neural path strengths, with training encouraging sparse, acyclic support. RISCAL returns both a predicted treated profile and a learned causal DAG. Note that we interpret inferred edges as model-derived biological hypotheses because hidden confounding and unmeasured regulators can induce apparent dependencies.

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

RISCAL retains informative structural recovery at 3,710 genes while matching strong predictive baselines. On LINCS, it yields reproducible graph cores enriched for curated directed interactions and coherent biological programs. We further test model by incorporating an unseen regulator, obtaining responses consistent with reported experimental phenotypes. These real-data results support reproducibility and biological plausibility, not edge-level causal correctness.

Our contributions are:

1. **Joint prediction and structure learning.** We introduce a nonlinear gene-level model whose shared mechanisms predict matched perturbation responses and learn directed gene-gene structure.
2. **Residual-activity intervention modeling.** We distinguish the target's effective activity propagated downstream from its measured treated expression, predicted by a separate readout.
3. **Biology-grounded benchmark.** We develop a configurable generator with ground-truth DAGs and biological properties, providing a testbed for causal discovery and perturbation prediction.

2 Related Work

Differentiable causal discovery. Classical constraint- and score-based methods for causal discovery can be difficult to scale, and rely on strong assumptions (Spirtes et al., 2000; Chickering, 2002). Differentiable approaches instead pose causal learning as continuous optimization, replacing discrete edge searches with gradient-based updates (Zheng et al., 2018; Yu et al., 2019; Lachapelle et al., 2020; Ng et al., 2020). Later methods, such as SDCD, replace costly matrix-exponential acyclicity constraints with spectral alternatives that improve stability and scalability (Nazaret et al., 2024).

Causal structure learning for perturbational transcriptomics. Interventional data, such as genetic interventions, can distinguish causal structures that are observationally indistinguishable by revealing how targeted perturbations alter downstream variables (Yang et al., 2018). In this context, DCDI established a differentiable framework for learning causal graphs across multiple known intervention regimes (Brouillard et al., 2020), and subsequent methods exploit intervention targets, causal ordering, exogenous noise, or alternative acyclicity parameterizations (Chevalley et al., 2025b; Xue et al., 2025; Lippe et al., 2022; Viñas Torné et al., 2026). Other approaches improve scalability in genomic settings through structured graph parameterizations (Lopez et al., 2022; Brown et al., 2025), while BaCaDI jointly infers causal structure and uncertain intervention effects (Hägele et al., 2023). Despite these advances, recovering gene regulatory structure remains challenging (Chevalley et al., 2025a).

Predicting Perturbation Responses. Recent models such as scGen, ChemCPA, CellOT, and GEARS predict cellular responses to perturbation, or identifies interventions that induce a desired expression profile, with varying generalization across cell types, contexts, and unseen interventions (Lotfollahi et al., 2019; Hetzel et al., 2022; Bunne et al., 2023; Roohani et al., 2024). Causal differential networks, on the other hand, identify intervention targets through changes in graph structure (Wu et al., 2024). When gene-gene connectivity is incorporated, as in GEARS or PDGrapher, it is typically supplied through ontology or co-expression networks (Gonzalez et al., 2026). Similarly, biologically informed neural networks constrain their architectures using predefined gene programs (Ma et al., 2018; Elmarakeby et al., 2021; Lotfollahi et al., 2023; de la Fuente et al., 2025). These approaches predict responses, learn signal propagation or provide interpretability within prescribed architectures, rather than jointly inferring directed gene-level causal connectivity from the same perturbational data.

Joint Structure Learning and Perturbation Modeling. Several approaches combine structure learning with perturbation modeling. DCD-FG combines genomic-scale structure learning and held-out intervention prediction (Lopez et al., 2022), while GIM learns causal models and feature-dependent interventions (Schneider et al., 2025). Discrepancy-VAE recovers latent causal factors from soft interventions (Zhang et al., 2023); deep SCMs and causal normalizing flows support intervention inference (Pawlowski et al., 2020; Javaloy et al., 2023). Bicycle learns cyclic regulation and predicts unseen responses under known, perfect interventions (Rohbeck et al., 2024). GPO-VAE couples response prediction with directed, GRN-aligned latent parameters using optimal-transport control-perturbation pairings (Baek et al., 2025). CASCADE learns a gene-level DAG, nonlinear mechanisms, and intervention coefficients for counterfactual prediction, constraining real-data graph search with a prior regulatory scaffold (Cao et al., 2025). Despite these advances, current joint models lack proper modeling of residual gene function while accurately inferring regulatory structure.

3 Problem Formulation

We study joint perturbation-response prediction and directed gene-regulatory structure discovery from paired control and treated expression profiles. For cell line c, let G_c be the number of measured genes and N_c the number of paired profiles. Thus, the dataset consists of the set of tuples

$$\mathcal{D}_c = \left\{ \left(x_n^{d,c}, u_n^c, x_n^{t,c} \right) \right\}_{n=1}^{N_c}, \tag{1}$$

where $x_n^{d,c} \in \mathbb{R}^{G_c}$ is the control expression profile, $x_n^{t,c} \in \mathbb{R}^{G_c}$ is the corresponding treated expression profile after a single-gene CRISPR loss-of-function (LoF) perturbation, and $u_n^c \in \{0,1\}^{G_c}$ is the one-hot vector identifying the intervention target (Figure 1-A). Each pair comprises measurements from distinct wells on the same cell line and plate. Section 6 describes the computational pairing and the grouped training, validation, and test splits. In what follows, we omit the cell line index c for clarity. Thus, our objective is to jointly learn (Figure 1-B,C):

- A nonnegative path-strength matrix $A \in \mathbb{R}_{\geq 0}^{G \times G}$, where A_{ij} represents the neural path strength connecting input gene i to child gene j.
- The complete perturbation-conditioned predictor $F_\theta(x^d, u) = \hat{x}^t$, which predicts the measured treated expression profile from the control expression profile and intervention identity.

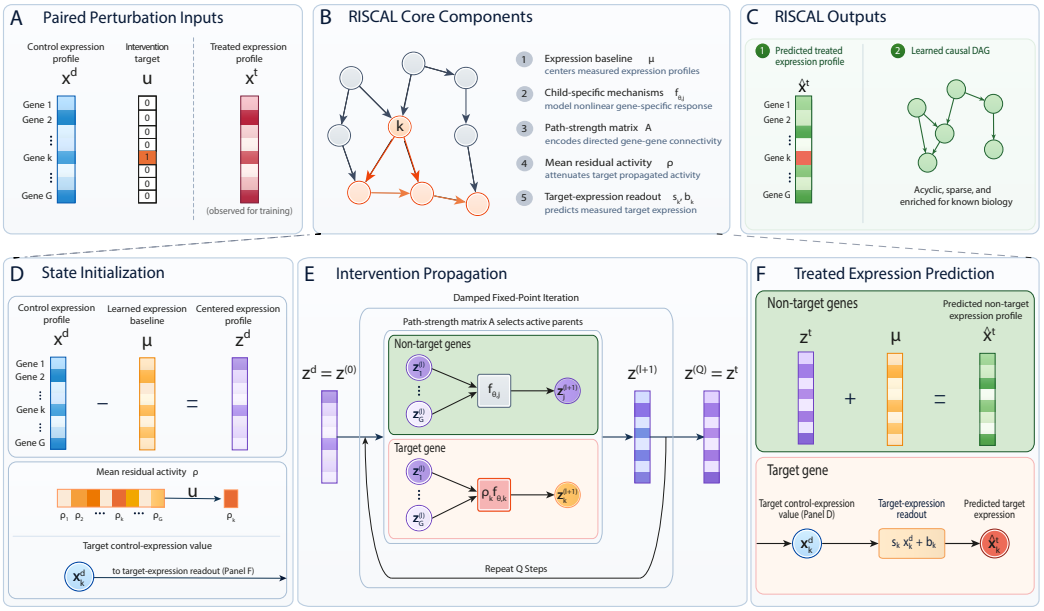

Figure 1: Overview of RISCAL. (A) Paired perturbation inputs. (B) Core model components. (C) Predicted treated profile and learned causal DAG. (D) State initialization using the learned expression baseline, target-specific residual activity, and target readout. (E) Intervention propagation through the learned structure by damped fixed-point iteration. (F) Treated-expression prediction combining non-target activities with a separate target-specific readout.

4 RISCAL

RISCAL jointly learns directed gene dependencies and predicts perturbation responses through shared nonlinear child-specific mechanisms (Figure 1). Unlike previous models, RISCAL distinguishes the effective activity propagated from an intervention target from its measured expression. To this end, we represent the measured expression profile relative to a learned expression baseline (Section 4.1) and define nonlinear child-specific mechanisms and their shared directed structure (Section 4.2). Then model incomplete LoF through attenuating target activity (Section 4.3). Given a control profile and an intervention, the coupled mechanisms propagate the perturbation (Section 4.4), and a target-specific readout completes the predicted treated expression profile (Section 4.5). Finally, joint training and validation select an acyclic graph and its associated perturbation-response model (Section 4.6).

4.1 PREPROCESSING OF MEASURED EXPRESSION PROFILES

We center expression profiles to separate the learned expression baseline from the variation modeled by the child-specific mechanisms. This gives each gene a zero reference, allowing the nonlinear child-specific mechanisms to describe regulatory deviations around a learned expression baseline.

Let $\mu \in \mathbb{R}^G$ be a learned expression baseline initialized using only control expression profiles in the training partition (Figure 1 panel D). For any measured expression profile x, its centered expression profile is $z = x - \mu$. Thus, z^d and z^t are the centered expression profiles of x^d and x^t, respectively. These centered expression profiles are the input to the nonlinear child-specific mechanisms (see Section 4.2), whose outputs are then returned to the measured scale by adding μ back.

4.2 CHILD-SPECIFIC NONLINEAR STRUCTURAL CAUSAL MECHANISMS

RISCAL assigns each gene its own nonlinear child-specific mechanism through a dedicated neural network (Figure 2). Each mechanism predicts the child gene's centered expression from the graph-gated centered expression of the other genes. These networks capture nonlinear responses, including saturation and interactions between regulators, following the child-specific graph-masked architectures of DCDI, GIM, and GraN-DAG (Brouillard et al., 2020; Schneider et al., 2025; Lachapelle et al., 2020).

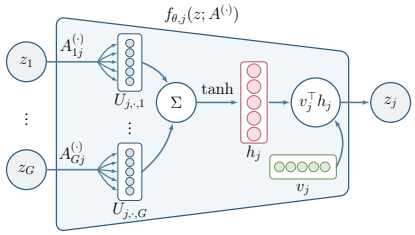

Figure 2: Child-specific mechanism $f_{\theta,j}$. The graph gate $A^{(\cdot)}$ selects inputs from the centered expression profile.

For gene j, let $U_j \in \mathbb{R}^{H \times G}$ and $v_j \in \mathbb{R}^H$ be the input and output weights of the NN's hidden layer with H hidden units. Following GraN-DAG, these weights define a directed neural path from gene i to gene j:

$$A_{ij} = \sum_{h=1}^{H} |v_{j,h}| \, |U_{j,h,i}|, \qquad A_{ii} = 0, \tag{2}$$

defining the path-strength matrix A. Note that $A_{ij} = 0$ guarantees that gene i cannot directly affect gene j. The same parameters determine both the child-specific mechanisms and the directed graph, albeit neural path strength measures potential connectivity, not causal effect sign or magnitude.

During training, A is converted into the soft gate matrix $A^{(g)} = \sigma\left(\frac{A_{ij} - q_{\text{gate}}}{\tau}\right)$, where q_{gate} and $\tau > 0$ are fixed gate hyperparameters. This soft transformation attenuates weak paths while retaining differentiability. For evaluation, we use the binary adjacency $A^{(b)} = \mathbf{1}[A_{ij} \geq q_{\text{eval}}]$ that selects active inputs through the threshold q_{eval}. In what follows, $A^{(\cdot)}$ refers to both $A^{(g)}$ and $A^{(b)}$.

Thus, we define the child-specific mechanism for gene j as

$$f_{\theta,j}(z; A^{(\cdot)}) = v_j^\top \tanh\left(\sum_{i \neq j} A^{(\cdot)}_{ij} U_{j,\cdot,i} z_i\right), \tag{3}$$

where $U_{j,\cdot,i} = (U_{j,1,i}, \ldots, U_{j,H,i})^\top$ is the vector containing the weights connecting input gene i to the hidden units of the NN of child gene j. We denote the vector of all child-specific mechanisms by $f_\theta(z; A^{(\cdot)}) \in \mathbb{R}^G$. Adding the baseline μ_j returns the output to the measured expression scale. Section 4.4 describes how these coupled mechanisms propagate an intervention through the network.

4.3 RESIDUAL LOF ACTIVITY

While a perfect gene knockout would completely eliminate the target's functional activity, CRISPR LoF can leave residual function through in-frame repair, transcriptional adaptation, and other responses (Ito et al., 2024). Importantly, measured target transcript abundance can be a poor proxy for the remaining gene activity, as mRNA may persist despite loss of protein function, while altered protein products can remain partially functional (Smits et al., 2019; Wang et al., 2024). To account for this, we introduce a latent residual-activity variable $r_k \in (0, 1)$, representing the retained fraction of target activity after intervention on gene k. Note that this quantity is not a direct measurement of editing efficiency, as it is inferred jointly with the causal model from transcriptional data.

We model residual activity as $r_k \sim \mathrm{Beta}(\rho_k\phi, (1-\rho_k)\phi)$, where the gene-specific mean ρ_k represents retained activity and the shared concentration ϕ controls its dispersion. Its moments and uncertainty propagation are detailed in Appendix A.1. We parameterize

$$\rho_k = \sigma(a + \widetilde{\eta}_k), \qquad \widetilde{\eta}_k = \eta_k - \frac{1}{G}\sum_{m=1}^{G}\eta_m, \qquad \phi = \mathrm{softplus}(\kappa), \tag{4}$$

where a, $\eta_1, \ldots, \eta_G$, and κ are learned jointly with the causal model. Centering ensures that the gene-specific deviations sum to zero, so a is the average residual-activity logit across genes.

During intervention propagation, ρ_k attenuates the target's effective activity, while uncertainty in r_k is propagated to its children using the Beta moments given in the appendix. Complete loss of the modeled target activity is recovered as $\rho_k \to 0$.

4.4 Intervention Propagation by Damped Fixed-Point Iteration

To model an intervention on gene k, RISCAL attenuates its mechanism output by the mean residual activity ρ_k (Figure 1, panel E):

$$x_k^{\mathrm{eff}} = \rho_k \left[\mu_k + f_{\theta,k}(z; A^{(\cdot)})\right]. \tag{5}$$

The intervention target still responds to its regulators, but its effective target activity is passed to downstream genes. All other genes retain their mechanisms.

During training, the graph may contain cycles, preventing evaluation in a single regulator-to-target order. We therefore approximate the effective treated activity profile using a damped fixed-point iteration (Bai et al., 2019), blending each non-target gene's current value with its mechanism output while directly refreshing the attenuated effective target activity. Specifically, we initialize non-target genes at their control expression values x_j^d and the target using Eq. (5) evaluated at z^d. At each step $\ell = 0, \ldots, Q-1$, we evaluate the mechanisms and update the effective activity profile:

$$\begin{aligned} n^{(\ell)} &= \mu + f_\theta(x^{(\ell)} - \mu; A^{(\cdot)}), \\ x_j^{(\ell+1)} &= \begin{cases} \rho_k n_k^{(\ell)}, & j = k, \\ (1-\delta)x_j^{(\ell)} + \delta n_j^{(\ell)}, & j \neq k. \end{cases} \end{aligned} \tag{6}$$

Here $n^{(\ell)}$ contains the mechanism outputs on the measured scale, and $0 < \delta \leq 1$ controls how far non-target entries move toward those outputs at each step. The target is reassigned its attenuated output at every step without damping. We monitor $\|x^{(Q)} - x^{(Q-1)}\|_1/G$ as a convergence diagnostic, and after Q steps, we obtain the effective treated activity profile $x^{t,\mathrm{eff}} = x^{(Q)}$.

Given an acyclic graph, the effective treated activity profile $x^{t,\mathrm{eff}}$ can be computed by evaluating each gene after its parents, starting with root nodes and applying residual attenuation to the target. This ordering ensures that every mechanism receives already computed input values. Our iterative procedure recovers the same effective activity profile upon convergence (Bongers et al., 2021), while a fixed budget of Q updates provides an approximation if convergence has not yet been reached.

4.5 Predicting Measured Perturbation Responses

The target's measured transcript abundance does not necessarily reflect the functional activity that drives downstream responses. RISCAL therefore predicts target expression through a separate gene-specific target-expression readout $\widehat{x}_k^t = s_k x_k^d + b_k$, where s_k and b_k are learned scale and bias parameters mapping the target's control expression to its treated expression (Figure 1, panel F).

The final predicted treated expression profile retains the non-target entries of the effective treated activity profile and replaces the target entry with the inferred readout $\widehat{x}_k^t$. The target-expression readout is used only for transcript prediction as downstream responses are computed from the effective target activity. This separates the target's measured expression from its modeled regulatory activity.

4.6 Joint Training and Acyclic Graph Selection

RISCAL jointly optimizes the child-specific mechanisms, residual-activity variables, target-expression readouts, learned expression baseline, and noise variances with Adam. The objective

combines expression-fitting losses with penalties for graph sparsity and directed cycles:

$$\mathcal{L}_{\text{train}} = \alpha_{\text{causal}}\Big[\alpha_{\text{SCM}}\mathcal{L}_{\text{SCM}} + \lambda_{\text{edges}}\mathcal{R}_{\text{sparse}}(A) + \lambda_t^{\text{acyc}}\widehat{h}(A) + \tfrac{1}{2}\mu_t^{\text{acyc}}\widehat{h}(A)^2\Big] + \alpha_{\text{target}}\mathcal{L}_{\text{target}}. \tag{7}$$

The α coefficients and λ_{edges} are fixed, while the acyclicity coefficients are updated during training. Their settings are reported in Supplementary Table 3.

Fitting intervention responses. We train each non-target mechanism using measured treated expression values as inputs, following interventional likelihood-based structure learning (Brouillard et al., 2020). The target input is replaced by its estimated effective target activity, as its measured transcriptional signal may not necessarily reflect the activity transmitted to downstream genes:

$$\check{x}_j^t = \begin{cases} x_j^t, & j \neq k, \\ \rho_k\left[\mu_k + f_{\theta,k}(z^t_{\setminus k}; A^{(g)})\right], & j = k, \end{cases} \qquad \check{z}^t = \check{x}^t - \mu. \tag{8}$$

This teacher-forced treated profile retains observed non-target expression while separating the target's transcript from its functional activity. For each training pair, the SCM NLL is

$$\mathcal{L}_{\text{SCM}} = -\frac{1}{G-1}\sum_{j\neq k}\log\mathcal{N}\Big(x_j^t; \mu_j + f_{\theta,j}(\check{z}^t; A^{(g)}), \sigma_j^2 + \nu_j\Big), \tag{9}$$

where σ_j^2 is the learned noise variance and ν_j accounts for locally propagated target uncertainty (Appendix A.1). The measured target is supervised separately through its target-expression readout:

$$\mathcal{L}_{\text{target}} = -\log\mathcal{N}\big(x_k^t; \widehat{x}_k^t, \varsigma_k^2\big), \tag{10}$$

with a separate learned variance ς_k^2. These losses are averaged over training pairs. The target-expression loss trains the readout without feeding its predictions into the child-specific mechanisms.

Structural regularization. The regulatory mechanisms within a cell are expected to be sparse, as each gene is regulated by only a limited subset of measured genes. Thus, we encourage sparse connectivity using $\mathcal{R}_{\text{sparse}}(A) = \frac{1}{G}\sum_{i,j} A_{ij}$, which penalizes total neural path strength. Note that when applied to $A^{(b)}$, it penalizes the number of selected edges.

To push A towards acyclicity, we use the spectral-radius $h(A)$ constraint from SDCD (Nazaret et al., 2024), as it is zero exactly when its support is acyclic. Specifically, we use a numerical estimate $\widehat{h}(A)$ and adaptive augmented-Lagrangian coefficients (Eq. (7)). This estimate guides optimization but does not certify acyclicity. Appendix A.2.1 gives the estimator and coefficient-update schedule.

DAG checkpoint selection. At each eligible epoch, only checkpoints whose binary adjacencies pass an acyclicity check are considered. Among these candidates, we select the one that minimize

$$\mathcal{S}_{\text{val}}(A^{(b)}) = \alpha_{\text{causal}}\left[\alpha_{\text{SCM}}\mathcal{L}_{\text{SCM,val}}(A^{(b)}) + \lambda_{\text{edges}}\mathcal{R}_{\text{sparse}}(A^{(b)})\right]. \tag{11}$$

It uses the binary graph, so selection balances SCM fit against the number of edges. The target-readout loss is excluded because it does not assess graph structure, also acyclicity penalties because candidates have passed the graph check. Algorithm 1 in Appendix A.3 summarizes the procedure. Threshold sensitivity, training convergence, and computational diagnostics are reported in Appendix A.4.

5 A Biology-Grounded Data Generator for Perturbation Data

Evaluating gene-level causal discovery requires a known ground-truth DAG, yet real perturbation screens lack complete ground-truth regulatory networks against which inferred edges can be assessed (Chevalley et al., 2025a). Synthetic benchmarks enable direct measurement of structural recovery (Nazaret et al., 2024), but their relevance to biological applications depends on the mechanisms, intervention semantics, and sampling conditions they represent. In particular, evaluating recovery from paired LoF screens requires accounting jointly for incomplete perturbations, differences between functional activity and measured expression, and the reuse of control profiles. We therefore develop a configurable synthetic generator that incorporates these features while retaining known ground-truth DAGs. Starting from a minimal linear SCM, we progressively introduce signed regulatory effects,

separate effective target activity and measured transcript readouts, gene-specific expression baselines and heterogeneous noise, variable residual activity under imperfect LoF, biologically motivated graph topologies with hubs and heterogeneous connectivity, child-specific nonlinear mechanisms, and reused controls. This incremental construction supports controlled assessment of how these features affect recovery (Appendix B details each stage). Configurable components and sampling regimes also make the generator a testbed for other causal-discovery and perturbation-response methods.

We evaluate causal recovery under two regimes that differ in dimensionality and sample availability (corresponding to the latest two stages of the synthetic generator Appendix B). The data-rich benchmark contains $G = 100$ genes and 300,000 paired profiles, assessing recovery with abundant observations relative to the modeled dimension. The PC3-scale benchmark contains $G = 3{,}710$ genes and 21,229 pairs, matching the gene universe and pair count of the PC3 prostate-cancer experiments and assessing recovery under real-data settings. Descriptive fidelity diagnostics for the PC3-scale benchmark are reported in Supplementary Table 9 and Supplementary Figures 19-20.

6 Experimental Settings

Real data (LINCS L1000) and Preprocessing. We evaluate on ten LINCS L1000 Level 3 cell-line datasets using PDGrapher-compatible profile selection (Gonzalez et al., 2026). For each cell line, we restrict features to the PDGrapher-compatible gene universe and then to genes targeted at least once, and transform expression as $\log_2(x + 1)$ (Appendix C.1 for detailed data acquisition, feature selection and filtering protocol). Each treated well is paired deterministically with a control well from the same detection plate through cyclic control reuse (Appendix C.2 for batch correction and pairing procedure). We split data by control identity to prevent shared controls across partitions, ensuring that every validation or test target is represented in training. Five train/validation fits per cell line share one fixed held-out test set.

Causal baselines comprise SDCD (Nazaret et al., 2024), InterSort (Chevalley et al., 2025b), GENIE3 (Huynh-Thu et al., 2010), and a Random DAG baseline. Full InterSort is used in the data-rich $G = 100$ benchmark and its scalable SORTRANKING mode in the PC3-scale benchmark. GENIE3 serves as baseline from the gene-regulatory-network literature and is the only causal baseline without an acyclicity mechanism.

Predictive baselines comprise linear regression, a multilayer perceptron, an axial transformer, PDGrapher (Gonzalez et al., 2026), and GEARS (Roohani et al., 2024). Among these baselines, we develop the axial transformer as a strong prediction-only reference, optimized exclusively for perturbation-response prediction and does not learn an interpretable gene-level DAG. Complete implementations, input contracts, support rules, and evaluation protocols are provided in Appendix D.1 for causal discovery baselines and E for predictive baselines. Selected evaluation metrics, their definitions, aggregation, and paired statistical inference are provided in Appendix F.

7 Results

We evaluate RISCAL along three complementary dimensions: recovery of known directed structure, prediction of perturbation responses, and biological plausibility of the learned graphs. We first assess structure recovery under the two synthetic regimes described in Section 5. We then compare predictive performance across ten LINCS cell lines and examine the reproducibility, curated-interaction enrichment, and functional organization of the inferred graphs. Finally, we incorporate an unseen regulator and recover responses consistent with experimental phenotypes reported for that regulator.

7.1 Causal Discovery and Ablation Studies on Synthetic Benchmarks

Data-rich $G = 100$ benchmark. At this scale, RISCAL attains the highest native-support AUPRC (0.850 ± 0.020), precision (0.778 ± 0.030), and F1 (0.812 ± 0.014). InterSort attains higher recall (0.940 ± 0.012), but its support is substantially denser (709.2 ± 44.2 predicted edges for 357.8 ± 3.0 ground-truth edges), whereas the RISCAL selected edge set uses 392.0 ± 23.8 edges. SDCD and GENIE3 are lower on all four structural metrics. Note that GENIE3 is the only method whose support is cyclic, since it has no mechanism that enforces acyclicity. Supplementary Table 13 reports the complete native-support comparison and the separately labelled post-hoc oracle-threshold diagnostic.

PC3-scale benchmark. Moving from the data-rich $G = 100$ benchmark to the PC3-scale regime makes DAG recovery more challenging, but RISCAL retains informative and balanced structural performance. It achieves an AUPRC of 0.633 ± 0.016 and an F1 of 0.649 ± 0.018, with similar precision (0.640 ± 0.022) and recall (0.659 ± 0.015). Its selected edge set also remains close in size to the ground-truth DAG, containing $14{,}144.0 \pm 290.6$ versus $13{,}732.4 \pm 23.2$ edges.

InterSort-SORTRANKING preserves some ranking signal (AUPRC 0.556 ± 0.008), but its support contains 4.93 million edges, producing high recall with negligible precision. SDCD remains close to the Random DAG baseline. Table 1 summarizes the primary comparison, while Supplementary Table 14 provides the complete results and post-hoc threshold diagnostics. PC3-scale ablations support the model design, while oracle substitutions localize residual error (Appendix G).

Table 1: Causal discovery in the PC3-scale benchmark ($G = 3710$) under fixed support rules. Values are mean $\pm$ standard deviation across five seeds; best means are bold.

Model	AUPRC ↑	Precision ↑	Recall ↑	F1 ↑
RISCAL	$\mathbf{0.633 \pm 0.016}$	$\mathbf{0.640 \pm 0.022}$	0.659 ± 0.015	$\mathbf{0.649 \pm 0.018}$
SDCD	0.001 ± 0.000	0.006 ± 0.004	0.000 ± 0.000	0.000 ± 0.000
InterSort-SORTRANKING	0.556 ± 0.008	0.003 ± 0.000	$\mathbf{0.946 \pm 0.002}$	0.005 ± 0.000
Random DAG baseline	0.001 ± 0.000	0.001 ± 0.000	0.001 ± 0.000	0.001 ± 0.000

7.2 Real Perturbation-Response Prediction

Following the protocol in Section 6, the primary endpoints are perturbation-response R^2 and macro intervention MSE, complemented by profile-level Pearson correlations.

Across the ten cell lines, the axial transformer is the strongest predictive reference, achieving the highest response-profile correlation and the lowest macro intervention MSE. Still, RISCAL, GEARS, and the axial transformer operate in a similar predictive regime, their treated and response correlations differ by less than 0.002, and all achieve macro intervention MSE ≈ 0.0075. PDGrapher remains competitive but shows slightly lower response correlation and higher error, whereas linear regression and the generic MLP perform markedly worse. Values are reported in Supplementary Table 23.

Figure 3 shows the cell-line-specific differences among the four competitive methods. The paired analysis defined in Appendix F confirms that RISCAL outperforms linear regression, the MLP, and PDGrapher on both primary endpoints in all ten lines, although its absolute advantage over PDGrapher is small. Results against GEARS are context dependent: RISCAL performs better on both endpoints in four lines, worse in two, shows no detectable difference in three, and yields a mixed result in one. The axial transformer has advantage in most lines, with HT29 favoring RISCAL and YAPC showing no detectable difference. Thus the result is not that RISCAL is the strongest dedicated predictor, but that learning a DAG preserves essentially the same predictive regime. Complete coverage, per-line metrics, confidence intervals, and corrected tests are reported in Supplementary Tables 22, 24 and 25.

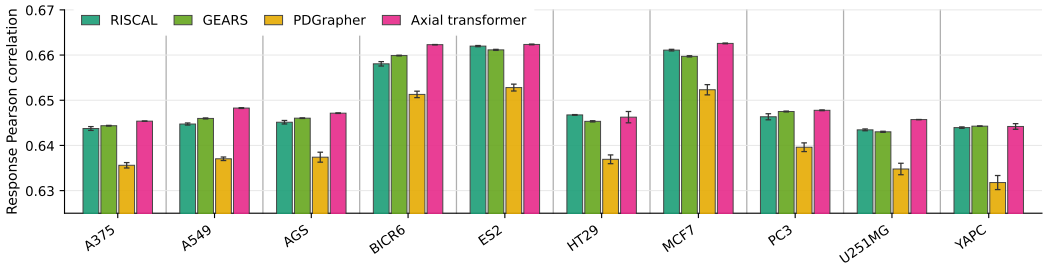

Figure 3: Samplewise response-profile Pearson correlation on each cell line's common held-out test. Bars show the mean $\pm$ one sample standard deviation across five train/validation fits. Linear regression and the MLP are omitted because their mean response correlations are 53.1% and 45.2% below RISCAL, respectively; the complete six-model comparison appears in Supplementary Figure 33.

7.3 Stability and Biological Plausibility of Learned Graphs

Real LINCS profiles do not provide a ground-truth DAG. We therefore seek converging evidence at four levels: reproducibility across trained fits, enrichment for curated directed interactions, functional coherence of the graph communities, and extending the results to an unseen regulator.

Table 2: Directed-edge enrichment.

	CollecTRI		SIGNOR	
Cell line	Enrichment	q	Enrichment	q
A375	3.07	0.002	2.50	0.008
A549	1.52	0.201	3.73	0.002
AGS	4.54	0.002	4.65	0.002
BICR6	3.80	0.002	2.58	0.014
ES2	1.74	0.196	4.18	0.002
HT29	5.46	0.002	2.83	0.017
MCF7	1.97	0.081	3.44	0.006
PC3	4.37	0.002	5.45	0.002
U251MG	4.49	0.002	5.09	0.002
YAPC	2.81	0.006	1.41	0.240

First, the learned structures contain substantial reproducible cores. The 3/5 acyclic consensus graphs contain 5,095-10,271 edges, while the five-fit consensus cores retain, on average, 46.5% of those edges. Complete core sizes and fit-to-fit stability are reported in Supplementary Tables 15 and 16 and Supplementary Figure 22.

Second, these recurrent edges concentrate independently curated directed biology. Relative to 1,000 degree-preserving acyclic null graphs with identical in/out-degree sequences, seven of ten cell lines are significantly enriched for CollecTRI (Müller-Dott et al., 2023) and nine for SIGNOR (Licata et al., 2020); every line is enriched for at least one resource and six for both (Table 2). Because these resources are incomplete and context aggregated, this supports biological plausibility at the graph level rather than edge-wise accuracy.

Moreover, the acyclic consensus graphs organize into functionally coherent communities. Figure 4 illustrates this with a HT29 module comprising 12 genes and 19 edges. The module is strongly enriched for defense response to virus (GO:0051607, $q = 5.30 \times 10^{-11}$).

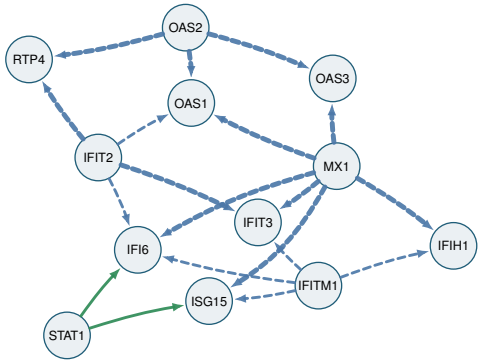

Figure 4: HT29 antiviral-response module learned by the SCM. Green edges are supported by CollecTRI or SIGNOR; thick and thin edges recur in five and four fits, respectively.

Widely, reciprocal matching across 247 modules shows that cell lines share functional programs more strongly than exact gene membership: median functional-profile cosine is 0.881, compared with a median gene-set Jaccard of 0.169. Thus the learned causal DAGs recover recurrent biological programs through partially context-specific gene modules. Diagnostics in Appendix D.3.

Finally, we used an independent LINC00525-knockdown experiment in A549 cells to incorporate this previously unseen regulator into the A549 frozen models, obtaining responses that generalized across held-out replicates (Pearson $r = 0.454 \pm 0.043$). 17 connections recurred across all five models and were enriched for stress, apoptosis, and DNA-damage responses, while intervention propagation indicated suppression of cell-cycle programs, consistent with reported antiproliferative and cell-cycle phenotypes in A549 and colorectal cancer (Fang et al., 2021; Zheng & Peng, 2019); complete results in Appendix H.

8 Conclusion

RISCAL shows that perturbation-response prediction and directed gene-structure learning can be pursued as complementary objectives within a single model. Because real perturbation screens lack complete ground-truth regulatory networks, we additionally introduce a biology-grounded synthetic generator that provides a reusable benchmark for controlled evaluation of causal discovery. Using this benchmark, RISCAL accurately recovers ground-truth DAGs and retains balanced structural recovery in the challenging 3,710-gene regime. On LINCS data, it operates in essentially the same predictive regime as strong prediction-focused models while additionally producing sparse acyclic graphs with reproducible cores, enrichment for curated directed interactions, and functionally coherent modules. Extending the model to the previously unseen regulator LINC00525 further recovered reproducible connections and propagated programs consistent with independent experimental evidence. Together, these results support RISCAL as a framework for turning perturbational expression data into accurate response predictions and interpretable, testable hypotheses about regulatory structure.

Reproducibility Statement

The complete implementation, self-contained experiment recipes, container specifications, data-acquisition instructions, and scripts used to regenerate datasets, results, tables, and figures are available at `https://anonymous.4open.science/r/RISCAL/`. Experimental settings are described in Sections 5 and 6, with detailed data provenance, preprocessing, model optimization, synthetic-data generation, and evaluation procedures provided in Appendices C.1, A.2, B, and F. Experiments were run on a shared server with two Intel Xeon Gold 6130 CPUs (32 physical cores and 64 hardware threads), 750 GiB of system memory, and an NVIDIA RTX PRO 5000 Blackwell GPU with 48 GB of memory. The host ran Ubuntu 26.04 LTS with NVIDIA driver 595.84; the main RISCAL container used Debian 13, Python 3.11.14, PyTorch 2.5.1, and CUDA 12.4. Complete software environments, including those of the baselines, are fixed by the provided lockfile and container specifications.

AI Use Statement

We used generative AI tools to assist with methodological and experimental planning, implementation and revision of analysis and reporting code, and drafting and editing portions of the manuscript. We also used these tools for literature discovery, manuscript organization, and language editing. All AI-assisted code was reviewed and verified by the authors. All numerical results, tables, and figures were generated from the version-controlled experimental and reporting pipeline, rather than supplied by a generative AI tool. The authors reviewed the final text, claims, citations, and artifacts, and take full responsibility for the content of this work.

References

Seungheun Baek, Soyon Park, Yan Ting Chok, Mogan Gim, and Jaewoo Kang. GPO-VAE: Modeling explainable gene perturbation responses utilizing GRN-aligned parameter optimization. *Bioinformatics*, 41(Supplement 1):i599–i608, 2025. doi: 10.1093/bioinformatics/btaf256. URL `https://doi.org/10.1093/bioinformatics/btaf256`.

Shaojie Bai, J. Zico Kolter, and Vladlen Koltun. Deep equilibrium models. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL `https://proceedings.neurips.cc/paper/2019/hash/01386bd6d8e091c2ab4c7c7de644d37b-Abstract.html`.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300, 1995. doi: 10.1111/j.2517-6161.1995.tb02031.x.

Vincent D. Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008 (10):P10008, 2008. doi: 10.1088/1742-5468/2008/10/P10008.

Stephan Bongers, Patrick Forré, Jonas Peters, and Joris M. Mooij. Foundations of structural causal models with cycles and latent variables. *The Annals of Statistics*, 49(5):2885–2915, 2021. doi: 10.1214/21-AOS2064.

Philippe Brouillard, Sébastien Lachapelle, Alexandre Lacoste, Simon Lacoste-Julien, and Alexandre Drouin. Differentiable causal discovery from interventional data. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL `https://proceedings.neurips.cc/paper/2020/hash/f8b7aa3a0d349d9562b424160ad18612-Abstract.html`.

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

Charlotte Bunne, Stefan G. Stark, Gabriele Gut, Jacobo Sarabia del Castillo, Mitch Levesque, Kjong-Van Lehmann, Lucas Pelkmans, Andreas Krause, and Gunnar Rätsch. Learning single-cell perturbation responses using neural optimal transport. *Nature Methods*, 20:1759–1768,

2023. doi: 10.1038/s41592-023-01969-x. URL https://www.nature.com/articles/s41592-023-01969-x.

Zhi-Jie Cao, Tianzi Guo, Cheng Li, Fan Zhang, Ping Lu, Chan Gu, Fan Guo, and Ge Gao. Causality-oriented regulatory inference and rational transcriptomic reprogramming design with CASCADE. *bioRxiv*, 2025. doi: 10.1101/2025.11.20.689377. URL https://doi.org/10.1101/2025.11.20.689377. Preprint.

Mathieu Chevalley, Yusuf H. Roohani, Arash Mehrjou, Jure Leskovec, and Patrick Schwab. A large-scale benchmark for network inference from single-cell perturbation data. *Communications Biology*, 8(1):412, 3 2025a. ISSN 2399-3642. doi: 10.1038/s42003-025-07764-y. URL https://doi.org/10.1038/s42003-025-07764-y.

Mathieu Chevalley, Patrick Schwab, and Arash Mehrjou. Deriving causal order from single-variable interventions: Guarantees & algorithm. In *The Thirteenth International Conference on Learning Representations*, 2025b. URL https://openreview.net/forum?id=u63OVngeSp.

David Maxwell Chickering. Optimal structure identification with greedy search. *Journal of Machine Learning Research*, 3:507–554, 2002.

Haoyue Dai, Peter Spirtes, and Kun Zhang. Independence testing-based approach to causal discovery under measurement error and linear non-gaussian models. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/b05bffeb1ef937677ef0e32f027b4c80-Abstract-Conference.html.

Jesus de la Fuente, Robert Lehmann, Carlos Ruiz-Arenas, Jan Voges, Irene Marin-Goñi, Xabier Martinez-de Morentin, David Gomez-Cabrero, Idoia Ochoa, Jesper Tegner, Vincenzo Lagani, and Mikel Hernaez. Interpretable causal representation learning for biological data in the pathway space, 2025. URL https://arxiv.org/abs/2506.12439. ICLR 2025.

Haitham A. Elmarakeby, Justin Hwang, Rand Arafeh, Jett Crowdis, Sydney Gang, David Liu, Saud H. AlDubayan, Keyan Salari, Steven Kregel, Camden Richter, Taylor E. Arnoff, Jihye Park, William C. Hahn, and Eliezer M. Van Allen. Biologically informed deep neural network for prostate cancer discovery. *Nature*, 598(7880):348–352, 9 2021. ISSN 1476-4687. doi: 10.1038/s41586-021-03922-4. URL https://doi.org/10.1038/s41586-021-03922-4.

Almudena Espín-Pérez, Chris Portier, Marc Chadeau-Hyam, Karin van Veldhoven, Jos C. S. Kleinjans, and Theo M. C. M. de Kok. Comparison of statistical methods and the use of quality control samples for batch effect correction in human transcriptome data. *PLOS ONE*, 13(8):e0202947, 2018. doi: 10.1371/journal.pone.0202947.

Panqi Fang, Hao Chen, Zhifei Ma, Chencheng Han, Wenda Yin, Siwei Wang, Hongyu Zhu, Wenjia Xia, Jie Wang, Lin Xu, Tongyan Liu, and Rong Yin. Lncrna LINC00525 suppresses p21 expression via mrna decay and triplex-mediated changes in chromatin structure in lung adenocarcinoma. *Cancer Communications*, 41(7):596–614, 2021. doi: 10.1002/cac2.12181.

Illés J. Farkas, Chuang Wu, Chakra Chennubhotla, Ivet Bahar, and Zoltán N. Oltvai. The topological basis of signal integration in the transcriptional-regulatory network of the yeast, *Saccharomyces cerevisiae*. *BMC Bioinformatics*, 7:478, 2006. doi: 10.1186/1471-2105-7-478.

Johann A. Gagnon-Bartsch and Terence P. Speed. Using control genes to correct for unwanted variation in microarray data. *Biostatistics*, 13(3):539–552, 2012. doi: 10.1093/biostatistics/kxr034.

Guadalupe Gonzalez, Xiang Lin, Isuru Herath, Kirill Veselkov, Michael Bronstein, and Marinka Zitnik. Combinatorial prediction of therapeutic perturbations using causally inspired neural networks. *Nature Biomedical Engineering*, 10:1008–1025, 2026. doi: 10.1038/s41551-025-01481-x. URL https://www.nature.com/articles/s41551-025-01481-x. Published online 9 September 2025.

Sam F. Greenbury, Iain G. Johnston, Matthew A. Smith, Jonathan P. K. Doye, and Ard A. Louis. The effect of scale-free topology on the robustness and evolvability of genetic regulatory networks. *Journal of Theoretical Biology*, 267(1):48–61, 2010. doi: 10.1016/j.jtbi.2010.08.006.

Alexander Hägele, Jonas Rothfuss, Lars Lorch, Vignesh Ram Somnath, Bernhard Schölkopf, and Andreas Krause. BaCaDI: Bayesian causal discovery with unknown interventions. In *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pp. 1411–1436. PMLR, 2023. doi: 10.48550/arXiv.2206.01665. URL https://proceedings.mlr.press/v206/hagele23a.html.

Leon Hetzel, Simon Boehm, Niki Kilbertus, Stephan Günnemann, Fabian Theis, et al. Predicting cellular responses to novel drug perturbations at a single-cell resolution. *Advances in Neural Information Processing Systems*, 35:26711–26722, 2022.

Vân Anh Huynh-Thu, Alexandre Irrthum, Louis Wehenkel, and Pierre Geurts. Inferring regulatory networks from expression data using tree-based methods. *PLOS ONE*, 5(9):e12776, 2010. doi: 10.1371/journal.pone.0012776. URL https://doi.org/10.1371/journal.pone.0012776.

Yusuke Ito, Satoshi Inoue, Takahiro Nakashima, Haosong Zhang, Yang Li, Hitomi Kasuya, Tetsuya Matsukawa, Zhiwen Wu, Toshiaki Yoshikawa, Mirei Kataoka, et al. Epigenetic profiles guide improved CRISPR/Cas9-mediated gene knockout in human T cells. *Nucleic Acids Research*, 52 (1):141–153, 2024. doi: 10.1093/nar/gkad1076.

Adrián Javaloy, Pablo Sánchez-Martín, and Isabel Valera. Causal normalizing flows: from theory to practice. In *Advances in Neural Information Processing Systems*, volume 36, pp. 58833–58864. Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2023. doi: 10.52202/075280-2566. URL https://doi.org/10.52202/075280-2566.

Marco Jost, Daniel A. Santos, Reuben A. Saunders, Max A. Horlbeck, John S. Hawkins, Sonia M. Scaria, Thomas M. Norman, Jeffrey A. Hussmann, Christina R. Liem, Carol A. Gross, and Jonathan S. Weissman. Titrating gene expression using libraries of systematically attenuated CRISPR guide RNAs. *Nature Biotechnology*, 38(3):355–364, 1 2020. ISSN 1546-1696. doi: 10.1038/s41587-019-0387-5. URL https://doi.org/10.1038/s41587-019-0387-5.

Alexandra B. Keenan, Sherry L. Jenkins, Kathleen M. Jagodnik, Simon Koplev, Edward He, Denis Torre, Zichen Wang, Anders B. Dohlman, Moshe C. Silverstein, Alexander Lachmann, Maxim V. Kuleshov, Avi Ma'ayan, Vasileios Stathias, Raymond Terryn, Daniel Cooper, Michele Forlin, Amar Koleti, Dusica Vidovic, Caty Chung, Stephan C. Schürer, et al. The library of integrated network-based cellular signatures NIH program: System-level cataloging of human cells response to perturbations. *Cell Systems*, 6(1):13–24, 1 2018. ISSN 2405-4712. doi: 10.1016/j.cels.2017.11.001. URL https://doi.org/10.1016/j.cels.2017.11.001.

Sébastien Lachapelle, Philippe Brouillard, Tristan Deleu, and Simon Lacoste-Julien. Gradient-based neural DAG learning. In *International Conference on Learning Representations*, 2020.

Hao-Chih Lee, Matteo Danieletto, Riccardo Miotto, Sarah T. Cherng, and Joel T. Dudley. Scaling structural learning with NO-BEARS to infer causal transcriptome networks. *Pacific Symposium on Biocomputing*, 25:391–402, 2020. URL https://psb.stanford.edu/psb-online/proceedings/psb20/Lee.pdf.

Luana Licata, Piero Lo Surdo, Marta Iannuccelli, Alessandro Palma, Enrico Micarelli, Livia Perfetto, Daniele Peluso, Alberto Calderone, Luisa Castagnoli, and Gianni Cesareni. Signor 2.0, the signaling network open resource 2.0: 2019 update. *Nucleic Acids Research*, 48(D1):D504–D510, 2020. doi: 10.1093/nar/gkz949.

Phillip Lippe, Taco Cohen, and Efstratios Gavves. Efficient neural causal discovery without acyclicity constraints. In *International Conference on Learning Representations (ICLR)*, 2022. doi: 10.48550/arXiv.2107.10483. URL https://openreview.net/forum?id=eYciPrLuUhG.

Romain Lopez, Jan-Christian Hütter, Jonathan K. Pritchard, and Aviv Regev. Large-scale differentiable causal discovery of factor graphs. In *Advances in Neural Information Processing Systems*, volume 35, pp. 19290–19303. Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2022. doi: 10.52202/068431-1402. URL https://doi.org/10.52202/068431-1402.

Mohammad Lotfollahi, Fabian A Wolf, and Fabian J Theis. scgen predicts single-cell perturbation responses. *Nature Methods*, 16:715–721, 2019.

Mohammad Lotfollahi, Sergei Rybakov, Karin Hrovatin, Soroor Hediyeh-zadeh, Carlos Talavera-López, Alexander V. Misharin, and Fabian J. Theis. Biologically informed deep learning to query gene programs in single-cell atlases. *Nature Cell Biology*, 25(2):337–350, 2 2023. ISSN 1476-4679. doi: 10.1038/s41556-022-01072-x. URL https://doi.org/10.1038/s41556-022-01072-x.

Jianzhu Ma, Michael Ku Yu, Samson Fong, Keiichiro Ono, Eric Sage, Barry Demchak, Roded Sharan, and Trey Ideker. Using deep learning to model the hierarchical structure and function of a cell. *Nature Methods*, 15(4):290–298, 3 2018. ISSN 1548-7105. doi: 10.1038/nmeth.4627. URL https://doi.org/10.1038/nmeth.4627.

Marija Milacic, Deidre Beavers, Patrick Conley, Chuqiao Gong, Marc Gillespie, Johannes Griss, Robin Haw, Bijay Jassal, Lisa Matthews, Bruce May, et al. The reactome pathway knowledgebase 2024. *Nucleic Acids Research*, 52(D1):D672–D678, 2024. doi: 10.1093/nar/gkad1025.

Sophia Müller-Dott, Eirini Tsirvouli, Miguel Väzquez, Ricardo O. Ramirez Flores, Pau Badia-i Mompel, Robin Fallegger, Dénes Türei, Yair Larriba, and Julio Saez-Rodriguez. Expanding the coverage of regulons from high-confidence prior knowledge for accurate estimation of transcription factor activities. *Nucleic Acids Research*, 51(20):10934–10949, 2023. doi: 10.1093/nar/gkad841.

Achille Nazaret, Justin Hong, Elham Azizi, and David Blei. Stable differentiable causal discovery. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 37413–37445. PMLR, 2024. URL https://proceedings.mlr.press/v235/nazaret24a.html.

M. E. J. Newman. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, 103(23):8577–8582, 2006. doi: 10.1073/pnas.0601602103.

Ignavier Ng, AmirEmad Ghassami, and Kun Zhang. On the role of sparsity and DAG constraints for learning linear DAGs. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/d04d42cdf14579cd294e5079e0745411-Abstract.html.

Nick Pawlowski, Daniel Coelho de Castro, and Ben Glocker. Deep structural causal models for tractable counterfactual inference. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/0987b8b338d6c90bbedd8631bc499221-Abstract.html.

Carlos Rodríguez-Caso, Bernat Corominas-Murtra, and Ricard V. Solé. On the basic computational structure of gene regulatory networks. *Molecular BioSystems*, 5(12):1617–1629, 2009. doi: 10.1039/B904960F.

Martin Rohbeck, Brian Clarke, Katharina Mikulik, Alexandra Pettet, Oliver Stegle, and Kai Ueltzhöffer. Bicycle: Intervention-based causal discovery with cycles. In *Proceedings of the Third Conference on Causal Learning and Reasoning*, volume 236 of *Proceedings of Machine Learning Research*, pp. 209–242. PMLR, 2024. URL https://proceedings.mlr.press/v236/rohbeck24a.html.

Yusuf Roohani, Kexin Huang, and Jure Leskovec. Predicting transcriptional outcomes of novel multigene perturbations with gears. *Nature biotechnology*, 42(6):927–935, 2024.

Nora Schneider, Lars Lorch, Niki Kilbertus, Bernhard Schölkopf, and Andreas Krause. Generative intervention models for causal perturbation modeling. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 53388–53412. PMLR, 2025. URL https://proceedings.mlr.press/v267/schneider25a.html.

Ian Smith, Peyton G. Greenside, Ted Natoli, David L. Lahr, David Wadden, Itay Tirosh, Rajiv Narayan, David E. Root, Todd R. Golub, Aravind Subramanian, and John G. Doench. Evaluation of RNAi and CRISPR technologies by large-scale gene expression profiling in the connectivity map. *PLOS Biology*, 15(11):e2003213, 2017. doi: 10.1371/journal.pbio.2003213.

Arne H. Smits, Frederik Ziebell, Gerard Joberty, Nico Zinn, William F. Mueller, Sandra Clauder-M"unster, Dirk Eberhard, Maria F"alth Savitski, Paola Grandi, Petra Jakob, et al. Biological plasticity rescues target activity in CRISPR knock outs. *Nature Methods*, 16:1087–1093, 2019. doi: 10.1038/s41592-019-0614-5.

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

Aravind Subramanian, Rajiv Narayan, Steven M. Corsello, David D. Peck, Ted E. Natoli, Xiaodong Lu, Joshua Gould, John F. Davis, Andrew A. Tubelli, Jacob K. Asiedu, David L. Lahr, Jodi E. Hirschman, Zihan Liu, Melanie Donahue, Bina Julian, Mariya Khan, David Wadden, Ian C. Smith, Daniel Lam, Arthur Liberzon, Courtney Toder, Mukta Bagul, Marek Orzechowski, Oana M. Enache, Federica Piccioni, Sarah A. Johnson, Nicholas J. Lyons, Alice H. Berger, Alykhan F. Shamji, Angela N. Brooks, Anita Vrcic, Corey Flynn, Jacqueline Rosains, David Y. Takeda, Roger Hu, Desiree Davison, Justin Lamb, Kristin Ardlie, Larson Hogstrom, Peyton Greenside, Nathanael S. Gray, Paul A. Clemons, Serena Silver, Xiaoyun Wu, Wen-Ning Zhao, Willis Read-Button, Xiaohua Wu, Stephen J. Haggarty, Lucienne V. Ronco, Jesse S. Boehm, Stuart L. Schreiber, John G. Doench, Joshua A. Bittker, David E. Root, Bang Wong, and Todd R. Golub. A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. *Cell*, 171(6): 1437–1452.e17, 11 2017. ISSN 0092-8674. doi: 10.1016/j.cell.2017.10.049. URL https://doi.org/10.1016/j.cell.2017.10.049.

Kazuhiro Takemoto and Chikoo Oosawa. Modeling for evolving biological networks with scale-free connectivity, hierarchical modularity, and disassortativity. *Mathematical Biosciences*, 208(2): 454–468, 2007. doi: 10.1016/j.mbs.2006.11.002.

The Gene Ontology Consortium. The gene ontology knowledgebase in 2023. *Genetics*, 224(1): iyad031, 2023. doi: 10.1093/genetics/iyad031.

Panagiotis Tigas, Yashas Annadani, Andrew Jesson, Bernhard Schölkopf, Yarin Gal, and Stefan Bauer. Interventions, where and how? experimental design for causal models at scale. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/98a5c0470e57d518ade4e56c6ee0b363-Abstract-Conference.html.

Ramon Viñas Torné, Sílvia Fàbregas Salazar, Soyon Park, Ivo Alexander Ban, Artyom Gadetsky, Nikita Doikov, and Maria Brbić. PACER: Acyclic causal discovery from large-scale interventional data. In *International Conference on Machine Learning (ICML)*, 2026. doi: 10.48550/arXiv.2605.15353. URL https://arxiv.org/abs/2605.15353.

Ying Wang, Yujing Zhai, Mingzhe Zhang, Chunlin Song, Yuqing Zhang, and Gang Zhang. Escaping from CRISPR–Cas-mediated knockout: The facts, mechanisms, and applications. *Cellular & Molecular Biology Letters*, 29:48, 2024. doi: 10.1186/s11658-024-00565-x.

Ethan Weinberger, Ryan Conrad, and Tal Ashuach. Modeling variable guide efficiency in pooled CRISPR screens with ContrastiveVI+. In *NeurIPS 2024 Workshop on AI for New Drug Modalities*, 2024. URL https://openreview.net/forum?id=Swe1i5VVaz.

Menghua Wu, Umesh Padia, Sean H Murphy, Regina Barzilay, and Tommi Jaakkola. Identifying biological perturbation targets through causal differential networks. *arXiv preprint arXiv:2410.03380*, 2024.

Albert Xue, Jingyou Rao, Sriram Sankararaman, and Harold Pimentel. dotears: Scalable and consistent directed acyclic graph estimation using observational and interventional data. *iScience*, 28(2):111673, 2025. doi: 10.1016/j.isci.2024.111673. URL https://doi.org/10.1016/j.isci.2024.111673.

Karren Yang, Abigail Katcoff, and Caroline Uhler. Characterizing and learning equivalence classes of causal DAGs under interventions. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 5541–5550. PMLR, 2018. URL https://proceedings.mlr.press/v80/yang18a.html.

Ying Yu, Naixin Zhang, Yuanbang Mai, Luyao Ren, Qiaochu Chen, Zehui Cao, Qingwang Chen, Yaqing Liu, Wanwan Hou, Jingcheng Yang, Huixiao Hong, Joshua Xu, Weida Tong, Lianhua Dong, Xiang Fang, Leming Shi, and Yuanting Zheng. Correcting batch effects in large-scale multiomics studies using a reference-material-based ratio method. *Genome Biology*, 24:201, 2023. doi: 10.1186/s13059-023-03047-z.

Yue Yu, Xun Zheng, Anima Anandkumar, and Yisong Yue. Dag-gnn: Dag structure learning with graph neural networks. In *International Conference on Machine Learning (ICML)*, pp. 7154–7163, 2019.

Jiaqi Zhang, Kristjan Greenewald, Chandler Squires, Akash Srivastava, Karthikeyan Shanmugam, and Caroline Uhler. Identifiability guarantees for causal disentanglement from soft interventions. In *Advances in Neural Information Processing Systems*, volume 36, pp. 50254–50292. Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2023. doi: 10.52202/075280-2186. URL https://doi.org/10.52202/075280-2186.

Q. Zheng and P. Peng. Long non-coding rna LINC00525 is an unfavorable prognostic factor and regulates proliferation and apoptosis in colorectal cancer. *Annals of Oncology*, 30(Supplement 4): iv8, 2019. doi: 10.1093/annonc/mdz155.027. Conference abstract P-028.

Xun Zheng, Bryon Aragam, Pradeep K Ravikumar, and Eric P Xing. Dags with no tears: Continuous optimization for structure learning. *Advances in neural information processing systems*, 31, 2018.

Appendix

Table of Contents

A Model Implementation and Optimization. 17

A.1 Intervention-Uncertainty Propagation and Numerical Safeguards 17

A.2 Joint Optimization and Hyperparameters 17

A.3 Training Algorithm 19

A.4 Threshold, Convergence, Runtime, and Memory Diagnostics 20

B Progressive Synthetic Benchmark 21

B.1 Benchmark Design and Reporting Protocol 21

B.2 Identifiability and Intervention Semantics-Stages 00-02 23

B.3 Measurement and Noise Realism-Stages 03-07 27

B.4 Imperfect Loss-of-Function-Stages 08-09 32

B.5 Biological Complexity and Scaling-Stages 10-13 36

C LINCS Data Provenance and Preprocessing 42

C.1 Dataset Provenance and Cohort Construction 42

C.2 Batch Correction and Contrast Preservation 43

D Causal Structure Evaluation and Biological Plausibility Analysis 46

D.1 Causal Baseline Implementation and Synthetic Recovery 46

D.2 Causal Benchmark Results 46

D.3 Real-Graph Stability and Biological Enrichment 46

E Post-Intervention Prediction Benchmarking 54

E.1 Common protocol and prediction baselines implementation 54

E.2 Complete fixed-test predictive results 56

F Evaluation Metrics and Statistical Analysis 59

F.1 Causal-discovery metrics 59

F.2 Perturbation-response prediction metrics 59

F.3 Achievable synthetic R^2 60

F.4 Residual-activity diagnostics 60

F.5 Exogenous-regulator case-study metrics 60

F.6 Aggregation and paired statistical inference 61

F.7 Biological plausibility metrics 61

G Model Ablations and Oracle Substitutions 63

G.1 Training-objective ablations. 63

G.2 Architectural-component ablations 63

G.3 Edge-control ablations 64

G.4 Oracle substitutions and error decomposition 65

H Exogenous-regulator Case Study: LINC00525 in A549 66

H.1 Question and external perturbation data 66

H.2 Integration into frozen A549 models 66

H.3 Held-out generalization and new connections 66

H.4 Propagated programs and experimental concordance 67

A Model Implementation and Optimization

A.1 Intervention-Uncertainty Propagation and Numerical Safeguards

For the residual-activity model $r_k \sim \mathrm{Beta}(\rho_k\phi, (1-\rho_k)\phi)$,

$$\mathbb{E}[r_k] = \rho_k, \qquad \mathrm{Var}(r_k) = \frac{\rho_k(1-\rho_k)}{\phi+1}. \tag{12}$$

The target mechanism also has its own SCM noise. For sample n and target gene k, let $f^{\mathrm{nat}}_{n,k} = \mu_k + f_{\theta,k}(z^t_{n,\setminus k}; A^{(g)})$ denote the target's natural conditional mean on the measured scale and let σ^2_k be its SCM noise variance. Assuming that r_k and the zero-mean target noise are independent, the variance of the attenuated target source is

$$\omega_{n,k} = \mathrm{Var}[r_k]\left(f^{\mathrm{nat}}_{n,k}\right)^2 + \mathbb{E}[r_k^2]\,\sigma^2_k, \qquad \mathbb{E}[r_k^2] = \mathrm{Var}[r_k] + \rho_k^2. \tag{13}$$

We propagate this source variance locally through the analytic Jacobian-vector product of the child-specific mechanisms, evaluated at the teacher-forced treated profile,

$$\nu_{n,j} = \omega_{n,k}\left(\left.\frac{\partial f_{\theta,j}(z; A^{(g)})}{\partial z_k}\right|_{z=\tilde{z}^t_n}\right)^2, \tag{14}$$

which is available in closed form because each mechanism is a single $\tanh$ layer. The one-hot intervention direction computes all local derivatives $\partial f_{\theta,j}(z; A^{(g)})/\partial z_k$ in one Jacobian-vector product (JVP). When $A^{(g)}$ is replaced by $A^{(b)}$ for official evaluation, the derivative is zero for mechanisms that do not directly use target k. In the variance path this JVP factor is detached from automatic differentiation: its numerical value routes uncertainty, but the graph and mechanism weights do not receive gradients through that factor solely to enlarge the tolerated variance. The term $\nu_{n,j}$ is added only to the gene-specific variance of the SCM NLL. It does not alter the deterministic fixed-point mean or predictive variance. This is a local first-order approximation, not recursive multistep uncertainty propagation through all descendants; the Gaussian log-variance term still penalizes arbitrary variance inflation.

Numerical safeguards. For numerical stability, learned log-variances are clipped to $[-12, 12]$, and the resulting total variances to $[e^{-12}, e^{12}]$, before evaluating either Gaussian likelihood. Masked losses are normalized by the number of valid sample-gene entries. The SCM negative log-likelihood includes the propagated variance $\nu_{n,j}$, whereas the target-expression readout likelihood uses $\nu_{n,j} = 0$; their learned log-variances are separate.

A.2 Joint Optimization and Hyperparameters

The final protocol uses one joint trajectory per fit. The child-specific weights U, v simultaneously define mechanism predictions and path-strength connectivity; nuisance and residual-activity parameters are updated in the same trajectory. No graph is sampled and no temperature is annealed. When enabled by a recipe, the deterministic path-strength gate uses a fixed temperature τ, gate center q_{gate}, and evaluation threshold q_{eval} throughout that trajectory.

The residual-activity continuation factor c_t first restricts the model to the shared mean activity and then releases gene-specific deviations. This reduces the risk that gene-specific residual activity prematurely absorbs variation that should inform the causal structure and child-specific mechanisms. In the final PC3-scale and real-data recipes, $c_t = 0$ for 20 epochs, increases linearly to one over the following 40 epochs, and remains one thereafter.

Supplementary Table 3 reports the final recipes for the stages of the synthetic progression defined in Appendix B. A row with $10, 100$ denotes identical settings at both synthetic dimensions.

Table 3: Optimization, path-strength gate, and acyclicity settings. Init/inc/max denote initialization/increment/maximum; mult denotes multiplier. Gate dashes indicate no soft gate. Stage 10 covers all three topologies; Stage 13 also applies to real-data fits.

Stage	G	Gate center	Evaluation threshold	Gate temperature	$\lambda_t^{\mathrm{acyc}}$ init	Fixed inc/max	μ_t^{acyc} init	μ_t^{acyc} mult/max	Start/every	Progress
Stage 00	10, 100	-	0.5	-	0.005	5e-05/0.025	0	1/0	0/0	0
Stage 01	10, 100	-	0.125	-	0.005	5e-05/0.025	0	1/0	0/0	0
Stage 02	10, 100	-	0.125	-	0.005	5e-05/0.025	0	1/0	0/0	0
Stage 03	10, 100	-	0.125	-	0.001	1e-05/0.005	0	1/0	0/0	0
Stage 04	10, 100	-	0.125	-	0.001	1e-05/0.005	0	1/0	0/0	0
Stage 05	10, 100	-	0.125	-	0.001	1e-05/0.005	0	1/0	0/0	0
Stage 06	10, 100	-	0.125	-	0.001	0.0003/0.1	0	1/0	0/0	0
Stage 07	10, 100	-	0.125	-	0.05	0/0	0.001	2/0.08	50/10	0.95
Stage 08	10, 100	-	0.125	-	0.001	0/0	0.25	3/2.25	10/5	0.9
Stage 09	10, 100	-	0.125	-	0.001	0/0	0.25	3/2.25	10/5	0.9
Stage 10	10, 100	-	0.125	-	0.001	0/0	0.25	3/2.25	10/5	0.9
Stage 11	10	0.025	0.025	0.05	0.001	0/0	0.25	3/100	10/5	0.9
Stage 11	100	0.025	0.025	0.01	0.001	0/0	0.25	3/100	10/5	0.9
Stage 12	10	0.025	0.025	0.05	0.001	0/0	0.25	3/20.25	10/5	0.9
Stage 12	100	0.025	0.025	0.01	0.001	0/0	0.25	3/20.25	10/5	0.9
Stage 13	3710	0.025	0.09	0.05	0.01	0/0	0.01	2/200	5/5	0.9

A.2.1 ACYCLICITY-COEFFICIENT SCHEDULES

The acyclicity contribution in Eq. (7) is written uniformly as

$$\lambda_t^{\mathrm{acyc}}\widehat{h}(A) + \frac{\mu_t^{\mathrm{acyc}}}{2}\widehat{h}(A)^2. \tag{15}$$

Supplementary Table 3 reports the initialization and update schedule of both coefficients for every resolved recipe.

When a fixed increment Δ_λ and maximum $\lambda_{\max}$ are specified,

$$\lambda_t^{\mathrm{acyc}} = \min\{\lambda_1^{\mathrm{acyc}} + (t-1)\Delta_\lambda, \lambda_{\max}\}, \qquad \mu_t^{\mathrm{acyc}} = 0. \tag{16}$$

These recipes retain a nonzero linear acyclicity penalty while omitting the quadratic component. A multiplier of one together with zero maximum, start, interval, and progress entries indicates that no adaptive update is applied.

For recipes with adaptive updates, the fixed increment is zero and $\lambda_t^{\mathrm{acyc}}$ and μ_t^{acyc} follow the augmented-Lagrangian schedule. Let e_0, m, r, κ_μ, and $\mu_{\max}$ denote the declared activation epoch, update interval, progress ratio, multiplicative increase, and maximum quadratic-penalty scale. Updates occur after epochs $t \geq e_0$ with $t \bmod m = 0$. At such an epoch, let $h_t = \widehat{h}_{\mathrm{val}}(A_t)$ be the detached continuous acyclicity diagnostic and let h_{t^-} be its value at the preceding update. Progress is adequate when $h_t \leq r h_{t^-}$. At the first update no preceding value exists, so μ_t^{acyc} is retained; thereafter,

$$\mu_{t+1}^{\mathrm{acyc}} = \begin{cases} \min\{\kappa_\mu \mu_t^{\mathrm{acyc}}, \mu_{\max}\}, & h_t > r h_{t^-}, \\ \mu_t^{\mathrm{acyc}}, & h_t \leq r h_{t^-}, \end{cases} \qquad \lambda_{t+1}^{\mathrm{acyc}} = \lambda_t^{\mathrm{acyc}} + \mu_{t+1}^{\mathrm{acyc}} h_t. \tag{17}$$

At all other epochs both quantities remain unchanged. This validation diagnostic updates only the training penalty state; it is excluded from checkpoint ranking and test evaluation.

A.2.2 DAG CHECKPOINT SELECTION CONTRACT

For B_{val} validation minibatches, the mask-normalized SCM loss is

$$\overline{\mathcal{L}}_{\mathrm{SCM,val}}(A^{(b)}) = \frac{1}{B_{\mathrm{val}}} \sum_{m=1}^{B_{\mathrm{val}}} \mathcal{L}_{\mathrm{SCM}}^{(m)}(A^{(b)}). \tag{18}$$

The fixed recipe threshold defines $A^{(b)} = \mathbf{1}[A \geq q_{\mathrm{eval}}]$ with zero diagonal. Eq. (11) ranks only acyclic candidates using this SCM average and the binary-support penalty. The minimum-score feasible DAG is stored in `checkpoint_causal_graph.pt` and `checkpoint_best.pt`; if none exists, the latter stores an explicitly labelled `non_dag_fallback` with no causal test result. Early stopping begins only after selection is active and an eligible DAG has appeared. Validation selects checkpoints; test profiles and synthetic ground truth do not enter training or selection.

A.2.3 Parameter Groups, Optimizer, and Scheduler

The nonlinear implementation exposes three disjoint code-level groups:

Graph. The input and output weights U, v. They jointly determine the nonlinear mechanisms and A; there is no independent mechanism group.

Nuisance. The learned expression baseline, target-expression readout, and separate SCM and predictive log-variance vectors.

Residual activity. The global mean logit, centered gene-specific deviations, and raw hierarchical-Beta concentration.

Adam is constructed once per fit. The residual-activity learning rate inherits the nuisance rate when unspecified. The current final implementation has no separate mechanism weight-decay setting.

After every epoch, ReduceLROnPlateau receives the continuous causal validation score with training-only acyclicity coefficients disabled. The code-level graph group uses `min_lr_graph`; the other groups use their configured learning rates as their minima. Consequently, scheduler reductions affect only U, v and do not select the official checkpoint.

A.3 Training Algorithm

Joint path-strength SCM training and DAG checkpoint selection

Input: $\mathcal{D}_{\mathrm{tr}}, \mathcal{D}_{\mathrm{val}}$; $\xi = (\{U_j, v_j\}_{j=1}^{G}$, nuisance, residual-activity variables); $q_{\mathrm{gate}}, q_{\mathrm{eval}}, \tau, Q, \delta, t_{\min}$; resolved objective, acyclicity, residual-activity continuation, and stopping schedules.
Initialize: Adam parameter groups, scheduler, and the recipe-defined acyclicity state $(\lambda^{\mathrm{acyc}}, \mu^{\mathrm{acyc}})$.
for epoch t **do**
  Neural path strengths and soft gate matrix: $A_{ij} \leftarrow \sum_h |v_{jh}|\,|U_{jhi}|$ $(i \neq j)$, $A_{ii} \leftarrow 0$, $A^{(g)}_{ij} \leftarrow \sigma((A_{ij} - q_{\mathrm{gate}})/\tau)$, $A^{(g)}_{ii} \leftarrow 0$. Refresh detached Perron vectors p_L, p_R for $\widehat{h}(A)$.
  for minibatch $\mathcal{B} = (x^d, x^t, u) \subset \mathcal{D}_{\mathrm{tr}}$ with target k **do**
    Residual moments: $c_t \leftarrow$ recipe-defined residual-activity continuation factor, $\widetilde{\eta} \leftarrow \eta - \mathrm{mean}(\eta)$, $\rho_k \leftarrow \sigma(a + c_t\widetilde{\eta}_k)$, $\phi \leftarrow \mathrm{softplus}(\kappa)$, $\mathrm{Var}(r_k) \leftarrow \rho_k(1-\rho_k)/(\phi+1)$.
    Interventional mechanism fit: construct $\check{x}^t$ by Eq. (8); $\check{z}^t \leftarrow \check{x}^t - \mu$; $\omega_k \leftarrow \mathrm{Var}(r_k)(f^{\mathrm{nat}}_k)^2 + [\mathrm{Var}(r_k) + \rho_k^2]\sigma_k^2$; $\nu_j \leftarrow \omega_k \left[\partial f_{\theta,j}/\partial z_k\big|_{\check{z}^t}\right]^2$; $\mathcal{L}_{\mathrm{SCM}} \leftarrow \mathrm{NLL}_{j\neq k}(\check{x}^t, \mu + f_\theta(\check{z}^t; A^{(g)}); \sigma_j^2 + \nu_j)$.
    Damped fixed-point iteration: $\bar{u} \leftarrow \mathbf{1} - u$, $z^d \leftarrow x^d - \mu$, $x^{(0)} \leftarrow \bar{u} \odot x^d + u \odot \rho_k[\mu_k + f_{\theta,k}(z^d; A^{(g)})]$.
    for $\ell = 0$ **to** $Q - 1$ **do**
      $z^{(\ell)} \leftarrow x^{(\ell)} - \mu$, $n^{(\ell)} \leftarrow \mu + f_\theta(z^{(\ell)}; A^{(g)})$, $\pi^{(\ell)} \leftarrow u \odot \rho_k n^{(\ell)}$, $\tilde{x}^{(\ell)} \leftarrow \bar{u} \odot n^{(\ell)} + \pi^{(\ell)}$;
      $x^{(\ell+1)} \leftarrow \bar{u} \odot [(1-\delta)x^{(\ell)} + \delta\tilde{x}^{(\ell)}] + \pi^{(\ell)}$.
    end for
    Target-expression readout: $\hat{x}^t \leftarrow \bar{u} \odot x^{(Q)} + u \odot (s \odot x^d + b)$; $\mathcal{L}_{\mathrm{target}} \leftarrow \mathrm{NLL}_u(x^t, \hat{x}^t)$.
    Objective and update: form Eq. (7) from $\mathcal{L}_{\mathrm{SCM}}$, $\mathcal{L}_{\mathrm{target}}$, $\mathcal{R}_{\mathrm{sparse}}(A)$, and $\widehat{h}(A)$; update $\xi \leftarrow \mathrm{Adam}(\nabla_\xi \mathcal{L})$.
  end for
  Validation: update ReduceLROnPlateau and the training-only Augmented-Lagrangian acyclicity state from the continuous validation diagnostic.
  if $t \geq t_{\min}$ **then**
    Candidate DAG: $A^{(b)}_t \leftarrow \mathbf{1}[A_t \geq q_{\mathrm{eval}}]$, $\mathrm{diag}(A^{(b)}_t) \leftarrow 0$.
    if $A^{(b)}_t$ is a DAG **then**
      evaluate Eq. (11); retain the minimum-score feasible candidate.
    else

Retain only the labelled non-DAG continuous fallback.
end if
end if
Stop if the recipe-defined patience is exhausted after an eligible DAG has been observed.
end for
Return: the best feasible checkpoint, or the labelled continuous fallback; $(A, A^{(b)})$ and all fitted parameters.

Algorithm 1: Joint training, acyclicity updates, and validation-based selection of evaluation-threshold DAG checkpoints.

A.4 Threshold, Convergence, Runtime, and Memory Diagnostics

Post-training diagnostics use the five PC3-scale benchmark runs defined in Section 5. They do not alter training, checkpoint selection, or the selected edge set.

Evaluation-threshold sensitivity. The fixed $q_{\text{eval}} = 0.09$ threshold converts the path-strength matrix of each validation-selected checkpoint into its reported support. It is distinct from the soft gate used during optimization. Supplementary Figure 5 varies this threshold post-training against the synthetic ground-truth DAG; the per-seed F1-maximizing values are diagnostic only.

Training and checkpoint-selection dynamics. Supplementary Figure 6 shows the training objective and the causal-validation score over eligible DAG checkpoints. Because the quantities have different scales, each seed-specific series maps its first relevant value to one and its attained minimum to zero before aggregation.

ixed-point convergence and computational footprint. Prediction uses $Q = 25$ damped updates with $\delta = 0.5$. Across the five selected checkpoints, the held-out final-update diagnostic $\|x^{(Q)} - x^{(Q-1)}\|_1/G$ is $(6.45 \pm 0.63) \times 10^{-8}$. Supplementary Table 4 reports convergence, runtime, and peak GPU memory.

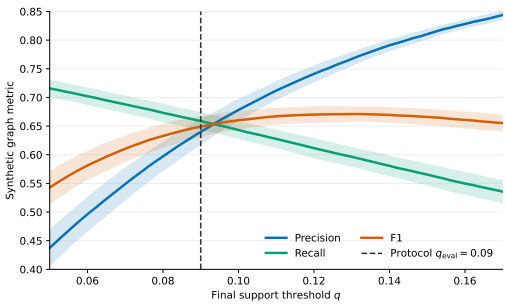

Figure 5: Post-training threshold sensitivity across five PC3-scale seeds. Curves show mean $\pm$ sample standard deviation; the dashed line marks $q_{\text{eval}} = 0.09$.

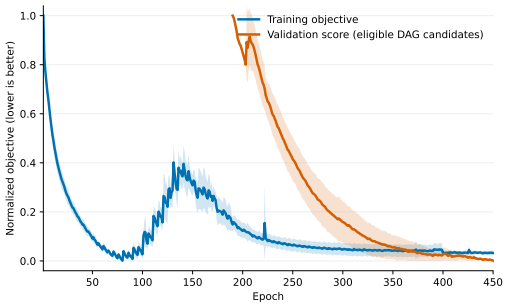

Figure 6: Normalized training objective and causal-validation score over eligible DAG checkpoints. Lines and bands show mean $\pm$ sample standard deviation across five seeds.

Table 4: Post-training technical diagnostics for the official five-seed PC3-scale benchmark cohort. The final fixed-point update is $\|x^{(Q)} - x^{(Q-1)}\|_1/G$ on held-out profiles. Runtime is descriptive because GPU resources were shared.

Cohort	Seeds	q_{eval}	Q	Damping δ	DAGs / 5	Final fixed-point update	s / epoch	Peak GPU GiB
PC3-scale benchmark	5	0.09	25	0.5	5/5	$(6.45 \pm 0.63) \times 10^{-8}$	45.9 ± 1.0	4.39 ± 0.00

B PROGRESSIVE SYNTHETIC BENCHMARK

B.1 BENCHMARK DESIGN AND REPORTING PROTOCOL

Supplementary Table 5 provides a compact Stage-by-Stage map of the progression and its generator-model contract; the corresponding Stage subsections provide the detailed biological and identifiability rationale. Supplementary Tables 6 and 7 report the official five-seed aggregates at $G = 10$ and $G = 100$, respectively. Supplementary Table 8 records residual-activity distributional calibration on a predeclared generator-derived cohort.

The official selected edge set is the evaluation-threshold edge set of the selected DAG checkpoint. Known ground-truth DAGs are used only after selection; no synthetic graph, structural metric, or achievable-response quantity enters training, checkpoint selection, threshold selection, or hyperparameter selection. Metric definitions, aggregation, and the Monte Carlo achievable-R^2 protocol are given in Appendix F.

Table 5: Compact map of the progressive synthetic benchmark, from an identifiable directed acyclic graph (DAG) anchor to PC3-like Library of Integrated Network-based Cellular Signatures (LINCS)-informed conditions. Detailed generator, model, biological, and identifiability rationale is provided in the corresponding Stage subsections.

Stage(s)	Generator	Model	Identifiability	Prior Work
Stage 00: minimal benchmark	Linear DAG; hard interventions; paired noise	Linear SCM; spectral acyclicity	Identifiable under the stated assumptions with all-target intervention coverage	**Direct:** DCDI, dotears, NOBEARS, SDCD Brouillard et al. (2020); Xue et al. (2025); Lee et al. (2020); Nazaret et al. (2024); **Related:** PDGrapher, DeepSCM Gonzalez et al. (2026); Pawlowski et al. (2020)
Stage 01: signed effects	Signed, nonzero-bounded edge weights	Unchanged real-valued edge matrix	Preserved under the stated nonzero-effect assumptions	**Direct:** DCDI, GIM Brouillard et al. (2020); Schneider et al. (2025)
Stage 02: effective vs. measured	Separate effective activity and measured expression	LoF-aware SCM; target-expression readout	Requires a detectable descendant intervention signal	**Related:** Dai et al. Dai et al. (2022)
Stage 03: target Laplace noise	Target-only paired Laplace noise	Exclude target structural residual	Target exclusion preserves the structural score; finite-sample prediction is harder	**Related:** Dai et al. Dai et al. (2022)
Stage 04: gene-specific readout	Gene-specific affine/Laplace target-expression readout	Gene-specific affine target-expression readouts	Separated from structural equations	**Related:** Dai et al. Dai et al. (2022)
Stage 05: measured scale	Baselines; log-scale effective activity	SEM intercepts; measured-space clamp	No intended change to the population identifiability assumptions	**None found**
Stages 06-07: pair noise	Gene-specific σ_j; pairwise ρ_{pair}	Learn ρ_{pair} with Gaussian NLL	Harder finite-sample recovery	**Direct:** DCDI, dotears Brouillard et al. (2020); Xue et al. (2025); **Related:** DeepSCM Pawlowski et al. (2020)
Stages 08-09: imperfect LoF	Retain treated-natural target activity	Learn gene-specific retention moments	Needs detectable intervention shifts	**Related:** DCDI, GIM, CRISPR escape Brouillard et al. (2020); Schneider et al. (2025); Smits et al. (2019)
Stage 10: biological topology	Hub-modular or scale-free DAGs	Topology robustness under the reduced objective	Hubs increase finite-sample difficulty	**Direct:** Tigas et al., GIM, PDGrapher, Inspre Tigas et al. (2022); Schneider et al. (2025); Gonzalez et al. (2026); Brown et al. (2025)
Stage 11: nonlinear mechanisms	Centered, child-specific masked MLPs	Path-strength matrix and soft gate matrix; Gaussian NLL; damped fixed-point iteration	$\mathcal{I}^*$-faithful detectable effects	**Direct:** DCDI, GIM Brouillard et al. (2020); Schneider et al. (2025)
Stage 12: reused controls	PC3-scaled pool; repeated baselines	Retain $\rho_{\text{pair}} = 0$; grouped splits	Preserved under target-independent reuse; lower effective sample size	**None found**
Stage 13: PC3-like scale	Fewer samples; up to 3,710 genes	No planned model change	Finite-sample recovery dominates	**Direct:** SDCD, Inspre Nazaret et al. (2024); Brown et al. (2025)

Table 6: Synthetic progression across Stages 00-12 at $G = 10$, aggregated across five dataset seeds. Entries are mean $\pm$ sample standard deviation, except DAGs / 5. Metric definitions and aggregation conventions are provided in Appendix F. Runtime per epoch is descriptive because GPU resources were shared.

Stage	DAGs / 5	AUPRC	Precision	Recall	F1	Pred. / ground-truth-edge ratio	Treated-profile R^2	$R^2_{\max}$	$R^2 / R^2_{\max}$	Runtime (s / epoch)
Stage 00	5/5	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	9.60 ± 0.20
Stage 01	5/5	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	9.65 ± 0.20
Stage 02	5/5	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	9.56 ± 0.14
Stage 03	5/5	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	0.984 ± 0.003	0.984 ± 0.003	1.000 ± 0.000	9.94 ± 0.11
Stage 04	5/5	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	0.984 ± 0.002	0.984 ± 0.002	1.000 ± 0.000	9.82 ± 0.25
Stage 05	5/5	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	0.998 ± 0.001	0.998 ± 0.001	1.000 ± 0.000	10.86 ± 0.25
Stage 06 ($\rho_{\text{pair}} = 1$)	5/5	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	0.998 ± 0.001	0.998 ± 0.001	1.000 ± 0.000	6.49 ± 0.13
Stage 07 ($\rho_{\text{pair}} = 0$)	5/5	1.000 ± 0.000	0.953 ± 0.073	1.000 ± 0.000	0.975 ± 0.040	1.054 ± 0.087	0.982 ± 0.022	0.987 ± 0.013	0.994 ± 0.010	6.51 ± 0.10
Stage 08	5/5	1.000 ± 0.000	0.989 ± 0.024	1.000 ± 0.000	0.995 ± 0.012	1.011 ± 0.025	0.943 ± 0.031	0.945 ± 0.033	0.999 ± 0.003	6.77 ± 0.19
Stage 09	5/5	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	0.942 ± 0.034	0.941 ± 0.034	1.001 ± 0.001	7.05 ± 0.05
Stage 10 (scale-free)	5/5	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	0.956 ± 0.017	0.955 ± 0.017	1.000 ± 0.000	7.02 ± 0.15
Stage 10 (hub-modular)	5/5	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	0.936 ± 0.044	0.935 ± 0.045	1.001 ± 0.001	7.48 ± 0.59
Stage 10 (Erdos-Renyi)	5/5	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	0.942 ± 0.034	0.941 ± 0.034	1.001 ± 0.001	16.21 ± 0.18
Stage 11	5/5	0.972 ± 0.023	0.932 ± 0.045	0.945 ± 0.043	0.938 ± 0.033	1.016 ± 0.067	0.823 ± 0.114	0.827 ± 0.108	0.994 ± 0.011	60.13 ± 5.21
Stage 12	5/5	0.972 ± 0.026	0.925 ± 0.049	0.946 ± 0.063	0.934 ± 0.038	1.025 ± 0.095	0.823 ± 0.115	0.827 ± 0.109	0.995 ± 0.010	67.02 ± 9.97

Table 7: Synthetic progression across Stages 00-12 at $G = 100$, aggregated across five dataset seeds. Entries are mean $\pm$ sample standard deviation, except DAGs / 5. Metric definitions and aggregation conventions are provided in Appendix F. Runtime per epoch is descriptive because GPU resources were shared.

Stage	DAGs / 5	AUPRC	Precision	Recall	F1	Pred. / ground-truth-edge ratio	Treated-profile R^2	$R^2_{\max}$	$R^2 / R^2_{\max}$	Runtime (s / epoch)
Stage 00	5/5	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	57.31 ± 4.85
Stage 01	5/5	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	57.23 ± 4.96
Stage 02	5/5	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	56.66 ± 1.30
Stage 03	5/5	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	0.999 ± 0.000	0.999 ± 0.000	1.000 ± 0.000	57.89 ± 4.91
Stage 04	5/5	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	0.999 ± 0.000	0.999 ± 0.000	1.000 ± 0.000	58.14 ± 4.61
Stage 05	5/5	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	58.38 ± 4.18
Stage 06 ($\rho_{\text{pair}} = 1$)	5/5	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	18.83 ± 0.48
Stage 07 ($\rho_{\text{pair}} = 0$)	5/5	1.000 ± 0.000	0.999 ± 0.003	1.000 ± 0.000	0.999 ± 0.001	1.001 ± 0.003	0.941 ± 0.009	0.950 ± 0.007	0.991 ± 0.010	14.11 ± 3.04
Stage 08	5/5	0.989 ± 0.003	0.987 ± 0.007	0.987 ± 0.005	0.987 ± 0.004	1.000 ± 0.008	0.790 ± 0.043	0.855 ± 0.015	0.924 ± 0.050	19.05 ± 0.94
Stage 09	5/5	0.994 ± 0.003	0.979 ± 0.011	0.991 ± 0.004	0.985 ± 0.007	1.013 ± 0.010	0.796 ± 0.047	0.854 ± 0.016	0.931 ± 0.050	18.81 ± 0.77
Stage 10 (scale-free)	5/5	0.991 ± 0.005	0.983 ± 0.007	0.988 ± 0.006	0.985 ± 0.005	1.005 ± 0.010	0.797 ± 0.077	0.874 ± 0.054	0.911 ± 0.044	18.97 ± 1.04
Stage 10 (hub-modular)	5/5	0.997 ± 0.006	0.969 ± 0.027	0.996 ± 0.008	0.982 ± 0.017	1.028 ± 0.023	0.819 ± 0.028	0.846 ± 0.029	0.968 ± 0.027	18.28 ± 0.59
Stage 10 (Erdos-Renyi)	5/5	0.994 ± 0.003	0.979 ± 0.011	0.991 ± 0.004	0.985 ± 0.007	1.013 ± 0.010	0.796 ± 0.047	0.854 ± 0.016	0.931 ± 0.050	18.74 ± 0.47
Stage 11	5/5	0.853 ± 0.023	0.763 ± 0.036	0.848 ± 0.023	0.803 ± 0.017	1.113 ± 0.074	0.450 ± 0.044	0.503 ± 0.052	0.897 ± 0.046	59.52 ± 5.81
Stage 12	5/5	0.850 ± 0.020	0.778 ± 0.030	0.851 ± 0.020	0.812 ± 0.014	1.095 ± 0.061	0.451 ± 0.042	0.502 ± 0.053	0.899 ± 0.036	57.78 ± 8.04

Table 8: Distributional calibration of residual activity across five dataset seeds. Stages 08-12 use $G = 100$ and Stage 13 uses $G = 3710$. The predeclared evaluation cohort contains all genes with at least one true direct child and is independent of the predicted graph. Each paired entry reports learned / true values. Definitions of the residual-activity diagnostics are provided in Appendix F.

Stage	True parent genes	$\hat{\mu}$ / μ true	SD across genes, pred. / true	$\hat{c}$ / c true	Mean Beta variance, pred. / true
Stage 08	81.4 ± 2.1	0.382 ± 0.006 / 0.400 ± 0.000	0.000 ± 0.000 / 0.000 ± 0.000	-	-
Stage 09	81.4 ± 2.1	0.363 ± 0.008 / 0.404 ± 0.002	0.056 ± 0.004 / 0.034 ± 0.003	29.607 ± 0.222 / 30.000 ± 0.000	0.007 ± 0.000 / 0.008 ± 0.000
Stage 10 (scale-free)	47.8 ± 5.0	0.365 ± 0.005 / 0.402 ± 0.003	0.053 ± 0.003 / 0.034 ± 0.004	29.524 ± 0.272 / 30.000 ± 0.000	0.008 ± 0.000 / 0.008 ± 0.000
Stage 10 (hub-modular)	79.4 ± 3.2	0.356 ± 0.008 / 0.402 ± 0.003	0.057 ± 0.003 / 0.034 ± 0.003	29.704 ± 0.377 / 30.000 ± 0.000	0.007 ± 0.000 / 0.008 ± 0.000
Stage 10 (Erdos-Renyi)	81.4 ± 2.1	0.363 ± 0.008 / 0.404 ± 0.002	0.056 ± 0.004 / 0.034 ± 0.003	29.607 ± 0.222 / 30.000 ± 0.000	0.007 ± 0.000 / 0.008 ± 0.000
Stage 11	47.8 ± 5.0	0.284 ± 0.051 / 0.402 ± 0.003	0.248 ± 0.029 / 0.034 ± 0.004	18.963 ± 4.219 / 30.000 ± 0.000	0.007 ± 0.001 / 0.008 ± 0.000
Stage 12	47.8 ± 5.0	0.289 ± 0.047 / 0.402 ± 0.003	0.252 ± 0.032 / 0.034 ± 0.004	15.132 ± 6.463 / 30.000 ± 0.000	0.010 ± 0.003 / 0.008 ± 0.000
Stage 13 (PC3-like)	1641.6 ± 24.7	0.496 ± 0.007 / 0.400 ± 0.000	0.007 ± 0.001 / 0.033 ± 0.001	29.878 ± 0.015 / 30.000 ± 0.000	0.008 ± 0.000 / 0.008 ± 0.000

The purpose is to build a controlled ladder of structural causal models (SCMs) in which each new generator feature breaks one explicit simplifying assumption, and model changes are justified when that broken assumption requires it. This makes the model's real-data interpretation more falsifiable: failure to recover the ground-truth DAG in a synthetic regime that satisfies the stated identifiability assumptions indicates a limitation due to model specification, optimization, sample size, or assumption violations before any biological interpretation is attempted.

B.2 Identifiability and Intervention Semantics-Stages 00-02

B.2.1 Stage 00: Identifiability Anchor and Minimal Generator

For these theoretical identifiability statements only, let d be the number of nodes, $[d] = \{1, \dots, d\}$ their index set, $X = (X_1, \dots, X_d)$ the measured gene-expression vector, and $G^* = ([d], E^*)$ the ground-truth DAG with edge set E^*. This local use of d is distinct from the gene count G used for reported synthetic and real-data cohorts. Let

$$X_j = \sum_{i \in \mathrm{Pa}_{G^*}(j)} B^*_{ij} X_i + \varepsilon_j, \qquad \varepsilon_j \sim \mathcal{N}(0, \sigma^2), \qquad B^*_{ij} = 0 \text{ if } i \not\to j. \tag{19}$$

The binary generating adjacency A^* and signed coefficient matrix B^* use the orientation

$$A^*_{ij} = 1 \iff i \to j, \qquad B^*_{ij} \text{ is the coefficient associated with } i \to j. \tag{20}$$

With observational data alone, the best one can generally identify under Markov and faithfulness assumptions is an observational Markov equivalence class. DCDI states this directly and motivates interventional data as a way to shrink the equivalence class Brouillard et al. (2020). In our setting the intervention family is

$$\mathcal{I}^* = (\emptyset, \{1\}, \dots, \{d\}), \tag{21}$$

Here $\emptyset$ is the observational, or baseline-control, regime and each singleton $\{k\}$ is a known intervention on one target gene. The observational regime is therefore complemented with a known singleton intervention on every gene. Each intervention replaces one target mechanism while leaving its outgoing causal effects observable, which provides the orientation signal required to distinguish the true DAG. DCDI's Theorem 1 states that exact score maximization recovers a DAG in the correct $\mathcal{I}^*$-Markov equivalence class under the following conditions Brouillard et al. (2020):

1. The interventional family contains the observational regime, $I^*_1 = \emptyset$.
2. The density model has enough capacity to represent the ground-truth interventional distributions (Assumption 1 (Sufficient capacity) DCDI).
3. $\mathcal{I}^*$-faithfulness holds, "the intervention actually change something" (Assumption 2 (I*-Faithfulness) DCDI).
4. The model densities are strictly positive (Assumption 3: Strict positivity).
5. The ground-truth densities have finite differential entropy (Assumption 4: Finite differential entropies).
6. The sparsity coefficient λ is small enough that the true DAG is not outscored by an overly sparse graph.

DCDI's Corollary 11 gives the key statement, if every node is individually targeted, as in Eq. (21), the true DAG is alone in its $\mathcal{I}$-Markov equivalence class. Under the theorem assumptions, exact recovery is therefore possible.

The important condition for the generator is $\mathcal{I}^*$-faithfulness, in DCDI, this has two parts. The first is ordinary observational faithfulness, conditional independences in the observational distribution must correspond to d-separations in the ground-truth DAG. The second rules out pathological interventions: if an intervention can affect a variable, the relevant interventional conditional or marginal distribution must actually change. In generator terms, the target mechanism must change, and descendants that should be affected must show distributional changes. Interventions that leave the target and all descendants unchanged contain no identifying signal.

Dotears adds a complementary linear-SEM perspective: it assumes known single-node hard interventions on all nodes, a linear SEM, sub-Gaussian samples, and sufficient samples per regime, and proves consistency of its weighted-DAG estimator in the large-sample limit (Xue et al., 2025). It also makes two practical points that matter for this benchmark: edge effects must not be arbitrarily close to zero, and interventional data should enter the loss rather than being used only to estimate nuisance quantities. This justifies the generator's finite-sample lower bound $\beta_{\min}$:

$$B^*_{ij} \sim \text{Unif}\left([-1, -\beta_{\min}] \cup [\beta_{\min}, 1]\right) \quad \text{for } i \to j. \tag{22}$$

The same issue appears in NOTEARS, whose recovery degrades as edge scales become too small relative to noise (Zheng et al., 2018).

Minimal Identifiable Generator The starting generator samples a DAG by drawing a random topological order and only allowing edges from earlier to later nodes. This guarantees acyclicity by construction. A sample has the form

$$(x^d, u_k, x^t), \tag{23}$$

where x^d is the observational control expression profile, u_k is a one-hot target indicator, and x^t is the profile generated after intervention on target k.

For the baseline we use linear SEM,

$$x^d_j = \sum_{i \in \text{Pa}(j)} B^*_{ij} x^d_i + \varepsilon^d_j, \tag{24}$$

$$x^t_k = 0, \tag{25}$$

$$x^t_j = \sum_{i \in \text{Pa}(j)} B^*_{ij} x^t_i + \varepsilon^t_j, \qquad j \neq k. \tag{26}$$

The paired prediction experiments (until Gene-specific noise scales pass) use shared noise,

$$\varepsilon^t = \varepsilon^d, \tag{27}$$

so that the treated expression profile is a counterfactual transformation of the same exogenous state. This is stronger than the usual independent setting, but it is useful for testing whether a model can both recover the DAG and predict paired treated outcomes.

Supplementary Figure 7 illustrates an eight-node Erdős-Rényi graph where the target gene X_4 has two parents and two children. Under actual settings, a perfect intervention replaces its structural equation, so its incoming mechanisms are cut while its outgoing causal edges remain present.

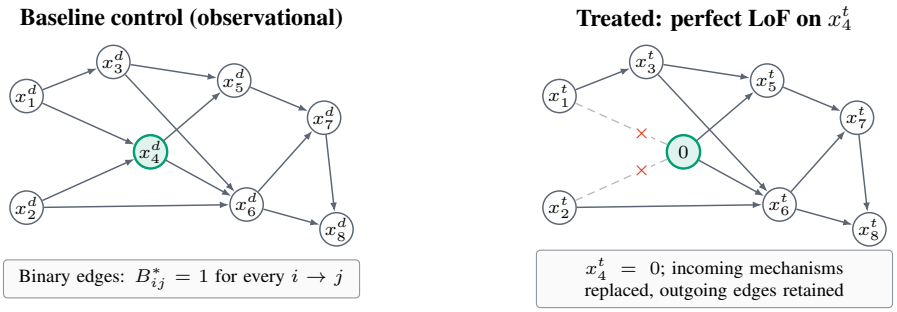

Figure 7: Minimal paired structural causal model (SCM). The crosses denote the two parent mechanisms replaced by the perfect intervention.

The generator is intentionally over-controlled. It is not a biological simulator yet, it is an identifiability control, if the model fails here, then the failure cannot be blamed on nonlinear mechanisms, latent confounding, noisy LoF readouts, or insufficient target coverage.

The Minimal Shared-Noise SCM Model The first model should be no more expressive than the generator. It learns a single real-valued matrix W with $W_{ij} = i \to j$ and zero diagonal. The sample-specific exogenous noise is inferred from the control expression profile:

$$\widehat{\varepsilon}^d = x^d - x^d W. \tag{28}$$

For a one-hot intervention vector u_k, define the non-target mask $m = 1 - u_k$ and $M = \mathrm{diag}(m)$. The treated prediction is the solution of

$$\widehat{x}^t = \widehat{x}^t W M + \widehat{\varepsilon}^d M, \qquad \widehat{x}^t_k = 0. \tag{29}$$

The implementation solves this as a linear system,

$$\left(I - (WM)^\top\right)(\widehat{x}^t)^\top = (\widehat{\varepsilon}^d M)^\top, \tag{30}$$

This keeps the forward pass aligned with the linear SEM and removes an unnecessary numerical approximation.

The early linear stages use the reduced objective specialized to a weighted adjacency W:

$$\mathcal{L} = \alpha_{\text{causal}}\left[\alpha_{\text{SCM}}\mathcal{L}_{\text{SCM}} + \lambda\|W\|_1 + \lambda_e^{\text{acyc}} h_{\text{spec}}(A_\theta) + \tfrac{1}{2}\mu_e^{\text{acyc}} h_{\text{spec}}(A_\theta)^2\right] + \alpha_{\text{target}}\mathcal{L}_{\text{target}}. \tag{31}$$

Here $A_\theta = |W|$, $\lambda\|W\|_1$ is the entrywise sparsity penalty, and the target-prediction coefficient is zero before the separate target-readout stage. For a cohort of N samples, let I_n be the singleton intervention target for sample n. The SCM NLL is

$$\mathcal{L}_{\text{SCM}} = \frac{1}{\sum_n (G - |I_n|)} \sum_{n=1}^{N} \sum_{j \notin I_n} \left(x^{t,\text{eff}}_{n,j} - \sum_i W_{ij} x^{t,\text{eff}}_{n,i} \right)^2. \tag{32}$$

The effective target activity is used here rather than the measured target readout. Excluding the intervened coordinate follows the standard perfect-intervention likelihood treatment: the target conditional is not used to fit the causal graph (Brouillard et al., 2020). The corresponding least-squares structural score is familiar from linear SEM discovery (Zheng et al., 2018; Xue et al., 2025).

Acyclicity is enforced through the spectral-radius constraint introduced for scalable structural learning by NO-BEARS and subsequently characterized as stable by SDCD Lee et al. (2020); Nazaret et al. (2024). Let $A_\theta \in \mathbb{R}^{G\times G}_{\geq 0}$ denote the non-negative structural-strength matrix: $A_\theta = |W|$ for the linear weighted model and $A_\theta = A$ for the nonlinear RISCAL model. We define

$$h_{\text{spec}}(A_\theta) = \mathrm{spr}(A_\theta) = \max_i |\lambda_i(A_\theta)|, \qquad h_{\text{spec}}(A_\theta) = 0 \iff \mathrm{supp}(A_\theta) \text{ is a DAG}. \tag{33}$$

The equivalence follows because a non-negative adjacency matrix represents a DAG if and only if it is nilpotent, and therefore all its eigenvalues are zero. Unlike trace constraints based on matrix

power series, whose values and gradients can vanish or explode with graph dimension and edge scale, SDCD proves that the spectral constraint remains stable under these changes. It also avoids explicitly evaluating a matrix exponential.

We approximate the dominant left and right Perron vectors, w_e and z_e, at the beginning of epoch e using 20 power iterations initialized from the positive uniform vector. The vectors are estimated without constructing an autograd graph and are held fixed across the minibatches of that epoch. The differentiable quotient

$$\widehat{h}_{\mathrm{spec}}(A_\theta) = \frac{w_e^\top A_\theta z_e}{w_e^\top z_e}, \tag{34}$$

provides the gradient with respect to the path-strength matrix. For fixed iteration count K, this requires $O(KG^2)$ dense computation rather than the $O(G^3)$ cost of matrix-exponential evaluation. The exact equivalence in Eq. (33) concerns the true spectral radius; the finite-iteration quotient is an optimization surrogate and does not itself certify a DAG. For the final nonlinear model, feasibility is checked separately on the deterministic support $A^{(b)} = \mathbf{1}[A \geq q_{\mathrm{eval}}]$, with q_{eval} fixed in path-strength units by the resolved recipe; only acyclic candidates are eligible for the official causal checkpoint.

The spectral constraint enters Eq. (7) through the recipe-specific coefficients $\lambda_t^{\mathrm{acyc}}$ and μ_t^{acyc}. Their fixed or adaptive schedules are defined in Appendix A.2.1 and reported for every stage and dimension in Supplementary Table 3. No acyclicity-coefficient update uses synthetic ground truth or test metrics.

Stage 00 result. The minimal shared-noise SCM recovers all five ground-truth DAGs exactly at both $G = 10$ and $G = 100$ (Supplementary Tables 6 and 7), establishing the intended identifiable anchor before biological complications are introduced.

B.2.2 Stage 01: From Adjacency to Signed Effects

The next step replaces binary edges with signed continuous weights:

$$B^*_{ij} \sim \mathrm{Unif}\left([-1, -\beta_{\min}] \cup [\beta_{\min}, 1]\right), \qquad \beta_{\min} = 0.25. \tag{35}$$

For the weighted linear stages, the resolved recipe fixes the reported support threshold at $q_{\mathrm{eval}} = 0.5\,\beta_{\min} = 0.125$. This preserves a margin below the minimum generating edge magnitude. Each present edge has a signed continuous coefficient; in Supplementary Figure 8, arrow width represents absolute effect strength, while colour and the explicit sign distinguish activating from repressing relationships.

This does not require a more complex model because the minimal SCM already learns real-valued W.

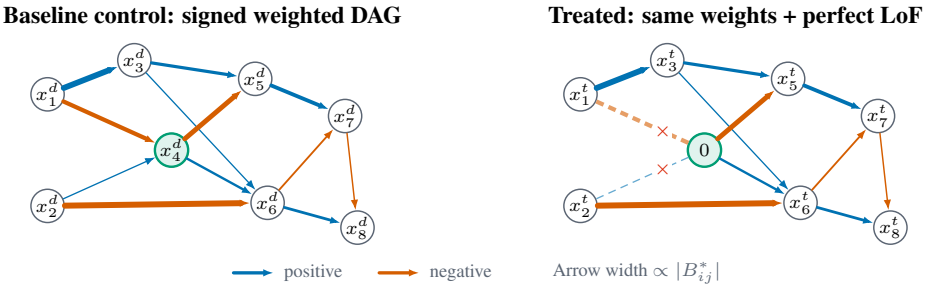

Figure 8: Blue and orange arrows denote positive and negative direct effects, respectively, while arrow width encodes the absolute coefficient magnitude $|B^*_{ij}|$.

This step is important for explainability: binary recovery says only that the model found an arrow. Signed recovery asks whether an up-regulating or down-regulating direct effect is learned and whether false positives remain small enough to be separated by thresholding. The finite lower bound $\beta_{\min}$ excludes arbitrarily weak effects from this controlled setting.

Stage 01 result. Allowing signed nonzero effects preserves exact five-seed recovery at both reported dimensions (Supplementary Tables 6 and 7).

B.2.3 Stage 02: Functional LoF from Measured Expression

A perfect knockout on measured expression, $x_k^t = 0$, is too clean for CRISPR loss-of-function (LoF) data. A biological LoF can eliminate or reduce functional activity without forcing the measured mRNA value to zero. The generator therefore introduces a measured/effective split:

$$z_j = \text{effective activity}, \qquad x_j = \text{measured expression}. \tag{36}$$

For an intervention on target k,

$$z_k^t = z_{\text{LoF}} = 0, \qquad x_k^t = x_k^d \tag{37}$$

The intervention now acts on the effective activity z_k, whereas the measured target expression x_k is produced by a separate readout. Children receive the clamped effective value, so nonzero measured mRNA cannot be mistaken for retained functional activity.

In the first functional-LoF benchmark stage, children receive $z_k^t = 0$, not the measured target value x_k^t. For non-target nodes, the measured expression and effective activity values still coincide.

This generator change justifies that the model must no longer apply the SCM directly to the measured target value. It predicts an effective treated activity profile and then maps the intervened target to measured expression:

$$\widehat{x}_k^{t,\text{eff}} = \text{control/effective clamp}, \tag{38}$$

$$\widehat{x}_k^t = g_k(x_k^d), \tag{39}$$

$$\widehat{x}_j^t = \widehat{x}_j^{t,\text{eff}}, \qquad j \neq k. \tag{40}$$

The minimal version uses $g_k(x_k^d) = x_k^d$. This preserves the interpretation of W as the graph of direct effects between effective activities, while a separate target-expression readout explains the measured target.

Supplementary Figure 9 illustrates this separation between the effective causal intervention and the target-expression readout.

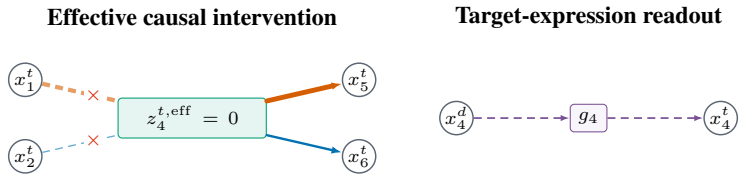

Figure 9: Functional-LoF measured/effective split. The perfect intervention replaces the two parent mechanisms of the effective target activity, $z_4^{t,\text{eff}} = 0$, while its signed outgoing effects remain part of the causal DAG. The measured target x_4^t is generated through the separate readout $g_4(x_4^d)$ and is therefore not propagated to the children as functional activity.

Stage 02 result. Separating effective loss from the measured target retains exact DAG recovery and treated prediction at both dimensions (Supplementary Tables 6 and 7).

B.3 Measurement and Noise Realism-Stages 03-07

B.3.1 Stage 03: Target-Expression Noise with a Global Laplace Delta

Real LINCS LoF pairs show that the measured target changes around the paired control expression value rather than remaining exactly equal to it. To ground this readout in a real paired perturbation setting without averaging over heterogeneous cell lines, Stages 03-13 use a frozen calibration artifact extracted from corrected PC3 profiles. PC3 is the declared reference line for this progression; the same extraction procedure could instead be applied to any other corrected cell line.

For the target delta relative to the paired control expression profile,

$$\Delta_k = x_k^t - x_k^d, \tag{41}$$

this controlled setting uses a heavy-tailed Laplace target-expression readout:

$$x_k^t = x_k^d + \delta, \qquad \delta \sim \mathrm{Laplace}(\ell_{\mathrm{PC3}}, b_{\mathrm{PC3}})\,. \tag{42}$$

The artifact computes ℓ_{PC3} and b_{PC3} directly from paired PC3 target deltas; no manually copied calibration constant enters the generator. The effective intervention and the causal DAG remain unchanged. The new component is confined to the measured target: a global Laplace perturbation replaces the deterministic copy and introduces heavy-tailed measurement variability without modifying the functional intervention.

Supplementary Figure 10 illustrates the resulting stochastic target-expression readout.



Figure 10: Global Laplace target-expression readout. The measured target is generated as $x_k^t = x_k^d + \delta$, with $\delta \sim \mathrm{Laplace}(\ell_{\mathrm{PC3}}, b_{\mathrm{PC3}})$. This adds irreducible heavy-tailed measurement variability without modifying the effective intervention or the signed causal DAG.

For the model, in the previous functional-LoF setting, the measured target was copied from the paired control expression profile,

$$x_k^t = x_k^d. \tag{43}$$

The measured target expression is no longer a deterministic copy of the control expression value. We therefore replace the copy operation by a gene-specific affine target-expression readout:

$$\widehat{x}_k^t = \widehat{a}_k x_k^d + \widehat{b}_k, \tag{44}$$

where the implementation initializes

$$\widehat{a}_k = 1, \qquad \widehat{b}_k = 0, \tag{45}$$

Since the random draw δ is not observable before prediction, the model cannot learn the sample-specific noise realization. Under an MSE prediction loss, the best it can learn is the conditional mean of the measured target:

$$\mathbb{E}[x_k^t \mid x_k^d] = x_k^d + \mathbb{E}[\delta]. \tag{46}$$

For a Laplace distribution with location ℓ_{PC3}, this gives

$$\mathbb{E}[x_k^t \mid x_k^d] = x_k^d + \ell_{\mathrm{PC3}}. \tag{47}$$

Therefore, in this global-delta benchmark stage, the expected solution is approximately

$$\widehat{a}_k \approx 1, \qquad \widehat{b}_k \approx \ell_{\mathrm{PC3}}, \tag{48}$$

Stage 03 result. The global Laplace target-expression readout leaves structural recovery exact; the small reduction in treated-profile R^2 is expected because the added target noise is not deterministic (Supplementary Tables 6 and 7).

B.3.2 Stage 04: Gene-Specific Target-Expression Head

A global delta distribution is still too simple: paired PC3 target deltas vary by target gene.

The frozen PC3 calibration artifact records an empirical target-delta mean for every gene. Stage 04 replaces the shared location with the selected target gene's PC3-specific location while retaining the common PC3 Laplace scale:

$$x_k^t = x_k^d + \delta_k, \qquad \delta_k \sim \mathrm{Laplace}\big(\bar{\delta}_k^{\mathrm{PC3}}, b_{\mathrm{PC3}}\big)\,. \tag{49}$$

The generator therefore replaces the shared location by the PC3-specific mean target delta of the selected target gene while retaining the common PC3 Laplace scale. The model continues to fit its target-expression readout with gene-specific $\widehat{a}_k$ and $\widehat{b}_k$ parameters.

Supplementary Figure 11 illustrates how the readout parameters depend on target identity while remaining separate from effective causal activity.

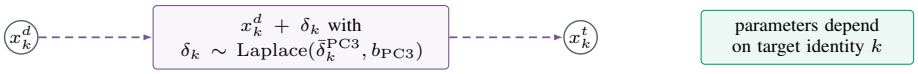

Figure 11: Gene-specific target-expression readout. The selected intervention target determines the PC3-calibrated Laplace location $\bar{\delta}_k^{\text{PC3}}$, while the artifact supplies the common scale b_{PC3}. This readout describes the observed target only; the effective target activity transmitted to descendants remains separate.

Stage 04 result. Gene-specific target-expression readouts preserve exact five-seed structural recovery, showing that target-specific measured noise need not enter the causal support (Supplementary Tables 6 and 7).

B.3.3 Stage 05: Measured Scale, Baselines, and SEM Intercepts

The centered SEM can generate negative values, real LINCS Level 3 profiles used here are in $\log_2(x+1)$ scale, where observed expression is positive. Thus, we change the generator to separate centered expression profile from measured expression:

$$z_j = \sum_i B_{ij}^* z_i + \varepsilon_j, \tag{50}$$

$$x_j = \mu_j + z_j, \tag{51}$$

Gene-specific baselines move the centered SCM into the positive measured expression scale, $x_j = \mu_j + z_j$. Consequently, zero effective measured activity corresponds to $z_k = -\mu_k$, while the separately generated target-expression readout may remain nonzero.

Here, μ_j is sampled with replacement from the gene-specific expression-baseline medians stored in the frozen PC3 calibration artifact.

This baseline shift also changes the interpretation of a perfect functional LoF. In the previous centered generator, setting $z_k^t = 0$ was a natural clamp. However, $z_j = 0$ corresponds to baseline expression, $x_j = \mu_j$, not to zero effective activity. Therefore, a perfect functional LoF is defined by the paired measured- and centered-scale assignments

$$\begin{aligned} x_k^{t,\text{eff}} &= 0, \\ z_k^t &= x_k^{t,\text{eff}} - \mu_k = -\mu_k. \end{aligned} \tag{52}$$

Children of the target gene receive this effective value, while the measured target expression can still be generated by a separate readout model. Thus, for an intervention on k,

$$z_k^{t,\text{eff}} = -\mu_k, \tag{53}$$

$$x_k^t = g_k(x_k^d), \tag{54}$$

$$x_j^t = \mu_j + z_j^t, \qquad j \neq k. \tag{55}$$

This preserves the biological interpretation of LoF; the target gene may still have nonzero measured mRNA expression, but its functional activity passed to its children is zero.

The model must not receive the generator's μ_j, because such privileged information is unavailable for biological DAG discovery. Instead, now it learns SEM intercepts:

$$x_j = b_j + \sum_i W_{ij} x_i + \varepsilon_j. \tag{56}$$

If Eq. (51) holds, then the equivalent measured-scale intercept is

$$b_j = \mu_j - \sum_i W_{ij} \mu_i. \tag{57}$$

Operationally, the model works in the measured scale, the learned expression baselines replace the unknown generator baselines, so the model does not need access to μ_j. For a perfect functional LoF on target k, the model uses two different quantities:

$$\widehat{x}_k^{t,\mathrm{eff}} = 0, \tag{58}$$

$$\widehat{x}_k^{t} = g_k(x_k^d). \tag{59}$$

The first quantity, $\widehat{x}_k^{t,\mathrm{eff}}$, is the value sent to the children of k and used in the SCM residual. The second, $\widehat{x}_k^t$, is the measured mRNA prediction for the target gene. Thus, the target can still have nonzero measured expression, while its functional activity in the causal system is zero.

Supplementary Figure 12 summarizes the relation between centered generator coordinates and the learned measured-scale SCM.

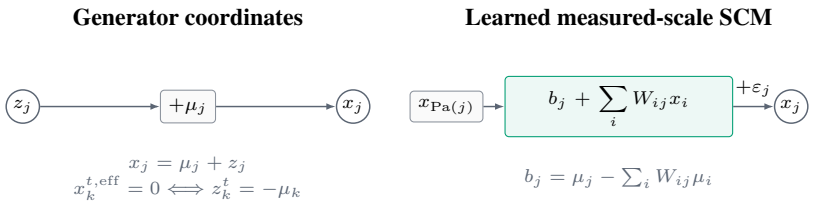

Figure 12: Transition from the centered causal coordinates z_j to measured expression $x_j = \mu_j + z_j$. A perfect functional LoF corresponds to $z_k^t = -\mu_k$, not to $z_k^t = 0$. Because the generator baselines μ_j are unavailable to the model, their measured-scale effect is represented through learned SEM intercepts b_j.

Stage 05 result. Moving to measured-scale baselines and learned expression baselines likewise retains exact DAG recovery at both dimensions (Supplementary Tables 6 and 7).

B.3.4 Stages 06-07: Gene-Specific Noise and Independent Treated Pair Noise

The next biological difficulty is heteroscedasticity: different genes have different residual variability.

Stage 06: Gene-Specific Noise Scales First, we change the generator to sample a gene-specific scale shape from the gene-specific PC3 noise-scale pool stored in the frozen calibration artifact and rescale it to a chosen median:

$$\sigma_j = \sigma_{\mathrm{median}} \frac{\tilde{\sigma}_j^{\mathrm{proxy}}}{\mathrm{median}(\sigma_{\mathrm{pool}}^{\mathrm{proxy}})}. \tag{60}$$

The graph, mechanisms, and intervention remain fixed, but every gene now has its own exogenous scale σ_j. In Supplementary Figure 13 unequal noise halos emphasize that heteroscedasticity changes finite-sample difficulty without adding or removing causal edges.

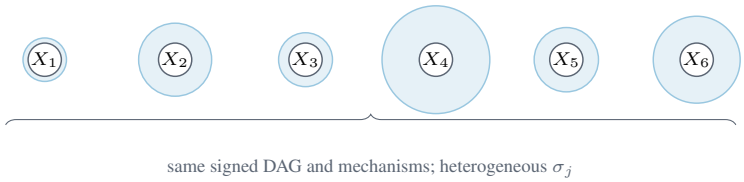

Figure 13: Gene-specific exogenous noise scales. Halo radius represents σ_j: genes may have markedly different residual variability even though the signed causal graph, child-specific mechanisms, and functional-LoF intervention remain unchanged.

For Stage 06, the frozen PC3 artifact supplies the relative gene-specific noise profile, which is rescaled to the declared median $\sigma_{\mathrm{median}} = 0.1$; the pair noise remains shared:

$$\varepsilon_{s,j}^t = \varepsilon_{s,j}^d \sim \mathcal{N}(0, \sigma_j^2). \tag{61}$$

At this stage, the causal mechanism and the measured/effective LoF split are kept unchanged. Stage 06 therefore tests finite-sample recovery under gene-specific noise scales; it does not establish identifiability without the separate population assumptions stated below.

Stage 07: Independent Treated Pair Noise To state the paired-noise family, let the treated residual retain a fraction ρ_{pair} of its paired control-profile residual:

$$\varepsilon^t_{s,j} = \rho_{\text{pair}}\varepsilon^d_{s,j} + \sqrt{1-\rho^2_{\text{pair}}}\,\eta_{s,j}, \qquad \eta_{s,j} \sim \mathcal{N}(0, \sigma^2_j). \tag{62}$$

The official Stage 07 generator evaluates the independent endpoint $\rho_{\text{pair}} = 0$; Stage 06 used the shared-noise endpoint $\rho_{\text{pair}} = 1$. Shared exogenous noise is a related counterfactual construction (Pawlowski et al., 2020); no intermediate correlations are reported. Supplementary Figure 14 illustrates the inherited and independently resampled components of the treated residual.

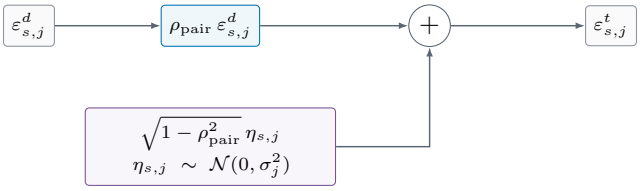

Figure 14: Treated residuals combine a predictable component inherited from the paired control expression profile and an independent innovation. ρ_{pair} controls how much of the control expression profile's exogenous noise is retained in its paired treated sample.

The deterministic model cannot predict $\eta_{s,j}$ from x^d_s. Under MSE, the best it can learn is the conditional mean:

$$\mathbb{E}[\varepsilon^t_{s,j} \mid \varepsilon^d_{s,j}] = \rho_{\text{pair}}\varepsilon^d_{s,j}. \tag{63}$$

Therefore, even with the correct graph and the correct ρ_{pair}, the model cannot reconstruct the sample-specific component $\eta_{s,j}$:

$$\widehat{x}^t_{s,j} \approx \mathbb{E}[x^t_{s,j} \mid x^d_s, I_s]. \tag{64}$$

The remaining error is part of the irreducible noise introduced by the independent treated pair-noise mechanism. From Stage 07 onward, the SCM fit therefore uses a Gaussian NLL over non-target coordinates,

$$\mathcal{L}_{\text{NLL}}(m) = \frac{1}{2\sum_j m_j}\sum_j m_j\left[\frac{(x^t_j - \widehat{x}^t_j)^2}{\widehat{\sigma}^2_j} + \log\widehat{\sigma}^2_j + \log(2\pi)\right], \qquad m_j = \mathbf{1}[j \neq k], \tag{65}$$

where $\widehat{x}^t_j$ is the predicted conditional mean and $\widehat{\sigma}^2_j = \exp(\widehat{\ell}_j)$ is a learned gene-specific SCM variance. The measured target is separately supervised by the target-masked readout NLL in Eq. (10).

This does not make the model predict the random draw $\eta_{s,j}$ itself. Instead, the conditional mean explains the predictable component and the learned variance calibrates the residual uncertainty. This is important in Stage 07 because the generating endpoint is $\rho_{\text{pair}} = 0$, so the best achievable sample-wise prediction error is bounded away from zero even for the correct causal graph. Gaussian likelihoods are standard in causal discovery (Brouillard et al., 2020; Zheng et al., 2018; Ng et al., 2020; Xue et al., 2025).

The model also learns a global $\widehat{\rho}_{\text{pair}}$ rather than receiving the generator value:

$$\widehat{\varepsilon}^t = \widehat{\rho}_{\text{pair}}\,\widehat{\varepsilon}^d, \qquad \widehat{\rho}_{\text{pair}} = \text{sigmoid}(\theta_{\rho_{\text{pair}}}). \tag{66}$$

Subsequent benchmark stages retain the generator setting $\rho_{\text{pair}} = 0$, so treated exogenous noise remains independent of control-profile noise across samples and genes while preserving the gene-specific variance σ^2_j.

Stages 06-07 result. Heterogeneous noise retains exact five-seed support recovery, while independent treated residuals remain near exact, with F1 0.975 ± 0.040 at $G = 10$ and 0.999 ± 0.001 at $G = 100$. As intended, the independent-noise setting lowers raw treated-profile R^2 while largely preserving the structural signal (Supplementary Tables 6 and 7).

B.4 IMPERFECT LOSS-OF-FUNCTION-STAGES 08-09

All previous LoF benchmark stages treat the functional intervention as perfect: the effective target activity is clamped to zero, $x_k^{t,\text{eff}} = 0$. This is a useful identifiability control, but it is still too strong biologically. Several CRISPR studies report that perturbations are not always complete, knockout cells can retain residual protein expression or functional activity, escape loss of function, or show variable guide efficiency across cells Smits et al. (2019); Wang et al. (2024); Weinberger et al. (2024). The next generator therefore introduces a residual-activity variable.

Let $r_{s,k} \in [0, 1]$ denote the residual-activity variable for target gene k in sample s. The interpretation is

$$\begin{aligned} r_{s,k} = 0 \quad &\iff \quad \text{perfect functional LoF,} \\ r_{s,k} = 1 \quad &\iff \quad \text{no functional perturbation.} \end{aligned} \tag{67}$$

The generator first evaluates the unperturbed target mechanism using treated parent states and treated exogenous noise,

$$z_{s,k}^{t,\text{nat}} = f_k^*\left(z_{s,\text{Pa}(k)}^t\right) + \varepsilon_{s,k}^t, \qquad x_{s,k}^{t,\text{nat}} = \mu_k + z_{s,k}^{t,\text{nat}}. \tag{68}$$

For an intervention on target k, the target effective measured activity becomes

$$x_{s,k}^{t,\text{eff}} = r_{s,k} x_{s,k}^{t,\text{nat}} + (1 - r_{s,k}) x_{\text{LoF}}, \qquad x_{\text{LoF}} = 0. \tag{69}$$

Equivalently,

$$x_{s,k}^{t,\text{eff}} = r_{s,k} x_{s,k}^{t,\text{nat}}. \tag{70}$$

This construction is a restricted soft intervention in the sense of general interventional causal models. Rather than removing the dependence of the target on its parents, the intervention replaces its observational conditional distribution with a different parent-dependent conditional distribution Yang et al. (2018); Brouillard et al. (2020). In our case, this interventional distribution is induced by the biologically constrained attenuation above, which shifts the target's effective-activity distribution towards lower activity while preserving its dependence on the original parents. Thus, the proposed imperfect LoF is not an unrestricted mechanism change, but a specific member of the broader family of imperfect interventions: the direction of the distributional change is fixed by the biological semantics of loss of function.

The parent set and natural generator mechanism f_k^* are shared with the observational regime, but the interventional mechanism applies a post-mechanism attenuation before the target signal reaches its children. It is therefore not a claim that CRISPR leaves the target mechanism literally unchanged; rather, we assume that upstream regulation first determines the activity the target would naturally produce and that the LoF edit retains only a fraction of its functional output. Biologically, residual protein and partial function after frameshift editing have been observed through mechanisms such as translation reinitiation and edited-exon skipping Smits et al. (2019), while broader knockout-escape mechanisms are reviewed in Wang et al. (2024). We state this as a modeling assumption, not as a universal molecular description of every CRISPR LoF.

This restriction need not remove the interventional orientation signal. If attenuation induces a non-degenerate change in the target conditional distribution and in affected descendants, and if causal sufficiency, the known singleton-intervention family, and the remaining $\mathcal{I}^*$-faithfulness conditions used by DCDI hold, then the population interventional family identifies the true DAG Yang et al. (2018); Brouillard et al. (2020). This conditional result does not by itself establish that the finite generated sample, the non-convex optimizer, or real CRISPR data satisfies those assumptions.

The target therefore transmits an attenuated, sample- and gene-dependent signal to its children while the measured target continues to use the separate expression readout.

Supplementary Figure 15 represents the imperfect intervention as an attenuation gate on the effective target activity. This avoids the misleading interpretation that LoF deletes or weakens a biological edge: the parent set, natural target mechanism, outgoing edge identities, and child mechanisms are retained, while the effective target activity supplied to the children changes.

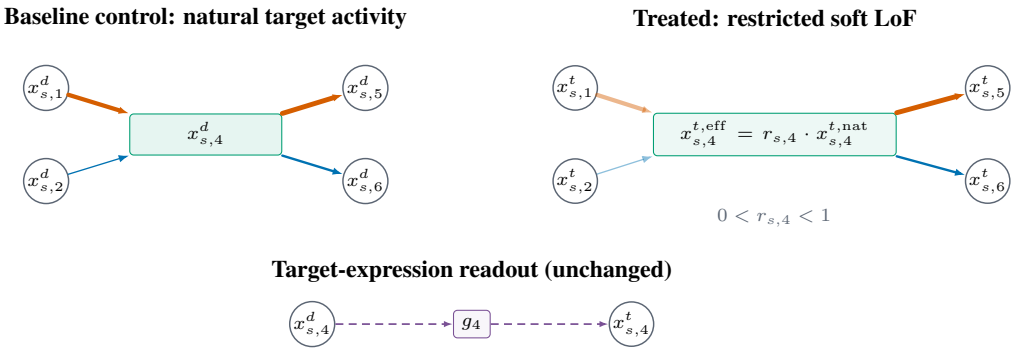

Figure 15: In the control expression profile, the natural target activity is propagated directly through its signed outgoing effects. In the treated expression profile, the same parent set first determines the natural target activity $x^{t,\mathrm{nat}}_{s,4}$, which is attenuated once by the sample-specific residual-activity variable $r_{s,4}$ before being supplied to the unchanged child mechanisms. The measured target $x^t_{s,4}$ is generated through the separate readout g_4 and is not propagated as functional activity.

In centered SEM coordinates, this same effective value is

$$z^{t,\mathrm{eff}}_{s,k} = x^{t,\mathrm{eff}}_{s,k} - \mu_k = r_{s,k} x^{t,\mathrm{nat}}_{s,k} - \mu_k = r_{s,k} z^{t,\mathrm{nat}}_{s,k} - (1 - r_{s,k})\mu_k. \tag{71}$$

Children of the target gene receive this effective target activity, implemented in the generator as the centered value $z^{t,\mathrm{eff}}_{s,k}$, whereas the measured target mRNA is still generated by the target-expression readout $g_k(x^d_k)$. Thus, imperfect LoF weakens the downstream interventional signal without forcing the observed target expression to follow the SCM mechanism. In particular, the Laplace target measurement is never propagated to the children, it is an observed readout distinct from the latent effective activity.

Because the residual-activity variable $r_{s,k}$ must lie between 0 and 1, we model it using a Beta distribution, whose support naturally matches this range,

$$r_{s,k} \sim \mathrm{Beta}(\alpha_k, \beta_k). \tag{72}$$

The controlled global setting uses

$$r_{s,k} \sim \mathrm{Beta}(8, 12), \qquad \mathbb{E}[r_{s,k}] = 0.40, \qquad \mathbb{E}[1 - r_{s,k}] = 0.60. \tag{73}$$

CRISPR LoF perturbations are generally imperfect (Smits et al., 2019). For this controlled setting, we choose a distribution that removes, on average, 60% of the target's effective activity while retaining the remaining 40%. This defines an imperfect-LoF stress test without claiming that the latent functional-loss fraction can be estimated directly from expression measurements in real cell lines.

To evaluate heterogeneous intervention strength across genes, we additionally consider a gene-specific version in which each target has its own expected functional loss,

$$p^{\mathrm{loss}}_k = 0.50 + 0.20 z_k, \qquad z_k \sim \mathrm{Beta}(4, 4), \tag{74}$$

and a fixed generating concentration $\phi^* = 30$,

$$r_{s,k} \sim \text{Beta}\left((1-p_k^{\text{loss}})\phi^*, p_k^{\text{loss}}\phi^*\right). \tag{75}$$

This allows different genes to exhibit different average levels of functional loss while keeping the overall mean close to the controlled 60% loss level.

Under those population conditions, imperfect LoF can preserve interventional orientation information, while still weakening the finite-sample signal. Expression measurements in real cell lines do not directly observe latent functional activity and are therefore never used to estimate the residual-activity distribution $r_{s,k}$. The frozen PC3 artifact calibrates observed measurement components introduced in Stages 03-07 and the control reuse introduced in Stage 12; it does not calibrate residual functional activity.

$$x_{s,k}^t = x_{s,k}^d + \delta_{s,k}, \qquad \delta_{s,k} \sim \text{Laplace}\big(\bar{\delta}_k^{\text{PC3}}, b_{\text{PC3}}\big). \tag{76}$$

This is a statistical calibration of observed readouts, not a claim to validate the latent functional-loss mechanism or to provide a complete biological simulator. The reproducible Stage 13 PC3-like comparison below aggregates all five declared generated seeds, checks the frozen-source hash and named-gene alignment, and reports descriptive agreement without using these quantities for model, threshold, or hyperparameter selection.

The model must now stop using a fixed hard clamp for the target effective value. In the global-residual version it learns a single retained-activity parameter,

$$\widehat{r} = \text{sigmoid}(\theta_r), \qquad \widehat{x}_{s,k}^{t,\text{eff}} = \widehat{r}\,\widehat{x}_{s,k}^{t,\text{nat}}. \tag{77}$$

Here $\widehat{x}_{s,k}^{t,\text{nat}}$ is obtained by evaluating the learned unperturbed mechanism of node k on the current treated-parent expression profile before applying the LoF attenuation. The paired control target is used only by the separate target-expression readout, not as the effective target activity propagated to descendants.

For the gene-specific setting, the final inference model represents target heterogeneity with a hierarchical Beta distribution. It learns a global logit a, raw gene-specific deviations η_k, and a shared concentration parameter κ. The deviations are centered across the G modeled genes before they enter the gene-specific logits,

$$\bar{\eta} = \frac{1}{G}\sum_{m=1}^{G}\eta_m, \qquad \widetilde{\eta}_k = \eta_k - \bar{\eta}, \qquad \rho_k = \text{sigmoid}(a + \widetilde{\eta}_k), \qquad \phi = \text{softplus}(\kappa). \tag{78}$$

$$r_k \sim \text{Beta}(\rho_k\phi, (1-\rho_k)\phi), \qquad \widehat{r}_k \equiv \mathbb{E}[r_k] = \rho_k, \qquad \widehat{\text{Var}}(r_k) = \frac{\rho_k(1-\rho_k)}{\phi+1}. \tag{79}$$

Centering gives $\sum_{k=1}^{G}\widetilde{\eta}_k = 0$, so the gene-specific terms describe deviations around the global logit a. Consequently, $\text{sigmoid}(a)$ is the mean residual activity for a gene with $\widetilde{\eta}_k = 0$; because the sigmoid is nonlinear, it is not generally the arithmetic mean of $(\rho_1, \ldots, \rho_G)$.

The shared concentration $\phi > 0$ controls uncertainty around each gene-specific mean: larger values concentrate the Beta distribution more tightly around ρ_k. Unlike the fixed concentration used by the synthetic generator, ϕ is learned from the training objective. The model does not sample r_k during inference. It uses $\rho_k = \mathbb{E}[r_k]$ to construct the effective target activity, while $\text{Var}(r_k) = \rho_k(1-\rho_k)/(\phi+1)$ is propagated analytically in the SCM negative log-likelihood.

Thus, ρ_k is the model-implied mean residual activity for the target. It is not a direct estimate of editing efficiency, protein abundance, or functional loss, and it cannot describe perturbations that alter the target mechanism in ways other than scalar attenuation.

Therefore, the effective target activity is,

$$\widehat{x}_{s,k}^{t,\mathrm{eff}} = \widehat{r}_k \widehat{x}_{s,k}^{t,\mathrm{nat}} = \rho_k \widehat{x}_{s,k}^{t,\mathrm{nat}}, \tag{80}$$

This effective value is the effective target activity supplied to downstream mechanisms. After intervention propagation, the measured target coordinate is produced by the learned gene-specific affine head

$$\widehat{x}_{s,k}^{t,\mathrm{meas}} = s_k x_{s,k}^{d} + b_k, \tag{81}$$

and the complete measured prediction is

$$\widehat{x}_s^t = (1 - u_s) \odot x_s^{t,\mathrm{eff}} + u_s \odot \left(s \odot x_s^d + b \right). \tag{82}$$

Here x_s^d is the observed control expression profile and is distinct from the learned expression baseline μ. The target-expression readout is trained through the predictive likelihood only; its output is not supplied to child mechanisms or to the SCM NLL.

The variance is used to calibrate the SCM negative log-likelihood. We reformulate the SCM NLL term as a Gaussian negative log-likelihood, following the same motivation as for the prediction loss after introducing irreducible treated noise. Once the residual activity is no longer deterministic, part of the SCM residual is an unavoidable consequence of uncertainty in the effective activity of the intervened target rather than graph misspecification. The predicted variance is therefore propagated through the SCM and incorporated into the interventional likelihood.

For the linear model used at this point in the progression, let $\widehat{x}_{s,k}^{t,\mathrm{nat}}$ be the target conditional mean, $\widehat{\sigma}_k^2$ its SCM noise variance, and $\widehat{v}_{r,k} = \widehat{\mathrm{Var}}(r_k)$. Because the paired exogenous noises are independent in this benchmark stage, the variance of the attenuated target source is

$$\widehat{\omega}_{s,k} = \widehat{v}_{r,k} \left(\widehat{x}_{s,k}^{t,\mathrm{nat}} \right)^2 + \widehat{m}_{r,k}^{(2)} \widehat{\sigma}_k^2, \qquad \widehat{m}_{r,k}^{(2)} = \widehat{v}_{r,k} + \rho_k^2. \tag{83}$$

If j is a candidate child, the resulting linear specialization is

$$\widehat{\sigma}_{s,j}^{2,\mathrm{LoF}} = \widehat{\omega}_{s,k} W_{kj}^2. \tag{84}$$

The first term in Eq. (83) is caused by uncertain residual activity acting on the target conditional mean. The second retains the target's own SCM noise after attenuation and therefore depends on the second moment of r_k, not only on its variance. Equation (84) is specific to the linear progression model; the final nonlinear model replaces W_{kj} by a local child-mechanism Jacobian, as detailed below.

For the interventional regime, the residual variance combines the learned gene-specific SCM variance with the propagated uncertainty from imperfect LoF,

$$(\widehat{\sigma}_{s,j}^{\mathrm{int}})^2 = \widehat{\sigma}_j^2 + (\widehat{\sigma}_{s,j}^{\mathrm{LoF}})^2. \tag{85}$$

The SCM term is therefore

$$\mathcal{L}_{\mathrm{SCM}} = \frac{1}{2\sum_s (G-1)} \sum_s \sum_{j \neq k} \left[\frac{(r_{s,j}^t)^2}{(\widehat{\sigma}_{s,j}^{\mathrm{int}})^2} + \log\left((\widehat{\sigma}_{s,j}^{\mathrm{int}})^2 \right) + \log(2\pi) \right]. \tag{86}$$

The intervened coordinate $j = k$ remains excluded from $\mathcal{L}_{\mathrm{SCM}}$. Its natural and effective target activities are latent, whereas the observed target follows the separate measured-expression readout. Including the observed target value as if it were the effective target activity would impose the wrong structural residual. The target-expression readout is instead trained by the target-masked predictive term in Eq. (10).

Stages 08-09 result. Imperfect residual activity lowers the achievable treated-profile signal while retaining high structural recovery. At $G = 100$, Stage 08 reaches AUPRC/F1 $0.989 \pm 0.003/0.987 \pm 0.004$, and Stage 09 reaches $0.994 \pm 0.003/0.985 \pm 0.007$; the learned uncertainty model therefore preserves most of the structural signal once gene-specific residual activity is introduced (Supplementary Tables 6 and 7).

B.5 Biological Complexity and Scaling-Stages 10-13

B.5.1 Stage 10: Biological Topology

This benchmark stage tests whether the model can recover the DAG across different topologies, rather than only under the simpler Erdős-Rényi setting. We consider hub-modular and scale-free DAGs as biology-informed stress tests. These topologies capture properties associated with gene regulatory networks, including sparsity, modularity, hierarchy, and heterogeneous degree distributions Farkas et al. (2006); Rodríguez-Caso et al. (2009); Takemoto & Oosawa (2007); Greenbury et al. (2010); Brown et al. (2025). They are not assumed to represent the true topology of biological regulation; the goal is to test whether recovery is robust across different graph topologies.

The generator continues to sample a random topological order and restricts edges to follow that order, thereby guaranteeing acyclicity. In the hub-modular variant, within-module and hub-outgoing edges are sampled with higher probability, whereas hub-incoming edges are downweighted. For $G = 100$, the generator produces 292 edges across five modules, with 5% of nodes designated as upstream hubs. These coarse properties were calibrated against PC3 PPI proxy.

In the scale-free generator, each new node selects $m = 3$ parents among preceding nodes with probability proportional to their current out-degree plus one. This produces upstream hubs through preferential attachment while preserving the DAG.

Supplementary Figure 16 contrasts the Erdős-Rényi, hub-modular, and scale-free topologies used in this progression.

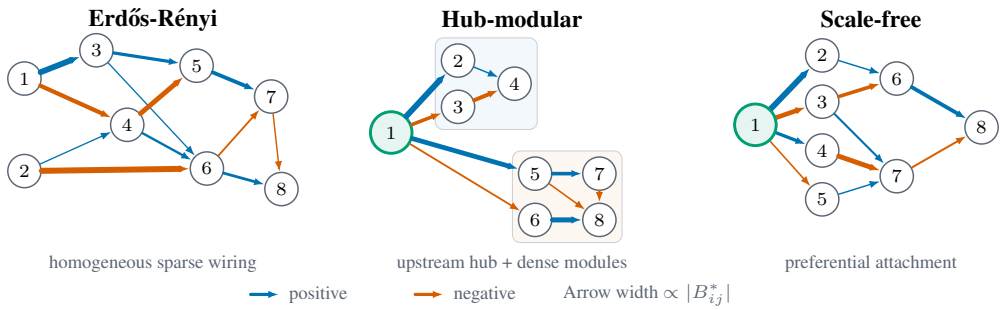

Figure 16: Stage 10 graph families: Erdős-Rényi, hub-modular, and scale-free DAGs. Arrow width is proportional to edge magnitude.

Changing topology does not modify the model architecture or objective. All subsequent stages use the scale-free construction: heterogeneous degree distributions and hubs provide a literature-supported approximation to one aspect of regulatory-network organization, rather than a universal biological law (Farkas et al., 2006; Rodríguez-Caso et al., 2009; Greenbury et al., 2010). The alternatives provide a topology robustness check (Supplementary Table 7).

Stage 10 result. All three topologies yield acyclic supports in every seed and retain high precision-recall balance. At $G = 100$, F1 is 0.985 ± 0.005 for scale-free, 0.982 ± 0.017 for hub-modular, and 0.985 ± 0.007 for Erdős-Rényi graphs (Supplementary Table 7).

B.5.2 Stage 11: Nonlinear Mechanisms

Linear structural equations assign a coefficient to every parent-child pair, but cannot represent saturation, thresholds, or interactions between parents. We instead assign a separate nonlinear mechanism to every child, following the child-specific, graph-masked MLP construction used by GIM and DCDI Schneider et al. (2025); Brouillard et al. (2020). The generator follows GIM's synthetic construction by instantiating one network per variable with one hidden layer of five units and Gaussian weights. The synthetic mechanism of node j is

$$u_j = \frac{D_j\left(A^*_{\cdot j} \odot z\right)}{\tau_{\text{in}}}, \tag{87}$$

$$f^*_j(z) = s_j\left[\text{MLP}_j(u_j) - \text{MLP}_j(0)\right], \tag{88}$$

$$z_j = f^*_j(z) + \varepsilon_j, \qquad x_j = \mu_j + z_j. \tag{89}$$

The true adjacency column $A^*_{\cdot j}$ masks every non-parent input. τ_{in} controls the input scale, D_j contains optional parent-specific calibration factors, and s_j controls the output amplitude. Subtracting $\text{MLP}_j(0)$ prevents the neural mechanism from replacing the separately calibrated baseline μ_j. Samples are generated by evaluating these equations in a topological order, and therefore constitute an exact ancestral evaluation of the true DAG.

Crucially, the parameters of MLP_j are not shared with any other child. Consequently, every gene has its own response surface: two genes with the same parent set can still respond differently to the same parent values, and the same parent can have a different sign, strength, saturation, and interaction pattern in each child. This child-specific conditional mechanism gives every gene a distinct mechanistic identity rather than treating all genes as copies of one shared nonlinear transformation.

Supplementary Figure 17 summarizes the child-specific nonlinear data-generating process and its exact ancestral evaluation over the true DAG.

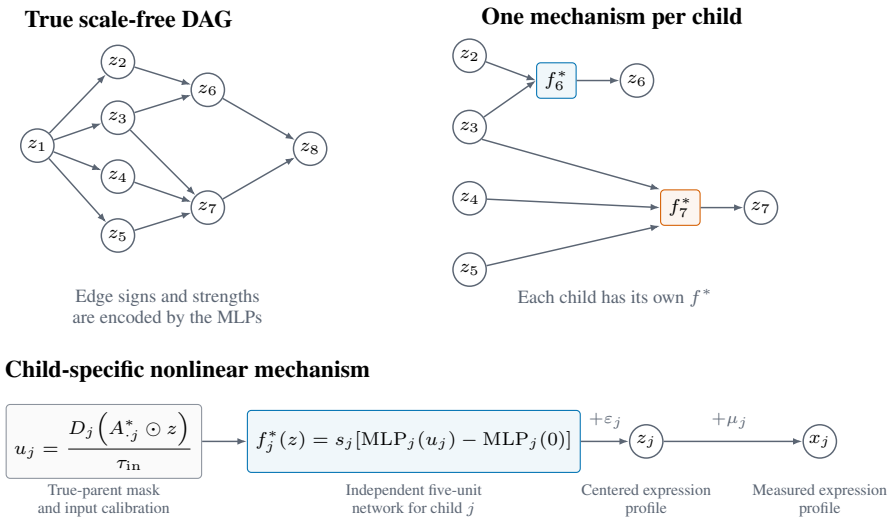

Figure 17: Stage 11 nonlinear generator. Each child receives only its true parents and has an independently parameterized mechanism. Samples are evaluated in topological order; arrows encode parenthood rather than a global effect sign or magnitude.

Random neural mechanisms can contain functionally silent nominal edges. We therefore screen each child mechanism on fixed sensitivity and intervention probe sets. For $i \in \text{Pa}(j)$, let z^π_i be coordinate i after a fixed permutation across observational probes and write $\widetilde{z}_i = r_i(\mu_i + z_i) - \mu_i$. We compute

$$D^{\text{mean}}_{ij} = \mathbb{E}_{\text{sens}}\left[\left|\frac{\partial f^*_j(z)}{\partial z_i}\right|\right], \qquad D^{\text{p10}}_{ij} = Q_{0.1,\text{sens}}\left(\left|\frac{\partial f^*_j(z)}{\partial z_i}\right|\right), \tag{90}$$

$$C_{ij} = \mathbb{E}_{\text{obs}}\left[\left(f^*_j(z) - f^*_j(z_{-i}, z^\pi_i)\right)^2\right], \tag{91}$$

$$E^{\text{rms}}_{ij} = \left\{\mathbb{E}_{\text{obs},r}\left[\left(f^*_j(z_{-i}, \widetilde{z}_i) - f^*_j(z)\right)^2\right]\right\}^{1/2}, \quad E^{\text{p10}}_{ij} = Q_{0.1,\text{obs},r}\left(\left|f^*_j(z_{-i}, \widetilde{z}_i) - f^*_j(z)\right|\right), \tag{92}$$

$$I_j = \left\{\sum_{i \in \text{Pa}(j)} \mathbb{E}_{\text{sens}}\left[\left(\frac{\partial f^*_j(z)}{\partial z_i}\right)^2\right]\right\}^{1/2}. \tag{93}$$

Here $Q_{0.1}$ is the empirical tenth quantile over the corresponding probe set. Acceptance requires

$$D_{ij}^{\mathrm{mean}} \geq \underline{D}_{\mathrm{mean}}, \qquad D_{ij}^{\mathrm{p10}} \geq \underline{D}_{\mathrm{p10}}, \qquad \underline{C} \leq C_{ij} \leq \overline{C}, \tag{94}$$

$$\underline{E}_{\mathrm{rms}} \leq E_{ij}^{\mathrm{rms}} \leq \overline{E}_{\mathrm{rms}}, \qquad E_{ij}^{\mathrm{p10}} \geq \underline{E}_{\mathrm{p10}}, \qquad I_j \leq \overline{I}, \tag{95}$$

for every true parent. The numerical bounds are resolved recipe parameters and are not optimized during model training.

The sampler scales parent-input columns, up to the recipe limit, to meet the lower derivative and LoF-effect bounds. One child-output rescaling then enforces the upper conditional-contribution, interventional-effect, and incoming-gain bounds. All summaries are recomputed, and a mechanism failing any bound is resampled. These checks operationalize detectable interventional change under $\mathcal{I}^*$-faithfulness; they are not an independent identifiability theorem (Brouillard et al., 2020).

The inference model retains the child-specific nonlinear family but not the generator's exact calibration. It centers expression by a learned expression baseline μ, uses bias-free internal layers, and writes every training mechanism with the soft gate matrix $A^{(g)}$:

$$f_{\theta,j}(z; A^{(g)}) = v_j^\top \tanh\left(\sum_{i \neq j} A_{ij}^{(g)} U_{j,\cdot i} z_i\right). \tag{96}$$

We write $f_\theta(z; A^{(g)})$ for the vector collecting all G centered child-specific mechanisms. Their measured-scale means are $\mu + f_\theta(z; A^{(g)})$. Conceptually, $U_j \in \mathbb{R}^{H \times G}$ and $v_j \in \mathbb{R}^H$ belong only to child j. The implementation stores them in tensors of shape $G \times H \times G$ and $G \times H$ for vectorized evaluation, without sharing parameters across children. Following GraN-DAG Lachapelle et al. (2020), the same weights define

$$A_{ij} = \sum_h |v_{jh}|\,|U_{jhi}|.$$

There is no independently parameterized structural gate. When enabled by the resolved recipe, the same neural path strength produces the soft gate matrix

$$A_{ij}^{(g)} = \sigma\left(\frac{A_{ij} - q_{\mathrm{gate}}}{\tau}\right), \qquad A_{ii}^{(g)} = 0,$$

which attenuates $U_{j,\cdot i}$ during continuous training. A separately declared q_{eval} produces the binarized adjacency $A_{ij}^{(b)} = \mathbf{1}[A_{ij} \geq q_{\mathrm{eval}}]$, which replaces $A^{(g)}$ during checkpoint selection and official evaluation. Neither matrix is learned independently. The gate is not sampled and its temperature is not annealed.

Initialization is specified by an expected parent count $d_{\mathrm{par}}^{(0)} > 0$ and expected initial neural path strength $s_A^{(0)} > 0$. Before a common rescaling, the Xavier-style bounds are

$$a_U = \sqrt{\frac{6}{d_{\mathrm{par}}^{(0)} + H}}, \qquad a_v = \sqrt{\frac{6}{H+1}}.$$

For independent centered uniforms, $\mathbb{E}[\sum_h |U_{jhi} v_{jh}|] = H a_U a_v / 4$. Both bounds are multiplied by

$$c_A = \sqrt{\frac{s_A^{(0)}}{H a_U a_v / 4}},$$

so that the declared initial path-strength scale does not shrink merely with the measured dimension. These are internal initialization controls, not a privileged sampled support.

Following the nonlinear Gaussian conditional model used by GIM, each structural conditional is

$$p_\theta(x_j \mid z; A^{(g)}) = \mathcal{N}\Big(x_j; \mu_j + f_{\theta,j}(z; A^{(g)}), \sigma_j^2\Big). \tag{97}$$

Here $\sigma_j^2 = \exp(\ell_j^{\mathrm{SCM}})$ is a learned gene-specific SCM variance. A separate learned vector $\varsigma_j^2 = \exp(\ell_j^{\mathrm{pred}})$ calibrates the predictive post-LoF likelihood; neither variance changes the deterministic mechanism mean. Both are initialized from control-profile variability in the training partition only.

Accordingly, the SCM fit is a Gaussian negative log-likelihood rather than a linear residual loss. It follows a teacher-forced causal path that is distinct from the intervention-propagation path. Writing $\bar{u} = 1 - u$ for the non-target indicator and $z^t_{\setminus k} = \bar{u} \odot (x^t - \mu)$, the teacher-forced treated profile is

$$\check{x}^t = \bar{u} \odot x^t + u \odot \rho_k \left[\mu_k + f_{\theta,k}\Big(z^t_{\setminus k}; A^{(g)} \Big) \right]. \tag{98}$$

Thus the observed non-target coordinates are retained, whereas the observed target coordinate is removed before evaluating its natural mechanism and is then replaced by its mean effective target activity. Let $\check{z}^t = \check{x}^t - \mu$. The interventional term excludes target k and evaluates the remaining structural conditionals on $\check{z}^t$:

$$\mathcal{L}_{\mathrm{SCM}} = -\frac{1}{G-1} \sum_{j \neq k} \log p\Big(\check{x}^t_j \mid \check{z}^t; A^{(g)} \Big). \tag{99}$$

For the final nonlinear model, uncertainty in residual target activity and target SCM noise is propagated locally through the child-mechanism Jacobians, as detailed in Appendix A.1. The resulting $\nu_{n,j}$ augments only the SCM variance; it does not alter the deterministic prediction or predictive variance.

The target remains excluded because its measured treated expression does not directly reveal its latent effective target activity. The model uses the mean residual activity ρ_k to construct that effective target activity; the measured target readout in Eq. (81) does not enter this causal path.

The nonlinear intervention-propagation step cannot use the closed-form linear solve based on $(I - W)^{-1}$ because the treated expression profile appears inside the MLPs and must satisfy an implicit equation. During continuous training the path-strength matrix may still have cyclic support; such intermediate states are optimized numerically but are not interpreted as causal DAGs. The predictive path starts from the observed control expression profile x^d. With $z^{(\ell)} = x^{(\ell)} - \mu$, $n^{(\ell)} = \mu + f_\theta(z^{(\ell)}; A^{(g)})$ and $\pi^{(\ell)} = u \odot \rho_k n^{(\ell)}$, the damped fixed-point iteration is initialized and updated over ℓ as

$$\begin{aligned} x^{(0)} &= \bar{u} \odot x^d + u \odot \rho_k \left[\mu_k + f_{\theta,k}(z^d; A^{(g)}) \right], \\ \widetilde{x}^{(\ell)} &= \bar{u} \odot n^{(\ell)} + \pi^{(\ell)}, \\ x^{(\ell+1)} &= \bar{u} \odot \left[(1-\delta) x^{(\ell)} + \delta \widetilde{x}^{(\ell)} \right] + \pi^{(\ell)}, \qquad 0 < \delta \leq 1. \end{aligned} \tag{100}$$

Only the non-target coordinates are damped. The effective target assignment $\pi^{(\ell)}$ is recomputed and imposed without damping at every iteration. The final effective treated activity profile is $x^{t,\mathrm{eff}} = x^{(Q)}$, where Q is the fixed iteration budget; only after this causal computation is complete is its target coordinate replaced by the target-expression readout according to Eq. (82).

Fixed-point evaluation is the standard numerical construction for implicit neural maps Bai et al. (2019). In our setting, it provides a differentiable intervention-propagation procedure throughout graph optimization; once the binary adjacency defines a selected DAG, its fixed point is the ancestral SCM solution. A cyclic intermediate mask does not in general guarantee existence or uniqueness of a solution Bongers et al. (2021), so we use damping, a finite iteration budget, and monitor the final update magnitude rather than interpreting intermediate cyclic masks as the recovered causal model.

Stage 11 result. Child-specific nonlinear mechanisms reduce finite-sample recovery relative to the linear stages but retain five DAGs out of five, with $G = 100$ AUPRC 0.853 ± 0.023 and F1 0.803 ± 0.017 (Supplementary Table 7).

B.5.3 Stage 12: Reused Controls

The real paired construction reuses a finite set of control expression profiles within each cell line and is not a true before/after measurement of the same cell. We reproduce this sampling design, the generator first creates a finite pool of independent observational SCM profiles and then assigns each control to multiple treated samples.

Supplementary Figure 18 illustrates how a finite control pool is reused across independently generated treated records.

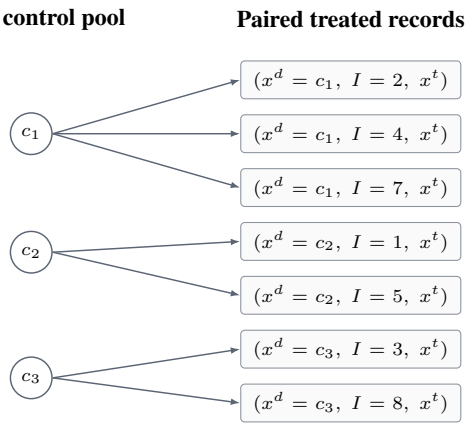

Figure 18: A finite set of independently generated observational profiles is assigned to multiple treated records. The assignment is independent of intervention target, and each treated well receives an independent exogenous realization.

The PC3 pair count, control-pool size, and reuse multiplicities are read from the frozen PC3 calibration artifact rather than from a cross-cell-line average. For a synthetic dataset containing N pairs, the generator sets

$$C = \mathrm{round}\left(N \frac{C_{\mathrm{PC3}}}{N_{\mathrm{PC3}}}\right). \tag{101}$$

The empirical PC3 reuse-count distribution is transferred by quantile matching and adjusted minimally so that the multiplicities sum exactly to N. The resulting control identifiers are shuffled independently of the intervention schedule, preventing control identity from becoming associated with a particular target.

Stage 12 result. Reusing controls retains five DAGs out of five and produces $G = 100$ AUPRC 0.850 ± 0.020 and F1 0.812 ± 0.014, close to the nonlinear Stage 11 setting while matching the paired-control design (Supplementary Table 7).

B.5.4 Stage 13: PC3-like Calibration Fidelity

Supplementary Table 9 and Supplementary Figures 19-20 compare all five Stage 13 datasets with the PC3 calibration source. The comparison measures fidelity to the calibrated statistics rather than biological realism beyond those features.

Table 9: Descriptive fidelity of the Stage 13 generator relative to its frozen PC3 calibration source. Distribution summaries use fixed declared samples; discrepancies are not selection criteria.

Metric	PC3	Stage 13 (5 seeds)	Discrepancy	Calibration role
Baseline expression	q10/50/90=2.485/2.924/3.505	q10/50/90=2.479/2.927/3.51	W1 (log2 expression): 0.003325 ± 0.000159	direct: PC3 control median and noise
Treated expression	q10/50/90=2.484/2.925/3.506	q10/50/90=2.479/2.927/3.51	W1 (log2 expression): 0.003163 ± 0.000253	emergent
Target delta	q10/50/90=-0.154/-0.01413/0.09372	q10/50/90=-0.1709/-0.009019/0.1392	W1 (log2 expression): 0.01964 ± 0.00171	direct: PC3 gene-specific readout
Paired correlation	q10/50/90=0.9437/0.969/0.9816	q10/50/90=0.948/0.963/0.9703	—difference—: 0.008052 ± 0.000585	emergent
Non-target effect RMS	0.1116	0.1166 ± 0.00226	—difference—: 0.004994 ± 0.00226	emergent
Target frequency	q10/50/90=4/6/6	q10/50/90=4/6/6	JS (nats): 0 ± 0	direct: PC3 target counts
Control reuse	ratio=0.06397; q10/50/90=12/15/19	ratio=0.06397 $\pm$ 0; q10/50/90=12/15/19	W1 (pairs/control): 0 ± 0	direct: PC3 reuse multiplicities

Stage 13 result. At PC3-like scale and the fixed $q_{\mathrm{eval}} = 0.09$ used in the real-data experiments, RISCAL returns five DAGs out of five with AUPRC 0.633 ± 0.016, precision 0.640 ± 0.022, recall 0.659 ± 0.015, and F1 0.649 ± 0.018 (Supplementary Table 14).

Control reuse and shared pair noise are distinct. If control c is assigned to several treated records, its complete measured expression profile, including its realized exogenous residual ε_c^d, is repeated in those records. Each treated expression profile nevertheless comes from a different well and receives an independent exogenous realization:

$$\varepsilon_{s,j}^t = \eta_{s,j}, \qquad \eta_{s,j} \sim \mathcal{N}(0, \sigma_j^2), \qquad \eta_{s,j} \perp\!\!\!\perp \varepsilon_{c,j}^d, \qquad \rho_{\mathrm{pair}} = 0. \tag{102}$$

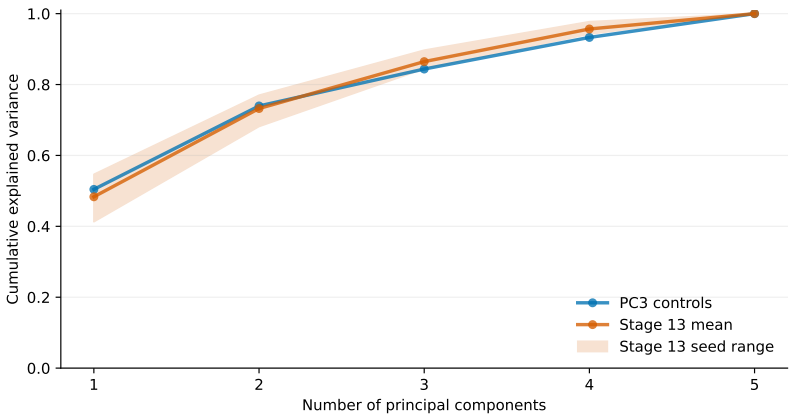

Figure 19: Cumulative PCA explained variance for PC3 and Stage 13 control profiles. The Stage 13 curve is the seed mean, with the full seed range shaded.

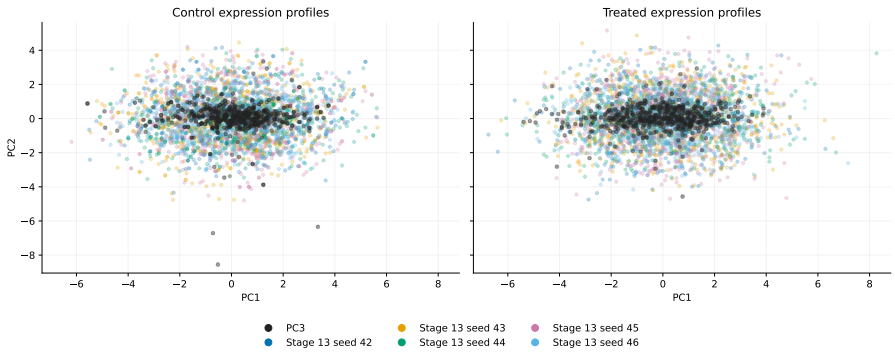

Figure 20: Independent two-component PCAs of PC3 and Stage 13 control and treated expression profiles. Common rendered limits compare coverage and scale; dataset-specific PCA bases preclude comparison of component directions.

This matches the construction of the real PC3 pairs: a control and its paired treated expression profile are not repeated measurements of the same biological unit, but expression profiles from different wells that are paired computationally. Reusing a control therefore repeats one observed baseline; it does not provide evidence that its sample-specific residual is inherited by any treated well. Gene-specific stable levels remain represented by μ_j, while variability among independently measured controls and treated wells remains represented by σ_j^2.

The conditional treated-noise mean under this already-defined independent-noise regime becomes

$$\mathbb{E}\left[\varepsilon_{s,j}^{t} \mid \varepsilon_{c,j}^{d}\right] = \rho_{\text{pair}}\varepsilon_{c,j}^{d} = 0, \tag{103}$$

while the Gaussian likelihood retains the irreducible treated variance σ_j^2. Thus, retaining $\rho_{\text{pair}} = 0$ removes only an unsupported predictable transport of the control residual; it does not remove gene-specific biological or measurement variability.

When assignment is target-independent, control reuse does not change the population causal graph or intervention semantics. It does, however, make pairs sharing a control statistically dependent and reduce the number of distinct observational profiles. This change requires no modification of the SCM architecture or training objective.

C LINCS Data Provenance and Preprocessing

C.1 Dataset Provenance and Cohort Construction

We reprocess the ten PDGrapher cell-line datasets (Gonzalez et al., 2026) from LINCS L1000 Level 3 expression profiles (Subramanian et al., 2017; Keenan et al., 2018). Each sample is an aggregate well-level expression profile from a perturbed or control cancer-cell population.

The ten cell lines are A375, A549, AGS, BICR6, ES2, HT29, MCF7, PC3, U251MG and YAPC. We retain this selection to make the experimental setting comparable to PDGrapher, while modifying the preprocessing and the perturbation definition.

Although PDGrapher describes the genetic datasets as single-gene CRISPR-Cas9 knockouts, the LINCS instances selected by the preprocessing pipeline correspond to the perturbation type `trt_xpr`. We therefore refer to these experiments as single-gene CRISPR loss-of-function (LoF) perturbations rather than strict knockouts. This distinction is biologically relevant: LoF perturbations are designed to reduce or disrupt gene function, but they do not necessarily imply complete deletion or complete loss of activity of the target gene. Control samples are selected from `ctl_vector` instances from the same cell line and are used as the baseline condition.

For cell line c, the supervised dataset is

$$\mathcal{D}_c = \left\{ (x_i^{d,c}, u_i^c, x_i^{t,c}) \right\}_{i=1}^{N_c}, \qquad u_i^c = e_{k_i^c} \in \{0,1\}^{G_c}, \tag{104}$$

where N_c is the number of paired examples, G_c the retained-gene count, i the sample index, $x_i^{d,c}, x_i^{t,c} \in \mathbb{R}^{G_c}$, k_i^c the perturbed-gene index, and $e_{k_i^c}$ the corresponding one-hot vector. The compact dataset artifact stores the integer index k_i^c; the data loader constructs u_i^c when collating each minibatch. We retain the superscript d for compatibility with the dataset and implementation; throughout the manuscript it denotes the baseline control context rather than a reference-state contrast.

The measurement unit is one aggregate well-level expression profile. The supervised unit is a constructed triplet containing measurements from a treated well and a distinct vector-control well. It is therefore not a before/after observation of the same cell or cell population. Because a control expression profile can be assigned to several treated expression profiles, triplets that share that control are dependent; the control identity defines the grouping unit used by the train, validation, and test split procedure.

PDGrapher operates on proxy causal graphs derived from protein-protein interaction networks (PPI) or gene regulatory networks (GRN). In contrast, our models do not use a PPI or GRN as an input graph. The only inherited use of the PDGrapher preprocessing is the initial gene-universe filtering based on PPIs, which restricts the expression space to the same 10,716 candidate genes used in that benchmark.

We start from LINCS Level 3 beta files for CRISPR LoF perturbations and vector controls. For each cell line, we select perturbation samples with `pert_type = trt_xpr` and control samples with `pert_type = ctl_vector`, remove failed experimental instances, intersect sample identifiers with the available binary GCTX-format expression matrices, and discard perturbations whose target gene is not present in the retained gene table.

Expression values are transformed as

$$x \leftarrow \log_2(x+1). \tag{105}$$

No additional min-max normalization, z-scoring, or feature-wise scaling is applied. This is a deliberate difference from the distributed PDGrapher preprocessing, which applies a logarithmic transformation followed by gene-specific min-max normalization. We remove the min-max step to avoid leakage through feature-wise extrema computed using held-out samples.

After intersecting with the PDGrapher-compatible candidate gene universe, samples whose target gene is not retained are removed. Supplementary Table 10 reports for every cell line: *Input LoF* is the treated expression profile count before the PPI-target filter; *PPI-supported* retains profiles whose target gene remains in that universe; *Plate excluded* counts profiles removed because their plate lacks eligible controls for same-plate pairing; and *Output pairs* is the final paired dataset after gene reduction.

Table 10: Cell-line-specific data-construction counts emitted from the extraction and preparation manifests. LoF denotes loss of function and PPI denotes protein-protein interaction.

Cell line	Input LoF	PPI-supported	Non-PPI excluded	Plate excluded	Output pairs	Final genes
A375	23523	21794	1729	26	21768	3706
A549	26094	24255	1839	3595	20660	3709
AGS	23071	21284	1787	0	21284	3712
BICR6	22981	21183	1798	0	21183	3711
ES2	25540	23708	1832	3751	19957	3595
HT29	22220	20525	1695	0	20525	3706
MCF7	20328	18774	1554	0	18774	3090
PC3	23040	21229	1811	0	21229	3710
U251MG	28337	26323	2014	3611	22712	3712
YAPC	21809	20135	1674	0	20135	3711

Supplementary Table 11 gives the complementary descriptive summary of the final corrected datasets. Cancer context is reported as a descriptive annotation. *Controls/pair* is the number of unique reused controls divided by output pairs; median and 90th-percentile (P90) reuse count the output pairs assigned to a single physical control. The latter quantities motivate keeping all triplets that share a control in the same data split.

Table 11: Descriptive properties of the final corrected cell-line datasets. Controls/pair is unique reused controls divided by output pairs; reuse quantiles count output pairs per control and P90 denotes the 90th percentile.

Cell line	Cancer context	Unique controls	Controls/pair	Median reuse	P90 reuse
A375	Skin cancer	1374	0.06312	15	20
A549	Lung cancer	1313	0.06355	15	19
AGS	Stomach cancer	1352	0.06352	15	19
BICR6	Head and neck cancer	1362	0.0643	15	19
ES2	Ovarian cancer	1277	0.06399	15	19
HT29	Colon cancer	1313	0.06397	15	19
MCF7	Breast cancer	1210	0.06445	15	19
PC3	Prostate cancer	1358	0.06397	15	19
U251MG	Brain cancer	1449	0.0638	15	19
YAPC	Pancreatic cancer	1275	0.06332	15	19

C.2 Batch Correction and Contrast Preservation

The extracted LINCS inputs do not themselves define treated-control pairs. For each LoF profile x_i^t, the target gene is obtained from the LINCS `cmap_name` field and encoded as the one-hot intervention vector u_i. The deterministic same-plate pairing contract, including cyclic control reuse and protocol-signature eligibility, is specified in Sections 3 and 6. This correction stage operates on those fixed pairs and never draws a baseline from the global control pool.

Same-plate pairing makes the treated-control difference within each pair robust to a plate-wide additive shift. However, the model learns jointly from pairs collected on many detection plates. Residual shifts between plates could then be mistaken for differences in cellular expression profiles rather than technical variation. We therefore align profiles across plates before fitting the SCM. The left panel of Supplementary Figure 21 visualizes this cross-plate separation for PC3 vector controls before correction.

For sample s, gene j, and detection-plate map $p(s)$, the accompanying correction assumes the additive measurement model

$$x_{sj}^{\log} = \mu_j + \delta_{sj} + b_{p(s),j} + \varepsilon_{sj}, \tag{106}$$

where $b_{p(s),j}$ is the technical offset for gene j on plate $p(s)$ and $\delta_{sj} = 0$ for vector controls.

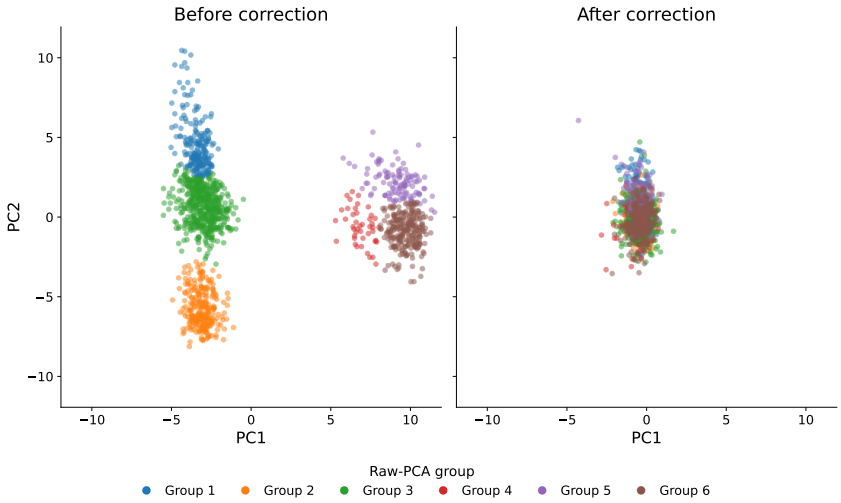

Figure 21: PC3 vector controls before and after plate-specific control-only correction. One principal component analysis (PCA) is fitted to the union of the raw and corrected controls, and both panels use the same axes and limits. Colors denote groups obtained by fitting an unsupervised Gaussian mixture to the raw the first two PCA coordinates (PC1 and PC2) coordinates, without using detection-plate or biological labels. The Bayesian information criterion (BIC) selects the number of components from two through eight; the selected raw-PCA assignments and colors are then held fixed in the corrected panel. PC3 is shown as a representative cell line; Supplementary Table 12 summarizes the same diagnostic for all ten cell lines.

This correction follows the established principle of estimating unwanted technical variation from negative-control features or reference samples, rather than from the biological conditions of interest. Remove-unwanted-variation (RUV) methods formalize a related version of this principle by using negative-control genes to isolate unwanted variation (Gagnon-Bartsch & Speed, 2012), while the L1000 processing pipeline itself uses invariant transcripts and plate-wise normalization to reduce technical variation (Subramanian et al., 2017). These methods motivate control-based estimation and plate-specific calibration.

Our estimator is most closely related to the quality-control calibration of Espín-Pérez et al. (2018), who estimate a gene-specific batch displacement from reference samples and apply it to every sample measured in that batch. Likewise, reference-material ratio correction has been shown to improve cross-batch comparability in large-scale transcriptomic and multi-omics studies, including settings where technical and biological factors are partially confounded (Yu et al., 2023). On the logarithmic scale, such ratio-based calibration corresponds to the additive plate term $b_{p(s),j}$ in Eq. (106). We adopt this control-reference formulation, replacing means with medians and using a plate-balanced global reference.

Crucially, plate effects are estimated separately for each cell line and exclusively from its negative controls, in the complete measured gene space before the PPI and intervention-supported reductions. Let C_p be the control set of plate p and $\mathcal{P}_{\mathrm{trt}}$ the set of plates containing retained treated expression profiles. We compute

$$\widetilde{c}_{pj} = \mathrm{median}_{c \in C_p}\, x_{cj}^{\log}, \qquad \widetilde{c}_{\mathrm{global},j} = \mathrm{median}_{p \in \mathcal{P}_{\mathrm{trt}}}\, \widetilde{c}_{pj}. \tag{107}$$

The second median gives equal weight to every treated plate. To regularize plates with few controls, we evaluate

$$\widehat{c}_{pj} = w_p \widetilde{c}_{pj} + (1 - w_p)\widetilde{c}_{\mathrm{global},j}, \qquad w_p = \frac{|C_p|}{|C_p| + \tau_{\mathrm{prep}}}, \tag{108}$$

For each cell line, the preprocessing shrinkage parameter τ_{prep} is selected from the declared grid by strict control leave-one-out (LOO): a held-out control is removed from its plate median and from the plate-balanced reference, then candidate values are ordered by median robust reference error, with the smaller τ_{prep} breaking ties. At most 64 controls per plate are scored in this selection diagnostic.

Neither treated expression profiles, intervention targets, learned causal DAGs, test metrics, nor model results enter this choice. Supplementary Table 12 reports the selected values for all ten cell lines. The resulting gene-specific offset is

$$\widehat{b}_{pj} = \widehat{c}_{pj} - \widetilde{c}_{\mathrm{global},j}. \tag{109}$$

Both members of a pair are corrected using their shared plate offset,

$$\widetilde{x}_i^t = x_i^t - \widehat{b}_{p(i)}, \qquad \widetilde{x}_i^d = x_i^d - \widehat{b}_{p(i)}. \tag{110}$$

The resulting paired intervention contrast is

$$\widetilde{x}_i^t - \widetilde{x}_i^d = x_i^t - x_i^d, \tag{111}$$

so applying the same plate offset to both members preserves each paired intervention contrast exactly while expressing both profiles relative to the same plate-balanced control reference. Repeated uses of one control remain exact repetitions after correction.

This control-only transformation aligns the plate-associated component of the control expression scale. The right panel of Supplementary Figure 21 shows the same PC3 controls after correction in the common PCA coordinate system. Supplementary Table 12 reports, for every cell line, the median gene-specific fraction of control variance explained by detection plate, η^2, before and after correction. Across the ten cell lines, this quantity falls from 0.646-0.712 before correction to 0.030-0.041 after correction, an approximately twenty-fold reduction.

Although the PDGrapher-compatible filtering step leaves 10,716 candidate genes, learning a causal graph over all of them would include many genes that are never directly perturbed within a given cell line. For causal graph learning, we therefore use an intervention-based dimensionality reduction step: for each cell line, we retain only genes that appear as perturbation targets at least once in that dataset. This preserves all supervised pairs but restricts $x_i^{d,c}$, u_i^c, and $x_i^{t,c}$ to the subset of genes that can be directly intervened upon. Besides reducing dimensionality, this choice better matches interventional causal-discovery assumptions, where direct interventions on variables provide information for orienting and constraining causal structure (Brouillard et al., 2020).

Table 12: Control-only correction diagnostics emitted from the declared manifests. Here τ_{prep} is the preprocessing shrinkage parameter; η^2 raw/corrected is the median gene-specific fraction of control variance explained by detection plate before/after correction; LOO denotes leave-one-out, RMS denotes root mean square, and p95 denotes the 95th percentile.

Cell line	τ_{prep}	Supported plates	Excluded plates	Excluded treated	Median plate η^2: raw/corrected	LOO RMS p95
A375	0	95	1	26	0.6842/0.03342	0.01453
A549	0	88	30	3595	0.6715/0.03624	0.0145
AGS	0	90	0	0	0.6926/0.03411	0.01466
BICR6	1	90	0	0	0.6721/0.04074	0.01201
ES2	0	87	35	3751	0.6887/0.03334	0.01416
HT29	0	89	0	0	0.6791/0.03562	0.01435
MCF7	0	75	0	0	0.6456/0.03025	0.009871
PC3	0	89	0	0	0.6872/0.03375	0.01416
U251MG	0	91	30	3611	0.7116/0.03127	0.011
YAPC	0	89	0	0	0.6843/0.03508	0.01562

D Causal Structure Evaluation and Biological Plausibility Analysis

D.1 Causal Baseline Implementation and Synthetic Recovery

The data-rich $G = 100$ and PC3-scale benchmarks compare RISCAL with each baseline under its stated support rule.

SDCD. We run SDCD version 0.1.4 (Nazaret et al., 2024) on the training split and retain the binary support and threshold emitted by its native procedure. The ground-truth DAG is used only for post-training structural metrics.

InterSort and InterSort-SORTRANKING. InterSort (Chevalley et al., 2025b) receives observational and single-gene interventional samples. We run the complete procedure, including local search, at $G = 100$. At PC3 scale ($G = 3710$), we report its scalable SORTRANKING mode separately because complete local search is not configured at this dimension.

GENIE3. GENIE3 (Huynh-Thu et al., 2010) fits one random forest per target using all other genes as predictors and ranks edges by impurity-based importance. We use scikit-learn with 1,000 trees and $\sqrt{G}$ candidate regulators per split. It does not use intervention identity or enforce acyclicity. Its dense output is thresholded to 3.5 edges per gene, matching the Random DAG budget. Because its top scores contain near-symmetric pairs, its oracle row is reported without an acyclicity constraint.

Random DAG baseline. The Random DAG chance reference samples an acyclic graph from a random topological order with an expected 3.5 edges per gene. Its budget is fixed independently of the ground-truth DAG.

D.2 Causal Benchmark Results

Data-rich $G = 100$ benchmark. Supplementary Table 13 reports the complete five-seed comparison. The native/protocol block is the primary comparison. Its oracle F1-maximizing threshold block is a strictly post-hoc synthetic diagnostic: it changes only the final displayed evaluation threshold separately for each seed and does not alter fitting, checkpoint selection, or optimization gates.

Table 13: Data-rich $G = 100$ causal-discovery comparison across five dataset seeds. The native/protocol block uses only each method's predeclared final support rule. The oracle F1-maximizing threshold block is a post-hoc synthetic diagnostic: for each seed it chooses only the final evaluation threshold that maximizes F1 against the known synthetic graph; it changes neither training nor checkpoint selection, and it does not alter optimization gates. AUPRC is threshold-invariant. Best values are bold within each block. GENIE3's oracle row alone is not acyclicity-constrained, since it has no acyclicity mechanism to respect (Appendix D.1).

Model	Support rule	DAGs / 5	Pred. / ground-truth edges	AUPRC	Precision	Recall	F1
Native/protocol support							
RISCAL	selected checkpoint, $q = 0.025$	5/5	392.0 ± 23.8 / 357.8 ± 3.0	$\mathbf{0.850 \pm 0.020}$	$\mathbf{0.778 \pm 0.030}$	0.851 ± 0.020	$\mathbf{0.812 \pm 0.014}$
SDCD	native support, $t = 0.1$	5/5	591.4 ± 45.6 / 357.8 ± 3.0	0.148 ± 0.012	0.199 ± 0.011	0.329 ± 0.032	0.247 ± 0.016
InterSort	native support, $\epsilon = 0.5$	5/5	709.2 ± 44.2 / 357.8 ± 3.0	0.779 ± 0.021	0.475 ± 0.024	$\mathbf{0.940 \pm 0.012}$	0.631 ± 0.020
GENIE3	predeclared 350-edge budget	0/5	350.0 ± 0.0 / 357.8 ± 3.0	0.190 ± 0.007	0.271 ± 0.010	0.266 ± 0.012	0.268 ± 0.011
Random DAG baseline	predeclared native support	5/5	350.0 ± 0.0 / 357.8 ± 3.0	0.039 ± 0.003	0.033 ± 0.003	0.032 ± 0.003	0.033 ± 0.003
Oracle F1-maximizing threshold diagnostic (post hoc; synthetic ground truth only)							
RISCAL	per-seed oracle F1-maximizing threshold (post hoc)	5/5	342.4 ± 18.1 / 357.8 ± 3.0	$\mathbf{0.850 \pm 0.020}$	$\mathbf{0.864 \pm 0.025}$	$\mathbf{0.826 \pm 0.022}$	$\mathbf{0.844 \pm 0.010}$
SDCD	per-seed oracle F1-maximizing threshold (post hoc)	5/5	564.6 ± 84.6 / 357.8 ± 3.0	0.148 ± 0.012	0.206 ± 0.015	0.323 ± 0.042	0.251 ± 0.017
InterSort	per-seed oracle F1-maximizing threshold (post hoc)	5/5	374.8 ± 48.3 / 357.8 ± 3.0	0.779 ± 0.021	0.685 ± 0.042	0.714 ± 0.053	0.697 ± 0.017
GENIE3	per-seed oracle F1-maximizing threshold (post hoc, no acyclicity constraint)	0/5	442.0 ± 68.2 / 357.8 ± 3.0	0.190 ± 0.007	0.252 ± 0.022	0.308 ± 0.031	0.276 ± 0.013

PC3-scale benchmark. Supplementary Table 14 reports the complete five-seed $G = 3710$ comparison. Its native/protocol block is primary; the oracle F1-maximizing threshold diagnostic uses the ground-truth DAG only after model and checkpoint selection to vary the final displayed threshold separately for each seed. It does not modify soft gate matrices or define the reported model. Because full InterSort local search is not run at this dimension, its rows are labelled InterSort-SORTRANKING.

D.3 Real-Graph Stability and Biological Enrichment

Learned causal DAGs are directed hypotheses under the SCM assumptions. We assess reproducibility, curated directed overlap, functional coherence, degree-controlled topology, and residual-activity

Table 14: PC3-scale causal-discovery comparison ($G = 3710$) across five dataset seeds. The native/protocol block uses each method's predeclared final support rule. At this dimension InterSort uses SORTRANKING without local search. The oracle F1-maximizing threshold block is a post-hoc synthetic diagnostic that changes only the final evaluation threshold; it affects neither training nor checkpoint selection. AUPRC is threshold-invariant. Best values are bold within each block.

Model	Support rule	DAGs / 5	Pred. / ground-truth edges	AUPRC	Precision	Recall	F1
Native/protocol support							
RISCAL	selected checkpoint, $q_{\text{eval}} = 0.09$	5/5	14144.0 ± 290.6 / 13732.4 ± 23.2	**0.633 ± 0.016**	**0.640 ± 0.022**	0.659 ± 0.015	**0.649 ± 0.018**
SDCD	native support, $t = 0.1$	5/5	516.2 ± 228.1 / 13732.4 ± 23.2	0.001 ± 0.000	0.006 ± 0.004	0.000 ± 0.000	0.000 ± 0.000
InterSort-SORTRANKING	native support, $\epsilon = 0.5$	5/5	4926234.0 ± 15259.6 / 13732.4 ± 23.2	0.556 ± 0.008	0.003 ± 0.000	**0.946 ± 0.002**	0.005 ± 0.000
Random DAG baseline	predeclared native support	5/5	12985.0 ± 0.0 / 13732.4 ± 23.2	0.001 ± 0.000	0.001 ± 0.000	0.001 ± 0.000	0.001 ± 0.000
Oracle F1-maximizing threshold diagnostic (post hoc; synthetic ground truth only)							
RISCAL	per-seed oracle F1-maximizing threshold (post hoc)	5/5	10711.2 ± 462.8 / 13732.4 ± 23.2	**0.633 ± 0.016**	**0.767 ± 0.011**	**0.598 ± 0.022**	**0.672 ± 0.013**
SDCD	per-seed oracle F1-maximizing threshold (post hoc)	5/5	183.8 ± 186.3 / 13732.4 ± 23.2	0.001 ± 0.000	0.073 ± 0.086	0.000 ± 0.000	0.000 ± 0.000
InterSort-SORTRANKING	per-seed oracle F1-maximizing threshold (post hoc)	5/5	12331.6 ± 725.8 / 13732.4 ± 23.2	0.556 ± 0.008	0.611 ± 0.021	0.548 ± 0.015	0.578 ± 0.007

stability. These graph-level analyses support biological plausibility rather than edge-wise causal correctness; biological resources do not define a complete cell-line-specific ground-truth DAG.

D.3.1 CONSENSUS STABILITY

For consensus level $k/5$, an edge becomes a candidate when it occurs in the selected edge sets of at least k fits. Candidates are ordered by frequency and continuous neural path strength and accepted only when they preserve acyclicity. Supplementary Table 15 summarizes core node and edge sizes at the 3/5, 4/5, and 5/5 levels; Supplementary Table 16 retains the complete candidate and acyclic-projection accounting. At the primary 3/5 level, pairwise directed-edge Jaccard ranges from a mean of 0.378 in PC3 to 0.467 in A549; five-fit consensus cores nevertheless contain 2,343-4,913 edges. Top-1% out-hub stability is more variable, so subsequent interpretation emphasizes consensus modules and enrichment rather than individual fitted hubs (Supplementary Figure 22).

Table 15: Size of reproducible graph cores across five fixed-test fits. A node is counted when it is incident to at least one directed edge in the deterministically acyclic consensus DAG. The final column is the strict 5/5 edge count relative to the primary 3/5 core.

Cell line	Nodes 3/5	Edges 3/5	Nodes 4/5	Edges 4/5	Nodes 5/5	Edges 5/5	Edges 5/5 / 3/5
A375	3042	9896	2856	7225	2562	4869	49.2%
A549	2978	9807	2821	7209	2537	4913	50.1%
AGS	2971	9014	2758	6369	2407	4194	46.5%
BICR6	3002	9530	2785	6724	2462	4339	45.5%
ES2	2882	8948	2678	6589	2379	4348	48.6%
HT29	2872	7994	2648	5819	2343	3887	48.6%
MCF7	2179	5095	1974	3649	1687	2343	46.0%
PC3	2909	8184	2554	5221	2173	3089	37.7%
U251MG	2809	7185	2577	5127	2219	3311	46.1%
YAPC	3007	10271	2754	7314	2436	4821	46.9%

Table 16: Acyclic consensus graph size across five fixed-test fits. Candidate edges meet the frequency rule before deterministic acyclic projection.

Cell	Consensus	Candidates	DAG edges	Cycle rejected	Present 5/5
A375	2/5	13974	13967	7	4869
A375	3/5	9896	9896	0	4869
A375	4/5	7225	7225	0	4869
A375	5/5	4869	4869	0	4869
A549	2/5	13865	13860	5	4913
A549	3/5	9807	9807	0	4913
A549	4/5	7209	7209	0	4913
A549	5/5	4913	4913	0	4913
AGS	2/5	12922	12915	7	4194
AGS	3/5	9014	9014	0	4194
AGS	4/5	6369	6369	0	4194
AGS	5/5	4194	4194	0	4194

Continued on next page

Table 16 continued

Cell	Consensus	Candidates	DAG edges	Cycle rejected	Present 5/5
BICR6	2/5	13841	13829	12	4339
BICR6	3/5	9530	9530	0	4339
BICR6	4/5	6724	6724	0	4339
BICR6	5/5	4339	4339	0	4339
ES2	2/5	12603	12596	7	4348
ES2	3/5	8948	8948	0	4348
ES2	4/5	6589	6589	0	4348
ES2	5/5	4348	4348	0	4348
HT29	2/5	11255	11247	8	3887
HT29	3/5	7994	7994	0	3887
HT29	4/5	5819	5819	0	3887
HT29	5/5	3887	3887	0	3887
MCF7	2/5	7466	7461	5	2343
MCF7	3/5	5095	5095	0	2343
MCF7	4/5	3649	3649	0	2343
MCF7	5/5	2343	2343	0	2343
PC3	2/5	12549	12545	4	3089
PC3	3/5	8184	8184	0	3089
PC3	4/5	5221	5221	0	3089
PC3	5/5	3089	3089	0	3089
U251MG	2/5	10427	10418	9	3311
U251MG	3/5	7185	7185	0	3311
U251MG	4/5	5127	5127	0	3311
U251MG	5/5	3311	3311	0	3311
YAPC	2/5	14715	14707	8	4821
YAPC	3/5	10271	10271	0	4821
YAPC	4/5	7314	7314	0	4821
YAPC	5/5	4821	4821	0	4821

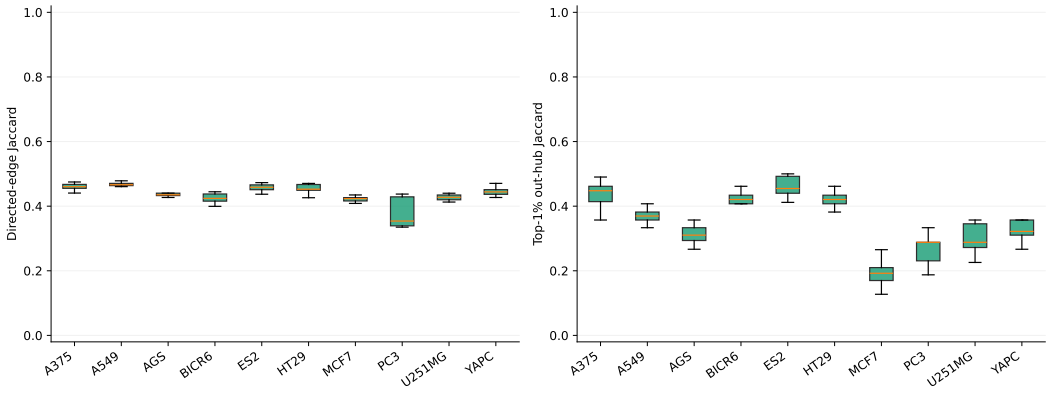

Figure 22: Stability across five fixed-test fits per cell line. Directed-edge Jaccard measures exact support overlap; top-1% out-hub Jaccard measures overlap among the highest-out-degree nodes. These are reproducibility diagnostics rather than evidence of biological correctness.

D.3.2 Directed biological enrichment

Supplementary Table 17 provides the primary 3/5 overlap counts underlying Table 2. Exact in/out-degree-preserving acyclic swaps generate each null ensemble. Significant enrichment in at least one resource for every cell line shows that curated interactions concentrate in the stable supports beyond expectations from sparsity and hub structure. Small absolute overlaps are expected from the sparse, incomplete, and context-aggregated nature of the resources and are not converted into edge-level accuracy.

Table 17: Directed-resource overlap at the primary consensus 3/5. Null DAGs preserve the exact in/out-degree sequence; q-values use Benjamini-Hochberg correction over the declared resource, cell-line and consensus tests.

Cell	Resource	Graph edges	Observed	Null mean	Fold	q
A375	CollecTRI	9896	9	2.94	3.07	0.002
A375	SIGNOR	9896	8	3.20	2.50	0.008
A549	CollecTRI	9807	6	3.95	1.52	0.201
A549	SIGNOR	9807	10	2.68	3.73	0.002
AGS	CollecTRI	9014	9	1.98	4.54	0.002
AGS	SIGNOR	9014	10	2.15	4.65	0.002
BICR6	CollecTRI	9530	13	3.42	3.80	0.002
BICR6	SIGNOR	9530	9	3.49	2.58	0.014
ES2	CollecTRI	8948	4	2.30	1.74	0.196
ES2	SIGNOR	8948	13	3.11	4.18	0.002
HT29	CollecTRI	7994	17	3.11	5.46	0.002
HT29	SIGNOR	7994	8	2.82	2.83	0.017
MCF7	CollecTRI	5095	6	3.05	1.97	0.081
MCF7	SIGNOR	5095	9	2.61	3.44	0.006
PC3	CollecTRI	8184	11	2.52	4.37	0.002
PC3	SIGNOR	8184	15	2.75	5.45	0.002
U251MG	CollecTRI	7185	10	2.23	4.49	0.002
U251MG	SIGNOR	7185	12	2.36	5.09	0.002
YAPC	CollecTRI	10271	11	3.92	2.81	0.006
YAPC	SIGNOR	10271	7	4.97	1.41	0.240

D.3.3 Functional modules within cell lines

Louvain communities in the undirected consensus projection are tested independently against Reactome (Milacic et al., 2024) and Gene Ontology Biological Process (The Gene Ontology Consortium, 2023). Supplementary Table 18 summarizes significant annotations; Supplementary Table 19 reports the illustrative module selected by the frozen rule for each line. The selected modules are displayed in Supplementary Figures 23, 24, 25, 26, 27, 28, 29, 30, and 31. The selected annotations span antigen presentation, cell cycle, antiviral and immune responses, transcriptional regulation, and related programs. Directed database overlap is kept separate from functional enrichment: a coherent pathway module need not have every internal direction catalogued in CollecTRI or SIGNOR. Across all module figures, solid green or red edges denote externally catalogued activation or inhibition, respectively, whereas dashed blue edges are learned-only relative to CollecTRI and SIGNOR.

Table 18: Numbers of significant Reactome and Gene Ontology Biological Process enrichment results after Benjamini-Hochberg correction. Counts are descriptive because functional terms are dependent and overlapping.

Cell	Significant terms	Modules	Hub target sets
A375	445	13	10
A549	583	14	11
AGS	358	12	7
BICR6	356	10	8
ES2	238	9	9
HT29	245	11	12
MCF7	176	16	2
PC3	351	17	5
U251MG	262	16	4
YAPC	126	8	4

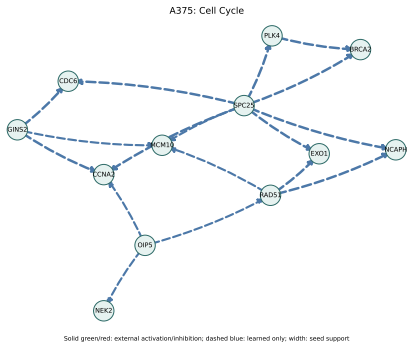

Figure 23: Illustrative reproducible module for A375, selected by the frozen reporting rule.

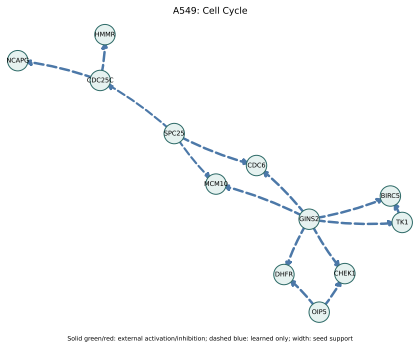

Figure 24: Illustrative reproducible module for A549, selected by the frozen reporting rule.

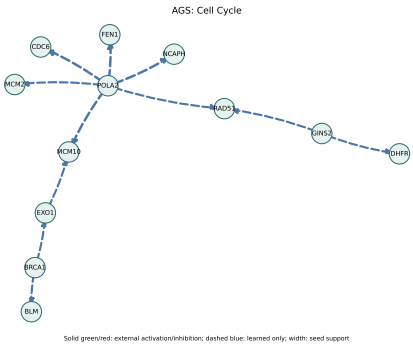

Figure 25: Illustrative reproducible module for AGS, selected by the frozen reporting rule.

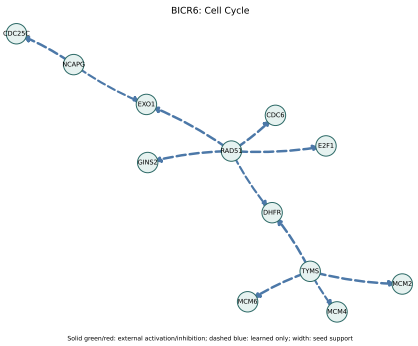

Figure 26: Illustrative reproducible module for BICR6, selected by the frozen reporting rule.

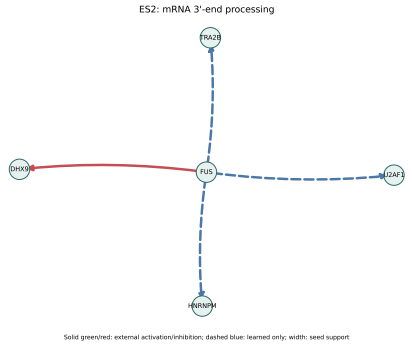

Figure 27: Illustrative reproducible module for ES2, selected by the frozen reporting rule.

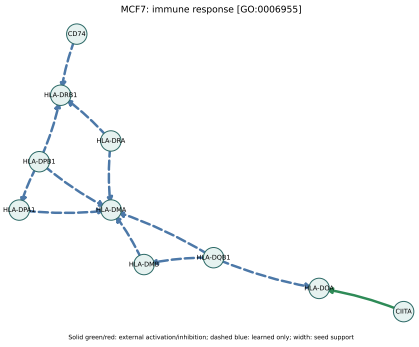

Figure 28: Illustrative reproducible module for MCF7, selected by the frozen reporting rule.

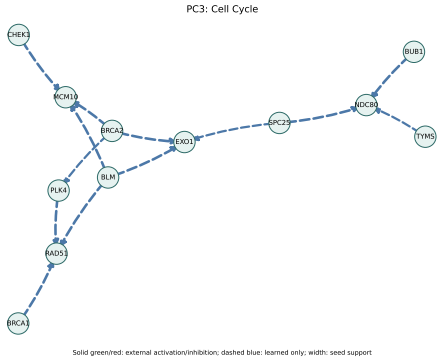

Figure 29: Illustrative reproducible module for PC3, selected by the frozen reporting rule.

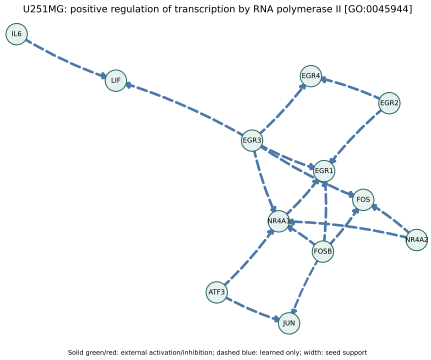

Figure 30: Illustrative reproducible module for U251MG, selected by the frozen reporting rule.

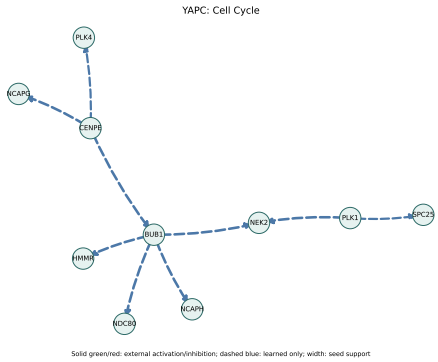

Figure 31: Illustrative reproducible module for YAPC, selected by the frozen reporting rule.

Table 19: Illustrative reproducible modules selected by the frozen consensus-4/5 reporting rule. Directed matches count displayed edges present in CollecTRI or SIGNOR; pathway coherence and directed-edge support are distinct evidence.

Cell	Annotation	Resource	q	Genes	Edges	Directed matches
A375	Cell Cycle	Reactome	< 0.001	12	17	0
A549	Cell Cycle	Reactome	< 0.001	12	14	0
AGS	Cell Cycle	Reactome	< 0.001	12	11	0
BICR6	Cell Cycle	Reactome	< 0.001	12	11	0
ES2	mRNA 3'-end processing	Reactome	0.029	5	4	1
HT29	defense response to virus [GO:0051607]	GO Biological Process	< 0.001	12	19	2
MCF7	immune response [GO:0006955]	GO Biological Process	< 0.001	10	12	1
PC3	Cell Cycle	Reactome	< 0.001	12	13	0
U251MG	positive regulation of transcription by RNA polymerase II [GO:0045944]	GO Biological Process	< 0.001	12	17	0
YAPC	Cell Cycle	Reactome	< 0.001	10	9	0

D.3.4 Programs conserved across cell lines

We next ask whether independently learned modules share genes or functional enrichment profiles across cell lines. The analysis compares all 27,400 module pairs, corrects gene and functional overlap tests over the full declared family, and retains reciprocal best matches satisfying the frozen size, shared-term, cosine, and corrected-overlap criteria. Every cell-line pair contains one to seven conserved programs. As Supplementary Figure 32 shows, functional similarity is consistently higher

than exact gene-set overlap: across the 45 line pairs, the median maximum reciprocal functional cosine is 0.881, compared with gene-set Jaccard 0.169. This supports conserved high-level programs with context-specific gene membership. Each heatmap entry is the best reciprocal match under its own criterion, so the two panels need not represent the same module pair; even a functional cosine of one indicates matching enrichment profiles, not identical gene membership or regulatory wiring.

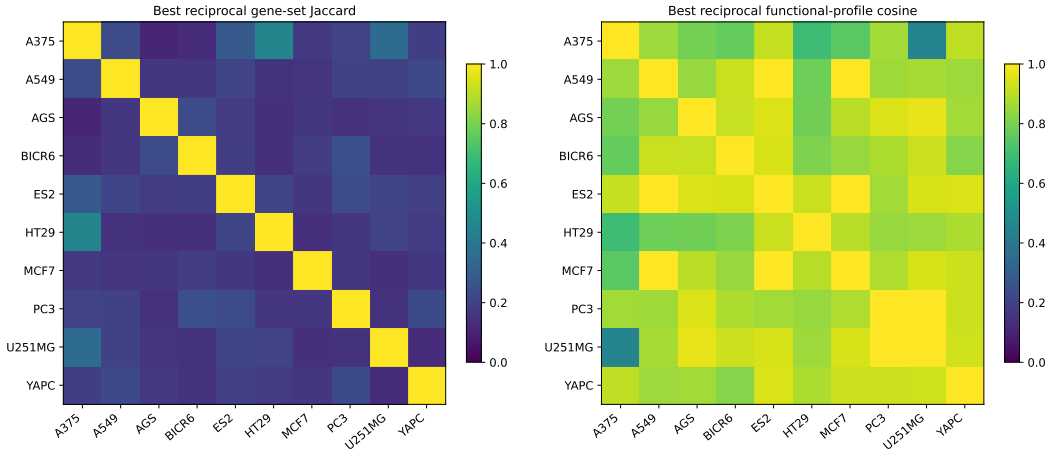

Figure 32: Cross-cell-line similarity of reciprocally matched consensus modules. The left matrix reports the largest reciprocal gene-set Jaccard for each pair of lines; the right reports the corresponding maximum cosine similarity between significant functional-enrichment profiles. Diagonal entries are one by definition.

D.3.5 Degree-controlled topology

The primary acyclic consensus graphs are more modular and clustered than acyclic nulls with identical in/out-degree sequences. Modularity is 1.055-1.118 times the null mean and transitivity is 1.732-2.980 times the null mean, with one-sided empirical $p = 0.0099$ throughout; mean path length is 1.016-1.085 times the null mean. Supplementary Table 20 therefore supports modular and clustered organization beyond density and degree alone, consistent with qualitative properties reported for gene-regulatory networks (Farkas et al., 2006; Rodríguez-Caso et al., 2009). This graph-level evidence does not establish recovery of a unique or true biological network topology.

Table 20: Topology of 3/5 acyclic consensus graphs relative to 100 degree-preserving acyclic nulls. Ratios divide the observed statistic by the null mean.

Cell	Modularity ratio	Transitivity ratio	Path-length ratio	Small-world coefficient
A375	1.097	2.167	1.016	2.133
A549	1.094	2.391	1.016	2.355
AGS	1.108	2.024	1.070	1.892
BICR6	1.109	2.007	1.060	1.893
ES2	1.118	1.983	1.054	1.882
HT29	1.088	2.475	1.065	2.325
MCF7	1.055	2.980	1.073	2.778
PC3	1.112	2.756	1.085	2.540
U251MG	1.063	2.779	1.029	2.701
YAPC	1.105	1.732	1.061	1.632

D.3.6 Learned residual activity

Supplementary Table 21 asks whether the learned residual-activity distribution is numerically stable across fits. Global means lie between 0.516 and 0.521, shared concentration remains near 29.93, and

across-gene dispersion is limited. These parameters provide a stable latent route for intervention uncertainty; they are not direct estimates of guide efficiency, protein abundance, or retained molecular function.

Table 21: Learned residual-activity distributions. These latent effective-activity parameters support uncertainty propagation and are not direct measurements of editing efficiency.

Cell line	Fits	Global mean	Concentration	Across-gene SD	Mean Beta variance
A375	5	0.5181 ± 0.0007	29.935 ± 0.003	0.0082 ± 0.0003	0.00807 ± 0.00000
A549	5	0.5176 ± 0.0009	29.940 ± 0.004	0.0078 ± 0.0004	0.00807 ± 0.00000
AGS	5	0.5195 ± 0.0017	29.930 ± 0.008	0.0086 ± 0.0008	0.00807 ± 0.00000
BICR6	5	0.5191 ± 0.0011	29.931 ± 0.005	0.0088 ± 0.0006	0.00807 ± 0.00000
ES2	5	0.5168 ± 0.0003	29.942 ± 0.001	0.0077 ± 0.0001	0.00807 ± 0.00000
HT29	5	0.5170 ± 0.0005	29.944 ± 0.002	0.0073 ± 0.0002	0.00807 ± 0.00000
MCF7	5	0.5158 ± 0.0004	29.944 ± 0.002	0.0076 ± 0.0002	0.00807 ± 0.00000
PC3	5	0.5179 ± 0.0007	29.939 ± 0.001	0.0080 ± 0.0002	0.00807 ± 0.00000
U251MG	5	0.5209 ± 0.0013	29.926 ± 0.006	0.0090 ± 0.0006	0.00807 ± 0.00000
YAPC	5	0.5180 ± 0.0004	29.936 ± 0.002	0.0081 ± 0.0002	0.00807 ± 0.00000

E Post-Intervention Prediction Benchmarking

All predictive baselines follow the evaluation metrics and paired inference in Appendix F.

E.1 Common protocol and prediction baselines implementation

All prediction-only baselines are evaluated within each cell line. Except for GEARS, their input for pair n is the paired control expression profile x_n^d and the one-hot target indicator u_n, and their output is a full treated-profile prediction $\widehat{x}_n^t \in \mathbb{R}^G$. Linear regression, the MLP, the axial transformer, and PDGrapher use exactly the same grouped-control outer splits as RISCAL. Their train, validation, and test partitions are therefore identical at the level of original pair and control identity. The neural baselines select the checkpoint minimizing validation macro intervention mean squared error (MSE), then evaluate that frozen checkpoint once on the corresponding outer test split. GEARS is the stated exception: its published condition-level validation split is nested within its admissible outer training pool.

The paired-control reference is not trained: for each held-out pair it sets

$$\widehat{x}_n^t = x_n^d. \tag{112}$$

It answers the simple question of whether a method improves upon carrying the matched vector-control expression profile forward unchanged. Its role in the definition of perturbation-response R^2 is given in Appendix F; it is a metric reference, not an additional learned expression baseline in the model-ranking figures.

Each baseline is implemented as:

Linear regression. The linear baseline is a multivariate ordinary least-squares regressor. Its feature vector concatenates the complete paired control expression and the one-hot intervention identity,

$$z_n = [x_n^d; u_n] \in \mathbb{R}^{2G}, \qquad \widehat{x}_n^t = W z_n, \tag{113}$$

where $W \in \mathbb{R}^{G \times 2G}$ is fit on the training partition with no intercept and no explicit regularizer. Thus it can learn a target-specific average response and a linear dependence on the particular control expression profile, but has no nonlinear interactions, message passing, or supplied graph. It is implemented with `sklearn.linear_model.LinearRegression` from scikit-learn with `fit_intercept=False`, and fit once per outer split; because ordinary least squares has no training epochs or validation checkpoint, its fitted training solution is evaluated directly on that split's validation and test pairs.

Multilayer perceptron. The multilayer perceptron (MLP) is a fully connected response predictor with separate encoders for control expression and intervention identity. The control vector is mapped to 256 hidden units and the one-hot intervention vector to 64 hidden units. The two encodings are concatenated and passed through two 512-unit ReLU layers, each followed by dropout with probability 0.3, before a linear output layer predicts all G treated expression coordinates. It is trained with AdamW on pointwise treated-profile MSE for at most 300 epochs (batch size 512, learning rate 10^{-4}). A validation macro-intervention-MSE criterion, a ReduceLROnPlateau scheduler, and early stopping with patience 10 choose the checkpoint.

Axial transformer. The axial transformer is a high-capacity, non-structural predictive reference trained with the same paired inputs, partitions, and checkpoint criterion as RISCAL.

The axial transformer uses the same input pair (x_n^d, u_n) but replaces the MLP's fusion network with attention over learned tokens. It first projects the control and intervention vectors to 1,024 and 64 scalar features, respectively, maps each scalar to a 256-dimensional token, concatenates the 1,088 tokens, and reshapes them into a 64×17 grid. Learned row and column embeddings are added before two axial-attention blocks; each block has eight attention heads and applies attention along both grid axes, followed by a feed-forward update. Mean pooling over the resulting tokens and a final linear layer produce $\widehat{x}_n^t$. The model is trained with AdamW on treated-profile MSE for at most 300 epochs (batch size 128, learning rate 3×10^{-5}); validation macro-intervention MSE selects the checkpoint, with learning-rate scheduling and early stopping of patience 15. This grid is an architectural reshaping of learned features, not a gene graph or an externally supplied biological network.

PDGrapher. PDGrapher (Gonzalez et al., 2026) is run through its response prediction module on the cell-line-specific PPI edge list retained by the preprocessing pipeline. For every pair, the response module receives one node feature pair $(x^d_{n,j}, u_{n,j})$ per gene, together with the PPI edge index, and returns a full predicted treated expression profile. Input binning thresholds are the 501 empirical percentiles of the flattened training-control expression only; they are recomputed inside every outer split.

The implementation uses one graph-neural-network layer of width 16 and two downstream neural layers. It is optimized with Adam on treated-profile MSE for at most 50 epochs (batch size 64, learning rate 10^{-2}), with the stated step scheduler and early stopping. The checkpoint with minimum validation macro-intervention MSE is evaluated on the untouched outer test pairs.

GEARS. GEARS (Roohani et al., 2024) is run in the pinned `perturbagenes-gears:0.1.2` container with provisioned GO resources. For every outer cell-line split, its split-local training `AnnData` contains one pooled control expression profile for each underlying control and the admissible treated training profiles. At inference, GEARS predicts a response conditional on the target gene and does not receive that profile's paired control as an input.

Targets absent from the GEARS GO graph are excluded separately in the outer train, validation, and test portions. Within the GO-supported training pool only, target conditions with fewer than two profiles are further excluded because GEARS requires replicated conditions to form its native differential-expression annotations. GEARS then performs its native condition-level `simulation` split only inside that filtered outer training pool, with 80% of admissible training genes in its native train portion, and uses the resulting native validation criterion to select `best_model`. The declared GEARS recipe uses a 64-dimensional hidden state, one GO-graph layer, one gene-graph layer, a 16-unit decoder, learning rate 10^{-3}, batch size 32, and at most 20 epochs. The outer validation metric is a post-selection diagnostic; it does not select the checkpoint. Test metrics are computed only on the GO-evaluable held-out profiles.

Supplementary Table 22 reports the declared and used outer training counts, the train exclusions split into GO and replication causes, and the evaluable/declared validation and test counts. The last column gives the numbers of GO-excluded targets in train/validation/test order; their identities remain in each fit's emitted `result.yaml`. These model-specific exclusions do not define a reduced common test subset for the other baselines.

Table 22: GEARS coverage. GO denotes Gene Ontology; "rep." denotes exclusions caused by insufficient replicated target conditions; and tr/val/te denote the outer training, validation, and test partitions. Counts separated by a slash follow the order stated in each header.

Cell line	Fit	Train declared	Train used	Train exclusions (GO/rep.)	Validation evaluable/declared	Test evaluable/declared	GO-excluded targets (tr/val/te)
A375	0	15218	14990	17/211	2190/2191	4356/4359	4/1/2
A375	1	15228	15006	18/204	2181/2181	4356/4359	4/0/2
A375	2	15217	15008	15/194	2189/2192	4356/4359	4/3/2
A375	3	15230	15009	16/205	2177/2179	4356/4359	4/1/2
A375	4	15219	14981	18/220	2190/2190	4356/4359	4/0/2
A549	0	14438	14259	12/167	2076/2078	4136/4144	4/2/4
A549	1	14440	14257	12/171	2074/2076	4136/4144	4/2/4
A549	2	14446	14274	12/160	2068/2070	4136/4144	4/1/4
A549	3	14431	14247	14/170	2085/2085	4136/4144	4/0/4
A549	4	14450	14270	13/167	2065/2066	4136/4144	4/1/4
AGS	0	14892	14738	18/136	2126/2128	4262/4264	4/2/2
AGS	1	14879	14719	18/142	2139/2141	4262/4264	4/2/2
AGS	2	14883	14730	18/135	2135/2137	4262/4264	4/2/2
AGS	3	14873	14713	16/144	2143/2147	4262/4264	4/4/2
AGS	4	14881	14712	17/152	2136/2139	4262/4264	4/2/2
BICR6	0	14811	14623	17/171	2119/2123	4248/4249	4/3/1
BICR6	1	14799	14625	19/155	2133/2135	4248/4249	4/2/1
BICR6	2	14807	14629	21/157	2127/2127	4248/4249	4/0/1
BICR6	3	14809	14622	18/169	2122/2125	4248/4249	4/3/1
BICR6	4	14814	14637	16/161	2115/2120	4248/4249	4/4/1
ES2	0	13959	13709	17/233	2001/2003	3991/3995	4/2/4
ES2	1	13956	13710	12/234	1999/2006	3991/3995	4/3/4
ES2	2	13964	13703	17/244	1996/1998	3991/3995	4/2/4
ES2	3	13966	13728	17/221	1994/1996	3991/3995	4/2/4

Continued on next page

Table 22 continued

Cell line	Fit	Train declared	Train used	Train exclusions (GO/rep.)	Validation evaluable/declared	Test evaluable/declared	GO-excluded targets (tr/val/te)
ES2	4	13964	13707	16/241	1995/1998	3991/3995	4/2/4
HT29	0	14354	14159	16/179	2060/2062	4103/4109	4/2/3
HT29	1	14360	14157	15/188	2053/2056	4103/4109	4/3/3
HT29	2	14355	14155	17/183	2060/2061	4103/4109	4/1/3
HT29	3	14355	14160	15/180	2058/2061	4103/4109	4/2/3
HT29	4	14361	14172	17/172	2054/2055	4103/4109	4/1/3
MCF7	0	13118	13047	16/55	1883/1886	3766/3770	4/2/3
MCF7	1	13114	13034	16/64	1887/1890	3766/3770	4/2/3
MCF7	2	13124	13038	16/70	1877/1880	3766/3770	4/2/3
MCF7	3	13127	13053	17/57	1875/1877	3766/3770	4/1/3
MCF7	4	13120	13026	19/75	1884/1884	3766/3770	4/0/3
PC3	0	14845	14675	21/149	2135/2136	4247/4248	4/1/1
PC3	1	14852	14666	21/165	2128/2129	4247/4248	4/1/1
PC3	2	14845	14684	20/141	2134/2136	4247/4248	4/2/1
PC3	3	14845	14683	17/145	2131/2136	4247/4248	4/4/1
PC3	4	14853	14656	20/177	2126/2128	4247/4248	4/2/1
U251MG	0	15887	15781	21/85	2276/2277	4547/4548	4/1/1
U251MG	1	15891	15794	19/78	2270/2273	4547/4548	4/3/1
U251MG	2	15893	15787	18/88	2267/2271	4547/4548	4/4/1
U251MG	3	15887	15784	20/83	2275/2277	4547/4548	4/2/1
U251MG	4	15885	15795	20/70	2277/2279	4547/4548	4/2/1
YAPC	0	14082	13851	13/218	2015/2020	4028/4033	4/3/3
YAPC	1	14081	13872	16/193	2019/2021	4028/4033	4/1/3
YAPC	2	14087	13860	16/211	2013/2015	4028/4033	4/1/3
YAPC	3	14088	13865	17/206	2013/2014	4028/4033	4/1/3
YAPC	4	14085	13866	16/203	2015/2017	4028/4033	4/2/3

E.2 Complete fixed-test predictive results

Supplementary Figure 33 shows the cell-line-specific response-profile correlations for all six models, while Supplementary Table 23 summarizes predictive performance across the ten cell lines. Supplementary Table 24 reports the complete cell-line and method-specific results. Five independently trained fits per line share the same held-out test; their mean and standard deviation therefore quantify sensitivity to the train/validation split rather than independent test uncertainty. Supplementary Table 25 gives the predeclared RISCAL-minus-baseline contrasts after alignment on original test indices. Positive perturbation-response R^2 and negative macro intervention MSE differences favor RISCAL. Control identities are the resampling units, and the same block operation is applied jointly to all five fits. GEARS exclusions are kept explicit in Supplementary Table 22 rather than imposed on the descriptive results of other methods.

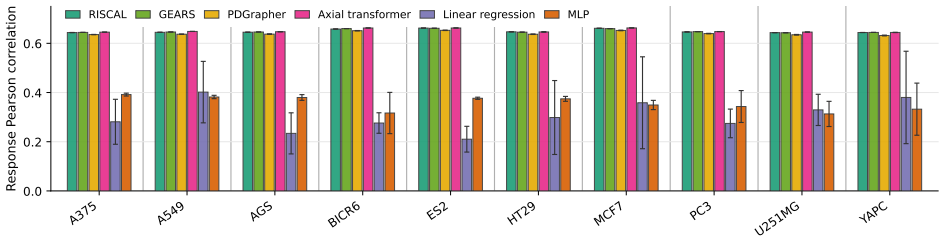

Figure 33: Samplewise response-profile Pearson correlation for all six models. Bars show the mean pm sample standard deviation across five fits per cell line.

Table 23: Aggregate perturbation-response prediction across ten cell lines. Values are unweighted means ± standard deviations across cell lines after averaging the five fits within each line. GEARS is evaluated on its natively supported test profiles.

Model	Treated-profile Pearson ↑	Response-profile Pearson ↑	Macro intervention MSE ↓
RISCAL	0.9794 ± 0.0018	0.6495 ± 0.0076	0.0075 ± 0.0007
Linear regression	0.8417 ± 0.0437	0.3043 ± 0.0622	0.1446 ± 0.0814
MLP	0.9471 ± 0.0194	0.3558 ± 0.0287	0.2379 ± 0.0550
Axial transformer	0.9794 ± 0.0018	0.6512 ± 0.0078	0.0075 ± 0.0007
GEARS	0.9794 ± 0.0018	0.6497 ± 0.0074	0.0075 ± 0.0007
PDGrapher	0.9790 ± 0.0019	0.6410 ± 0.0080	0.0076 ± 0.0007

Table 24: Fixed-test predictive results by cell line and method. Values are mean ± sample standard deviation across the five train/validation fits.

Cell line	Model	Treated-profile Pearson	Response-profile Pearson	Macro intervention MSE	Non-target MSE
A375	RISCAL	0.97694 ± 0.00002	0.6437 ± 0.0004	0.008512 ± 0.000007	0.0087 ± 0.0000
A375	Linear regression	0.81652 ± 0.09087	0.2810 ± 0.0914	0.152930 ± 0.119556	0.1548 ± 0.1191
A375	MLP	0.95992 ± 0.00021	0.3916 ± 0.0058	0.215944 ± 0.024032	0.2164 ± 0.0240
A375	Axial transformer	0.97701 ± 0.00000	0.6454 ± 0.0000	0.008488 ± 0.000000	0.0086 ± 0.0000
A375	GEARS	0.97695 ± 0.00000	0.6444 ± 0.0001	0.008507 ± 0.000001	0.0087 ± 0.0000
A375	PDGrapher	0.97656 ± 0.00003	0.6356 ± 0.0006	0.008661 ± 0.000012	0.0088 ± 0.0000
A549	RISCAL	0.97748 ± 0.00001	0.6447 ± 0.0002	0.008339 ± 0.000004	0.0084 ± 0.0000
A549	Linear regression	0.91359 ± 0.06462	0.4017 ± 0.1249	0.047350 ± 0.046904	0.0471 ± 0.0464
A549	MLP	0.95793 ± 0.00025	0.3818 ± 0.0068	0.219616 ± 0.037324	0.2197 ± 0.0373
A549	Axial transformer	0.97761 ± 0.00000	0.6483 ± 0.0001	0.008286 ± 0.000001	0.0083 ± 0.0000
A549	GEARS	0.97749 ± 0.00000	0.6460 ± 0.0001	0.008327 ± 0.000002	0.0084 ± 0.0000
A549	PDGrapher	0.97710 ± 0.00002	0.6370 ± 0.0004	0.008478 ± 0.000008	0.0085 ± 0.0000
AGS	RISCAL	0.98036 ± 0.00001	0.6451 ± 0.0004	0.007359 ± 0.000005	0.0072 ± 0.0000
AGS	Linear regression	0.81308 ± 0.07969	0.2339 ± 0.0835	0.143409 ± 0.089532	0.1400 ± 0.0873
AGS	MLP	0.96492 ± 0.00047	0.3798 ± 0.0117	0.195186 ± 0.043256	0.1952 ± 0.0433
AGS	Axial transformer	0.98041 ± 0.00000	0.6471 ± 0.0001	0.007343 ± 0.000001	0.0072 ± 0.0000
AGS	GEARS	0.98036 ± 0.00000	0.6460 ± 0.0001	0.007354 ± 0.000002	0.0072 ± 0.0000
AGS	PDGrapher	0.98004 ± 0.00002	0.6374 ± 0.0011	0.007485 ± 0.000012	0.0074 ± 0.0000
BICR6	RISCAL	0.98133 ± 0.00002	0.6581 ± 0.0005	0.006522 ± 0.000007	0.0064 ± 0.0000
BICR6	Linear regression	0.84880 ± 0.04556	0.2760 ± 0.0417	0.093816 ± 0.046723	0.0917 ± 0.0456
BICR6	MLP	0.92095 ± 0.05778	0.3167 ± 0.0838	0.212995 ± 0.025695	0.2130 ± 0.0257
BICR6	Axial transformer	0.98149 ± 0.00000	0.6623 ± 0.0000	0.006463 ± 0.000000	0.0063 ± 0.0000
BICR6	GEARS	0.98140 ± 0.00000	0.6599 ± 0.0001	0.006493 ± 0.000001	0.0064 ± 0.0000
BICR6	PDGrapher	0.98108 ± 0.00003	0.6513 ± 0.0007	0.006615 ± 0.000011	0.0065 ± 0.0000
ES2	RISCAL	0.98220 ± 0.00000	0.6620 ± 0.0001	0.007202 ± 0.000002	0.0073 ± 0.0000
ES2	Linear regression	0.75486 ± 0.07553	0.2102 ± 0.0524	0.273490 ± 0.140357	0.2798 ± 0.1437
ES2	MLP	0.96437 ± 0.00032	0.3766 ± 0.0044	0.212036 ± 0.023156	0.2122 ± 0.0232
ES2	Axial transformer	0.98221 ± 0.00000	0.6623 ± 0.0001	0.007200 ± 0.000001	0.0073 ± 0.0000
ES2	GEARS	0.98216 ± 0.00000	0.6611 ± 0.0001	0.007217 ± 0.000002	0.0073 ± 0.0000
ES2	PDGrapher	0.98185 ± 0.00003	0.6528 ± 0.0008	0.007352 ± 0.000013	0.0075 ± 0.0000
HT29	RISCAL	0.98031 ± 0.00000	0.6468 ± 0.0001	0.007086 ± 0.000001	0.0071 ± 0.0000
HT29	Linear regression	0.83163 ± 0.11991	0.2984 ± 0.1503	0.131821 ± 0.120305	0.1300 ± 0.1193
HT29	MLP	0.95948 ± 0.00052	0.3742 ± 0.0100	0.169260 ± 0.039502	0.1693 ± 0.0395
HT29	Axial transformer	0.98029 ± 0.00005	0.6463 ± 0.0013	0.007097 ± 0.000020	0.0071 ± 0.0000
HT29	GEARS	0.98025 ± 0.00001	0.6453 ± 0.0001	0.007102 ± 0.000002	0.0071 ± 0.0000
HT29	PDGrapher	0.97990 ± 0.00005	0.6369 ± 0.0010	0.007235 ± 0.000016	0.0072 ± 0.0000
MCF7	RISCAL	0.97983 ± 0.00001	0.6611 ± 0.0002	0.006736 ± 0.000002	0.0067 ± 0.0000
MCF7	Linear regression	0.83136 ± 0.19834	0.3584 ± 0.1868	0.280595 ± 0.528213	0.2710 ± 0.5060
MCF7	MLP	0.95391 ± 0.00799	0.3493 ± 0.0188	0.334736 ± 0.014428	0.3347 ± 0.0144
MCF7	Axial transformer	0.97989 ± 0.00000	0.6626 ± 0.0001	0.006719 ± 0.000001	0.0067 ± 0.0000
MCF7	GEARS	0.97973 ± 0.00001	0.6597 ± 0.0001	0.006766 ± 0.000002	0.0067 ± 0.0000
MCF7	PDGrapher	0.97943 ± 0.00006	0.6523 ± 0.0011	0.006873 ± 0.000017	0.0068 ± 0.0000
PC3	RISCAL	0.97924 ± 0.00003	0.6463 ± 0.0007	0.007237 ± 0.000008	0.0071 ± 0.0000
PC3	Linear regression	0.86236 ± 0.04783	0.2744 ± 0.0582	0.073880 ± 0.039423	0.0747 ± 0.0408
PC3	MLP	0.94419 ± 0.04615	0.3431 ± 0.0648	0.305423 ± 0.062036	0.3053 ± 0.0620
PC3	Axial transformer	0.97928 ± 0.00000	0.6478 ± 0.0001	0.007227 ± 0.000002	0.0071 ± 0.0000
PC3	GEARS	0.97925 ± 0.00000	0.6475 ± 0.0001	0.007230 ± 0.000001	0.0071 ± 0.0000
PC3	PDGrapher	0.97894 ± 0.00003	0.6396 ± 0.0010	0.007344 ± 0.000013	0.0073 ± 0.0000
U251MG	RISCAL	0.97939 ± 0.00001	0.6434 ± 0.0002	0.007647 ± 0.000002	0.0078 ± 0.0000
U251MG	Linear regression	0.88820 ± 0.04216	0.3294 ± 0.0635	0.067528 ± 0.038878	0.0710 ± 0.0423
U251MG	MLP	0.93674 ± 0.02717	0.3130 ± 0.0512	0.301456 ± 0.035764	0.3016 ± 0.0358
U251MG	Axial transformer	0.97946 ± 0.00000	0.6457 ± 0.0000	0.007618 ± 0.000000	0.0078 ± 0.0000
U251MG	GEARS	0.97933 ± 0.00001	0.6430 ± 0.0001	0.007664 ± 0.000002	0.0078 ± 0.0000
U251MG	PDGrapher	0.97904 ± 0.00005	0.6348 ± 0.0013	0.007787 ± 0.000021	0.0080 ± 0.0000
YAPC	RISCAL	0.97678 ± 0.00001	0.6439 ± 0.0002	0.008091 ± 0.000002	0.0080 ± 0.0000
YAPC	Linear regression	0.85606 ± 0.17721	0.3800 ± 0.1878	0.181651 ± 0.335440	0.1808 ± 0.3341
YAPC	MLP	0.90848 ± 0.07502	0.3323 ± 0.1059	0.212762 ± 0.017720	0.2127 ± 0.0177
YAPC	Axial transformer	0.97679 ± 0.00003	0.6442 ± 0.0006	0.008092 ± 0.000011	0.0080 ± 0.0000
YAPC	GEARS	0.97679 ± 0.00000	0.6443 ± 0.0000	0.008088 ± 0.000001	0.0080 ± 0.0000
YAPC	PDGrapher	0.97620 ± 0.00010	0.6318 ± 0.0016	0.008298 ± 0.000029	0.0082 ± 0.0000

Table 25: Paired fixed-test contrasts (RISCAL minus baseline). Positive response R^2 and negative macro intervention MSE differences favor RISCAL. Intervals and Holm-adjusted p-values use 10,000 control-block bootstrap and joint sign-permutation samples, respectively.

Cell	Baseline	Δ perturbation-response R^2	95% CI	Holm p	Δ macro intervention MSE	95% CI	Holm p	Blocks	Profiles
A375	Linear regression	9.7081	[8.5822, 10.8996]	< 0.001	−0.144418	[−0.163010, −0.127126]	< 0.001	273	4359
A375	MLP	13.7981	[13.5331, 14.0585]	< 0.001	−0.207432	[−0.214133, −0.200829]	< 0.001	273	4359
A375	Axial transformer	−0.0016	[−0.0019, −0.0013]	< 0.001	0.000024	[0.000019, 0.000028]	< 0.001	273	4359
A375	GEARS	−0.0005	[−0.0009, −0.0001]	0.010	0.000006	[−0.000000, 0.000013]	0.056	273	4356
A375	PDGrapher	0.0098	[0.0093, 0.0103]	< 0.001	−0.000149	[−0.000158, −0.000140]	< 0.001	273	4359
A549	Linear regression	2.6350	[2.3059, 2.9923]	< 0.001	−0.039011	[−0.044867, −0.033760]	< 0.001	262	4144
A549	MLP	14.4000	[14.0889, 14.7183]	< 0.001	−0.211277	[−0.219138, −0.203634]	< 0.001	262	4144
A549	Axial transformer	−0.0036	[−0.0042, −0.0030]	< 0.001	0.000053	[0.000044, 0.000062]	< 0.001	262	4144
A549	GEARS	−0.0009	[−0.0016, −0.0001]	0.029	0.000014	[0.000001, 0.000025]	0.028	262	4136
A549	PDGrapher	0.0095	[0.0089, 0.0102]	< 0.001	−0.000139	[−0.000149, −0.000129]	< 0.001	262	4144
AGS	Linear regression	10.7191	[9.5307, 11.9833]	< 0.001	−0.136050	[−0.154248, −0.119904]	< 0.001	270	4264
AGS	MLP	15.1739	[14.8622, 15.4918]	< 0.001	−0.187827	[−0.193532, −0.182169]	< 0.001	270	4264
AGS	Axial transformer	−0.0016	[−0.0020, −0.0011]	< 0.001	0.000016	[0.000010, 0.000023]	< 0.001	270	4264
AGS	GEARS	−0.0006	[−0.0016, 0.0003]	0.192	0.000007	[−0.000006, 0.000020]	0.280	270	4262
AGS	PDGrapher	0.0097	[0.0091, 0.0105]	< 0.001	−0.000126	[−0.000137, −0.000116]	< 0.001	270	4264
BICR6	Linear regression	7.5070	[6.7315, 8.3597]	< 0.001	−0.087294	[−0.097694, −0.077439]	< 0.001	270	4249
BICR6	MLP	18.1870	[17.8570, 18.5122]	< 0.001	−0.206474	[−0.211546, −0.201351]	< 0.001	270	4249
BICR6	Axial transformer	−0.0053	[−0.0058, −0.0049]	< 0.001	0.000059	[0.000053, 0.000064]	< 0.001	270	4249
BICR6	GEARS	−0.0025	[−0.0032, −0.0018]	< 0.001	0.000028	[0.000020, 0.000037]	< 0.001	270	4248
BICR6	PDGrapher	0.0079	[0.0074, 0.0084]	< 0.001	−0.000093	[−0.000100, −0.000087]	< 0.001	270	4249
ES2	Linear regression	20.3098	[17.6966, 23.2197]	< 0.001	−0.266288	[−0.303242, −0.231656]	< 0.001	258	3995
ES2	MLP	15.2670	[14.9719, 15.5673]	< 0.001	−0.204834	[−0.211370, −0.198530]	< 0.001	258	3995
ES2	Axial transformer	−0.0003	[−0.0005, −0.0001]	0.003	0.000002	[−0.000001, 0.000005]	0.130	258	3995
ES2	GEARS	0.0009	[0.0006, 0.0013]	< 0.001	−0.000014	[−0.000019, −0.000009]	< 0.001	258	3991
ES2	PDGrapher	0.0110	[0.0106, 0.0115]	< 0.001	−0.000150	[−0.000157, −0.000144]	< 0.001	258	3995
HT29	Linear regression	9.9947	[9.0061, 11.0595]	< 0.001	−0.124735	[−0.139210, −0.111667]	< 0.001	264	4109
HT29	MLP	13.1882	[12.9177, 13.4673]	< 0.001	−0.162175	[−0.167235, −0.157320]	< 0.001	264	4109
HT29	Axial transformer	0.0009	[0.0006, 0.0012]	< 0.001	−0.000011	[−0.000015, −0.000008]	< 0.001	264	4109
HT29	GEARS	0.0015	[0.0006, 0.0023]	< 0.001	−0.000015	[−0.000026, −0.000003]	0.013	264	4103
HT29	PDGrapher	0.0121	[0.0116, 0.0125]	< 0.001	−0.000149	[−0.000156, −0.000142]	< 0.001	264	4109
MCF7	Linear regression	22.0383	[18.0603, 26.6787]	< 0.001	−0.273858	[−0.336101, −0.220976]	< 0.001	241	3770
MCF7	MLP	27.3472	[26.7506, 27.9396]	< 0.001	−0.328000	[−0.337312, −0.318533]	< 0.001	241	3770
MCF7	Axial transformer	−0.0015	[−0.0018, −0.0012]	< 0.001	0.000017	[0.000013, 0.000021]	< 0.001	241	3770
MCF7	GEARS	0.0027	[0.0015, 0.0038]	< 0.001	−0.000028	[−0.000044, −0.000013]	< 0.001	241	3766
MCF7	PDGrapher	0.0112	[0.0107, 0.0118]	< 0.001	−0.000137	[−0.000145, −0.000130]	< 0.001	241	3770
PC3	Linear regression	5.4331	[4.6115, 6.4767]	< 0.001	−0.066643	[−0.077477, −0.057263]	< 0.001	272	4248
PC3	MLP	23.9675	[23.4519, 24.4681]	< 0.001	−0.298186	[−0.307712, −0.288475]	< 0.001	272	4248
PC3	Axial transformer	−0.0010	[−0.0014, −0.0006]	< 0.001	0.000010	[0.000006, 0.000015]	< 0.001	272	4248
PC3	GEARS	−0.0008	[−0.0016, 0.0001]	0.087	0.000008	[−0.000002, 0.000018]	0.112	272	4247
PC3	PDGrapher	0.0085	[0.0080, 0.0090]	< 0.001	−0.000107	[−0.000114, −0.000100]	< 0.001	272	4248
U251MG	Linear regression	4.7358	[3.6375, 6.5305]	< 0.001	−0.059882	[−0.079174, −0.047479]	< 0.001	294	4548
U251MG	MLP	22.0059	[21.6024, 22.4199]	< 0.001	−0.293810	[−0.302622, −0.284879]	< 0.001	294	4548
U251MG	Axial transformer	−0.0022	[−0.0025, −0.0019]	< 0.001	0.000028	[0.000025, 0.000032]	< 0.001	294	4548
U251MG	GEARS	0.0014	[0.0006, 0.0022]	< 0.001	−0.000016	[−0.000027, −0.000005]	0.005	294	4547
U251MG	PDGrapher	0.0105	[0.0101, 0.0109]	< 0.001	−0.000141	[−0.000147, −0.000135]	< 0.001	294	4548
YAPC	Linear regression	12.5240	[10.4922, 14.7650]	< 0.001	−0.173560	[−0.204406, −0.145584]	< 0.001	257	4033
YAPC	MLP	14.8310	[14.4930, 15.1734]	< 0.001	−0.204671	[−0.212285, −0.197162]	< 0.001	257	4033
YAPC	Axial transformer	−0.0001	[−0.0005, 0.0002]	0.535	−0.000000	[−0.000006, 0.000005]	1.000	257	4033
YAPC	GEARS	−0.0004	[−0.0010, 0.0003]	0.535	0.000003	[−0.000007, 0.000012]	1.000	257	4028
YAPC	PDGrapher	0.0148	[0.0142, 0.0154]	< 0.001	−0.000207	[−0.000217, −0.000196]	< 0.001	257	4033

F Evaluation Metrics and Statistical Analysis

Structural recovery is evaluated only on synthetic datasets with known ground-truth directed acyclic graphs (DAGs); predictive metrics are evaluated on held-out treated expression profiles. Ground-truth DAGs, oracle quantities, and held-out profiles are never used for training, checkpoint selection, hyperparameter selection, or selection of an evaluation threshold.

F.1 Causal-discovery metrics

Let $G^* = ([G], E^*)$ denote a synthetic ground-truth DAG, and let $A^{(b)}$ be the binary adjacency of a selected model checkpoint, with edge set $\widehat{E} = \{(i,j) : A^{(b)}_{ij} = 1\}$. A true positive (TP) is an edge in $\widehat{E} \cap E^*$, a false positive (FP) is an edge in $\widehat{E} \setminus E^*$, and a false negative (FN) is an edge in $E^* \setminus \widehat{E}$.

DAG count. For a declared multi-seed cohort, DAGs / S is the number of seeds whose official binary adjacency $A^{(b)}$ is acyclic, out of the S completed seeds. It records feasibility of the reported selected edge set rather than a continuous training surrogate.

AUPRC. Area under the precision-recall curve (AUPRC) evaluates the ranking of all possible directed edges against E^*. RISCAL ranks edges by continuous neural path strength A_{ij}; each baseline uses its declared continuous native score. AUPRC is independent of the final evaluation threshold and therefore complements thresholded-support metrics.

Precision, recall, and F1. For the official selected edge set,

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \qquad \text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \qquad \text{F1} = \frac{2\,\text{precision}\,\text{recall}}{\text{precision} + \text{recall}}. \tag{114}$$

Precision measures selected-edge specificity, recall measures recovered generating edges, and F1 summarizes their balance. When an oracle F1-maximizing threshold is shown, it is an explicitly labelled post-hoc synthetic diagnostic that changes only the displayed final selected edge set and is never an official result.

Predicted-to-true edge ratio. The selected-edge density diagnostic is

$$\frac{|\widehat{E}|}{|E^*|}. \tag{115}$$

It is computed within each seed before aggregation, so every ground-truth DAG has equal weight. It helps distinguish a high-recall but excessively dense support from a sparse selected edge set with similar thresholded scores.

F.2 Perturbation-response prediction metrics

For N held-out profiles, let x^d_n, x^t_n, and $\widehat{x}^t_n$ denote the control, treated, and predicted treated expression profiles, respectively. Let k_n be the intervention target and $e_n = \widehat{x}^t_n - x^t_n$. Unless stated otherwise, metrics aggregate over all held-out profiles and measured genes.

Treated MSE and RMSE.

$$\text{MSE}_{\text{treated}} = \frac{1}{NG} \sum_{n=1}^{N} \sum_{j=1}^{G} e^2_{nj}, \qquad \text{RMSE}_{\text{treated}} = \sqrt{\text{MSE}_{\text{treated}}}. \tag{116}$$

These evaluate absolute agreement with the measured treated expression profile.

Treated-profile R^2. Let $\overline{x}^t_{\cdot j} = N^{-1} \sum_n x^t_{nj}$. The treated-profile coefficient of determination is

$$R^2_{\text{treated}} = 1 - \frac{\sum_{n,j} e^2_{nj}}{\sum_{n,j} (x^t_{nj} - \overline{x}^t_{\cdot j})^2}. \tag{117}$$

It is a descriptive secondary metric for real data and predictive context for synthetic causal experiments.

Perturbation-response R^2. Let $\Delta x_n = x_n^t - x_n^d$ and $\Delta \widehat{x}_n = \widehat{x}_n^t - x_n^d$. We compute the same coefficient of determination using Δx as the outcome and $\Delta \widehat{x}$ as the prediction. This evaluates how well a model predicts the perturbation-induced response relative to carrying the paired control forward unchanged. It is a predeclared primary real-data endpoint.

Macro intervention MSE. Let $\mathcal{K}$ be the set of targets represented in the held-out split and $N_k = |\{n : k_n = k\}|$. We give every intervention target equal weight,

$$\mathrm{MSE}_{\mathrm{macro}} = \frac{1}{|\mathcal{K}|} \sum_{k \in \mathcal{K}} \frac{1}{N_k G} \sum_{n:k_n=k} \sum_{j=1}^{G} e_{nj}^2. \tag{118}$$

This prevents frequently observed intervention targets from dominating the reported error and is the other predeclared primary real-data endpoint.

Target and non-target MSE. Target-gene MSE averages e_{n,k_n}^2. Non-target-gene MSE averages e_{nj}^2 over $j \neq k_n$. They diagnose the separately modelled target-expression readout and the downstream intervention propagation, respectively.

Samplewise Pearson correlations. For each held-out profile, Pearson correlation is computed across genes after centering that profile, then averaged over profiles with nonzero variance. Treated-profile Pearson compares $\widehat{x}^t$ with x^t; response-profile Pearson compares $\Delta \widehat{x}$ with Δx. These are secondary shape-of-profile diagnostics rather than substitutes for error metrics.

F.3 Achievable synthetic R^2

Monte Carlo achievable-R^2 estimate. For synthetic data, the frozen generator defines a stochastic conditional law $p_{\mathrm{gen}}(x^t \mid x^d, k)$. For every held-out (x_n^d, k_n), the versioned oracle draws generator randomness, including any exogenous noise, residual-activity variables, and target-expression readout randomness declared by that generator, and estimates the conditional mean by Monte Carlo. If $\widetilde{x}_n^{t,(r)}$ is the mean of the samples in replicate r, the oracle prediction is

$$\widehat{x}_n^{t,\mathrm{oracle}} = \frac{1}{R} \sum_{r=1}^{R} \widetilde{x}_n^{t,(r)}, \qquad \widehat{R}^2_{\max} = 1 - \frac{\sum_{n,j} (x_{nj}^t - \widehat{x}_{nj}^{t,\mathrm{oracle}})^2}{\sum_{n,j} (x_{nj}^t - \overline{x}_{\cdot j}^t)^2}. \tag{119}$$

The sample count, replicate count, seed derivation, and replicate-range stability criterion are recorded in the immutable generator oracle contract. The oracle is run after model selection and reads no trained-model state.

Because $\widehat{R}^2_{\max}$ is a finite-sample Monte Carlo estimate scored against one realized stochastic treated expression profile, it is not a deterministic upper bound for every evaluated model on that split. Accordingly, $R^2_{\mathrm{treated}}/\widehat{R}^2_{\max}$ can be slightly larger than one (for example, 1.001) through finite-sample and Monte Carlo variability. Such a value does not mean that a learned model exceeds the population conditional mean; the ratio is descriptive context only and is never used for fitting or selection.

F.4 Residual-activity diagnostics

Residual-activity evaluation is a synthetic diagnostic rather than a graph or profile-prediction metric. On the predeclared generator-derived cohort of genes with at least one true direct child, we compare the learned and generating mean residual activity, its across-gene standard deviation, shared Beta concentration, and mean implied Beta variance. These distributional summaries do not claim recovery of a gene-specific biological activity measurement.

F.5 Exogenous-regulator case-study metrics

The LINC00525 case study reuses the Pearson-correlation, edge-consensus, Benjamini-Hochberg, and Gene Ontology enrichment definitions in this appendix. Only its case-specific evaluation quantities are defined here.

Leave-one-replicate-out correlation. For each frozen A549 model and candidate sparsity coefficient, the new outgoing weights are fitted to the mean of two LINC00525-knockdown response profiles and scored on the third. The score is the Pearson correlation across evaluable genes between the predicted expression change and the held-out measured change, after centering each vector across genes. We average over the three held-out replicates to select the coefficient and report mean $\pm$ sample standard deviation across the five frozen models.

Gene-label-permutation null. A null realization independently permutes the gene labels of each measured response vector while retaining its empirical values, the frozen A549 models, the fitting procedure, and the number of trainable parameters. For a statistic T where larger values favor the observed mapping, its one-sided empirical value over B null realizations is

$$p_{\mathrm{emp}} = \frac{1 + \sum_{b=1}^{B} \mathbb{1}[T_b \geq T_{\mathrm{obs}}]}{B + 1}.$$

With the 25 declared null realizations, the minimum attainable value is $1/(25+1) = 0.0385$.

New-edge consensus and functional enrichment. New LINC00525 edges are selected with the same q_{eval} threshold as the frozen A549 graph. Primary targets are genes selected in all five augmented models; the 4/5 set is a sensitivity analysis. Gene Ontology enrichment uses the hypergeometric test and correction already defined in Appendix F.7.

F.6 Aggregation and paired statistical inference

Synthetic tables report mean $\pm$ sample standard deviation across completed dataset seeds, except DAG counts. Real-data descriptive metrics retain each method's native held-out coverage and report mean $\pm$ sample standard deviation across declared fits within cell line; raw error scales are never pooled across cell lines.

For predeclared RISCAL-versus-baseline contrasts, saved predictions are aligned on the pairwise overlap of original held-out indices. The reported contrast is computed with shared control identities as blocks, using a joint control-block sign-permutation test and control-block bootstrap confidence interval across fits. Holm correction is applied within each cell-line and primary-metric family. Coverage, excluded profiles, and the number of control blocks are reported with every inferential result.

Reported wall-clock runtimes are descriptive rather than hardware-normalized benchmarks because experiments were executed on shared computing resources.

F.7 Biological plausibility metrics

Edge-frequency consensus. For selected edge sets from the fits $\widehat{E}^{(1)}, \ldots, \widehat{E}^{(S)}$, edge frequency is

$$\chi_{ij}^{\mathrm{cons}} = S^{-1} \sum_{s=1}^{S} \mathbf{1}[(i,j) \in \widehat{E}^{(s)}]. \tag{120}$$

A k/S frequency-consensus candidate contains edges with $\chi_{ij}^{\mathrm{cons}} \geq k/S$. Candidates are processed in decreasing frequency and continuous-score order, and an edge is accepted only when adding it preserves acyclicity. Thus consensus level measures reproducibility; it is not selected using biological databases.

Directed-resource enrichment. For a consensus edge set E and eligible directed resource R, observed overlap is $O = |E \cap R|$. Degree-matched enrichment is

$$\mathrm{FE} = \frac{O}{B^{-1} \sum_{b=1}^{B} |E_b^{(0)} \cap R|}, \tag{121}$$

where $E_b^{(0)}$ is an acyclic null graph with exactly the same in/out-degree sequence. The one-sided Monte Carlo p value uses the plus-one correction; Benjamini-Hochberg-adjusted values (Benjamini & Hochberg, 1995) are denoted q. Signed agreement is computed only among overlapping edges with an eligible database sign and stable model-derived sign.

Functional enrichment and module matching. Louvain communities (Blondel et al., 2008) are computed on the undirected consensus projection. For each community, one-sided hypergeometric overlap with Reactome or Gene Ontology Biological Process sets is adjusted by Benjamini-Hochberg (Benjamini & Hochberg, 1995). Cross-cell gene similarity is Jaccard overlap. A functional profile assigns each significant term the capped score $\min\{20, -\log_{10} q\}$; functional similarity is the cosine between these term-score vectors. Conserved programs require reciprocal best matches and the frozen size, shared-term, cosine, and corrected-overlap criteria.

Local-effect sign. For edge $i \to j$, the profile-dependent local effect is the exact child-mechanism derivative

$$J_{ji}(z) = \frac{\partial f_{\theta,j}(z; A^{(b)})}{\partial z_i}. \tag{122}$$

Its sign is evaluated over held-out centered control expression profiles. An edge is profile-stable within a fit when at least 80% of eligible profiles share a sign and fit-stable when at least four of five fits share it. This sign is local to the fitted nonlinear mechanism and is not a globally calibrated activation or inhibition label.

Topology. Modularity (Newman, 2006) and transitivity are computed on the undirected consensus projection; approximate mean path length uses its largest connected component. Each degree-controlled ratio divides the observed statistic by the mean over degree-preserving acyclic nulls.

Real residual-activity summaries. For learned Beta parameters α_k, β_k, the gene mean is $\rho_k = \alpha_k/(\alpha_k + \beta_k)$, concentration is $\phi = \alpha_k + \beta_k$, and variance is $\alpha_k\beta_k/[\phi^2(\phi + 1)]$. We report fit summaries of the global mean, shared concentration, across-gene dispersion of ρ_k, and mean Beta variance. These are latent-model diagnostics rather than externally calibrated biological measurements.

G Model Ablations and Oracle Substitutions

We ablate the final RISCAL model in the PC3-scale benchmark using the same five generated datasets, training protocol, checkpoint rule, and fixed $q_{\text{eval}} = 0.09$ threshold as in the primary experiment. The results support three complementary conclusions. First, the SCM data-fit term and the acyclicity constraint are required to recover a predictive DAG, whereas target supervision is deliberately separated from structural selection. Second, centering, nonlinear child-specific mechanisms, residual target attenuation, and the target readout perform distinct functions that the remaining architecture cannot absorb without loss. Third, the soft gate and sparsity penalty provide partly overlapping control of the learned support, with the gate having the larger individual effect. Oracle substitutions then localize the remaining gap to the generating process. Together, these results provide an independent check of the components motivated by the synthetic construction.

Unless stated otherwise, each variant changes only the named component and is trained independently on seeds 42-46. We report mean $\pm$ sample standard deviation. AUPRC measures threshold-free edge ranking; F1 and the number of selected edges describe the support at the fixed threshold. Predictive quality is summarized by samplewise response-profile Pearson correlation, macro intervention MSE, and target-gene MSE. All definitions follow Appendix F.

G.1 Training-objective ablations

This experiment asks which terms in the final training objective, Eq. (7), are required for structural recovery and intervention prediction. Each variant sets one coefficient to zero without retuning the remaining recipe.

As shown in Supplementary Table 26, removing $\mathcal{L}_{\text{SCM}}$ causes path strengths to collapse below the fixed threshold and eliminates recovery of true edges, despite only a small change in aggregate prediction. The SCM term is therefore the principal signal for structural learning rather than merely another predictive loss. Removing acyclicity yields cyclic supports in all five seeds; their apparently similar thresholded F1 is driven by higher recall from a denser graph and does not satisfy the DAG contract.

Table 26: Training-objective ablations in the PC3-scale benchmark. Dashes mark predictive quantities not reported because removing acyclicity produced no eligible DAG in any seed.

Variant	AUPRC ↑	F1 ↑	Selected edges	DAGs / 5	Response-profile Pearson r ↑	Macro intervention MSE ↓	Target-gene MSE ↓
RISCAL	0.633 ± 0.016	0.649 ± 0.018	$14{,}144 \pm 291$	5/5	0.7006 ± 0.0018	0.00687 ± 0.00026	0.0168 ± 0.0008
$\alpha_{\text{SCM}} = 0$	0.001 ± 0.000	0.000 ± 0.000	0 ± 0	5/5	0.6973 ± 0.0016	0.00697 ± 0.00025	0.0166 ± 0.0008
$\alpha_{\text{target}} = 0$	0.633 ± 0.016	0.649 ± 0.018	$14{,}144 \pm 291$	5/5	0.7006 ± 0.0018	0.00687 ± 0.00026	0.0170 ± 0.0007
No acyclicity	0.526 ± 0.024	0.652 ± 0.022	$16{,}970 \pm 559$	0/5	-	-	-

The target loss is the only objective term that directly trains the measured target readout. When $\alpha_{\text{target}} = 0$, its scale and bias remain at their identity initialization, so the predicted target copies its paired control value, $\widehat{x}_k^t = x_k^d$. This strong baseline explains why target-gene MSE increases only from 0.0168 to 0.0170, approximately 1.2%, while the learned DAG remains unchanged. In contrast, the no-readout architectural ablation below shows that a separately trained readout is needed when measured target expression must be recovered without repurposing the effective activity propagated through the SCM. We therefore retain $\mathcal{L}_{\text{target}}$ despite its negligible effect on structural recovery.

G.2 Architectural-component ablations

We next ask whether the main architectural choices capture distinct aspects of the intervention rather than merely adding capacity. Each architectural ablation changes one component while retaining the same training objective, graph parameterization, checkpoint rule, and evaluation threshold:

- **No centering, with biases.** We set the learned expression baseline to zero, $\mu = 0$, so the mechanisms operate directly on uncentered expression profiles. To preserve ordinary affine flexibility, we add an ablation-specific hidden bias $\beta_j^{(h)}$ and output bias $\beta_j^{(o)}$ to every

child-specific mechanism:

$$f_{\theta,j}^{\text{no-center}}(x; A^{(\cdot)}) = v_j^\top \tanh\left(\sum_{i \neq j} A_{ij}^{(\cdot)} U_{j,\cdot i} x_i + \beta_j^{(h)}\right) + \beta_j^{(o)}. \tag{123}$$

This ensures that the comparison tests the baseline-relative coordinate system rather than merely removing ordinary affine offsets.

- **Linear mechanisms.** Each nonlinear child-specific mechanism is replaced by a masked linear aggregation of its candidate parents:

$$f_{\theta,j}^{\text{linear}}(z; A^{(\cdot)}) = \sum_{i \neq j} A_{ij}^{(\cdot)} w_{ji} z_i. \tag{124}$$

The path-strength graph, soft gate, residual-activity model, training objective, and DAG checkpoint selection remain unchanged. This isolates the contribution of nonlinear conditional mechanisms.

- **No target readout.** We remove the separate mapping from the paired control profile to measured target expression. The effective target activity propagated through the SCM is therefore also used as the measured target prediction, $\widehat{x}_k^t = x_k^{t,\text{eff}}$. This tests whether one quantity can simultaneously represent functional activity and measured target expression.
- **No target attenuation.** We fix $\rho_k = 1$, making the intervened target use its unattenuated natural mechanism output, as for a non-intervened structural assignment. Any intervention-specific reduction in its effective activity must therefore be inferred indirectly by the target's child-specific mechanism from its expression inputs, because no explicit attenuation remains.

Supplementary Table 27 shows that removing centering causes both structural and predictive failure even after adding biases, supporting the baseline-relative coordinate system used by the mechanisms. Linear mechanisms preserve much of the aggregate predictive signal but approximately halve structural AUPRC and substantially reduce recall, showing that nonlinear conditional mechanisms matter most for distinguishing direct dependencies. Fixing $\rho_k = 1$ also degrades graph recovery, showing that the child-specific mechanism cannot reliably absorb this intervention-specific change and supporting an explicit residual-activity variable.

Table 27: Architectural-component ablations in the PC3-scale benchmark.

Variant	AUPRC ↑	F1 ↑	Selected edges	Response-profile Pearson r ↑	Macro intervention MSE ↓	Target-gene MSE ↓
RISCAL	0.633 ± 0.016	0.649 ± 0.018	$14{,}144 \pm 291$	0.7006 ± 0.0018	0.00687 ± 0.00026	0.0168 ± 0.0008
No centering, with biases	0.002 ± 0.001	0.008 ± 0.007	$37{,}551 \pm 6{,}768$	0.5752 ± 0.2622	0.10688 ± 0.22325	0.0168 ± 0.0008
Linear mechanisms	0.294 ± 0.024	0.426 ± 0.024	$5{,}803 \pm 444$	0.6988 ± 0.0016	0.00693 ± 0.00025	0.0168 ± 0.0008
No target readout	0.632 ± 0.016	0.649 ± 0.018	$14{,}118 \pm 252$	0.6741 ± 0.0013	0.00740 ± 0.00026	1.9689 ± 0.0234
No target attenuation ($\rho_k = 1$)	0.513 ± 0.025	0.545 ± 0.025	$14{,}705 \pm 271$	0.6972 ± 0.0015	0.00697 ± 0.00025	0.0168 ± 0.0008

Without the target readout, the model must use the same quantity both as the effective target activity propagated through the SCM and as the prediction of measured target expression. These roles need not coincide. The model could move this quantity toward the measured transcript, potentially distorting the causal mechanisms and the learned DAG, or preserve the effective activity required downstream and accept worse target prediction. The results support the second behavior: graph recovery is almost unchanged, whereas target-gene MSE increases approximately 117-fold. Consistently, the effective target activities learned by the complete and ablated models remain highly correlated on 256 common test profiles per seed (0.9897 ± 0.0005; MAE 0.0884 ± 0.0174). The separate readout therefore allows RISCAL to predict measured target expression without replacing the effective activity transmitted through the causal mechanisms.

G.3 Edge-control ablations

The soft gate and sparsity penalty both act on weak paths: the gate controls their participation in the mechanisms, whereas $\lambda_{\text{edges}} \mathcal{R}_{\text{sparse}}(A)$ penalizes their aggregate strength. We remove each

individually and then jointly to test their complementarity. The fixed q_{eval} remains unchanged and determines the reported binary support in every variant.

The fixed q_{eval} continues to remove weak paths in all variants, which limits changes in the reported edge count. As shown in Supplementary Table 28, removing either edge-control mechanism lowers threshold-free AUPRC, and removing both gives the weakest edge ranking. Removing the soft gate degrades structural recovery, reducing AUPRC from 0.633 to 0.598 and shifting the selected support toward higher precision but substantially lower recall. The gate has the larger individual effect, whereas the sparsity penalty provides a smaller complementary improvement. Thus, the two mechanisms overlap in controlling weak edges but are not interchangeable.

Table 28: Individual and joint removal of the two edge-control mechanisms.

Variant	AUPRC ↑	Precision ↑	Recall ↑	F1 ↑	Selected edges
RISCAL	0.633 ± 0.016	0.640 ± 0.022	0.659 ± 0.015	0.649 ± 0.018	$14{,}144 \pm 291$
No sparsity ($\lambda_{\mathrm{edges}} = 0$)	0.620 ± 0.020	0.629 ± 0.018	0.646 ± 0.024	0.637 ± 0.017	$14{,}117 \pm 521$
No soft gate	0.598 ± 0.030	0.709 ± 0.013	0.591 ± 0.043	0.644 ± 0.027	$11{,}460 \pm 864$
Neither	0.594 ± 0.029	0.691 ± 0.020	0.595 ± 0.044	0.638 ± 0.025	$11{,}844 \pm 1{,}026$

G.4 Oracle substitutions and error decomposition

Finally, we replace selected learned quantities with their generating values to ask where the remaining error originates. These are diagnostic substitutions, not competing models: supplying the true graph makes structural metrics perfect by construction, so only predictive and likelihood quantities are informative. Supplying the true mechanisms also requires the true graph because the generator mechanisms are defined on its parent sets.

Supplementary Table 29 shows that supplying the true residual activities alone has negligible effect on prediction and only a small structural effect (AUPRC 0.638 ± 0.016 versus 0.633 ± 0.016), so their exact gene-wise recovery is not the principal bottleneck. The true graph provides the largest likelihood improvement, and true mechanisms yield a further but smaller predictive gain. Even the fully specified oracle improves response correlation only from 0.7006 to 0.7043, showing that the learned model already approaches the predictive regime permitted by the generator, while the larger likelihood change localizes the remaining discrepancy to fine-grained structural and conditional-mechanism estimation.

Table 29: Oracle substitutions in the PC3-scale benchmark. L and T denote learned and true quantities, respectively.

Variant	Graph	Mechanisms	ρ	Response-profile Pearson r ↑	Macro intervention MSE ↓	SCM NLL ↓
RISCAL	L	L	L	0.7006 ± 0.0018	0.00687 ± 0.00026	-1.723 ± 0.023
True residual activity	L	L	T	0.7007 ± 0.0018	0.00687 ± 0.00026	-1.723 ± 0.021
True graph	T	L	L	0.7025 ± 0.0022	0.00682 ± 0.00028	-1.832 ± 0.002
True graph and residual	T	L	T	0.7029 ± 0.0027	0.00680 ± 0.00029	-1.832 ± 0.003
True graph and mechanisms	T	T	L	0.7042 ± 0.0019	0.00675 ± 0.00026	-1.838 ± 0.000
All three true	T	T	T	0.7043 ± 0.0019	0.00675 ± 0.00026	-1.838 ± 0.000

H Exogenous-regulator Case Study: LINC00525 in A549

H.1 Question and external perturbation data

We test whether a regulator absent from the original gene universe can be attached to trained cell-line-specific models while keeping their learned causal DAGs and child-specific mechanisms fixed. We study LINC00525, a long non-coding RNA implicated in lung adenocarcinoma proliferation and cell-cycle progression (Fang et al., 2021).

We use GSE171460, comprising three control and three LINC00525-knockdown A549 arrays on platform GPL15207 from the same study. The submitter-normalized RMA log-expression values are mapped to the A549 model genes; when multiple probes map to one gene, we retain the probe with the highest mean expression across the six arrays. This covers 3,657 of the 3,709 A549 genes. Each knockdown response is its log-expression difference from the mean of the three controls. The LINC00525 probe itself changes by approximately -0.30 on the log2 scale, corresponding to only about 19% measured depletion. We therefore do not interpret the experiment as a perfect intervention or assign this value as a known functional activity.

The contrast contains a dominant type-I-interferon signature, a documented concern for RNA-interference expression profiling (Smith et al., 2017). The primary analysis excludes 36 covered genes in the declared Hallmark interferon-alpha core before fitting. This prevents the reagent-associated program from determining the new connections; it does not remove those genes from the frozen A549 models during intervention propagation.

H.2 Integration into frozen A549 models

We independently augment each of the five official A549 models. All original parameters are frozen, and each learned DAG is replaced by its official binary support at $q_{\mathrm{eval}} = 0.09$. LINC00525 is introduced as an exogenous root with no parents or child-specific mechanism. Its centered input is $u_{\mathrm{LNC}} = 0$ for the control condition and $u_{\mathrm{LNC}} = -1$ for knockdown. The only fitted model parameters are its input weights w_{jh}^{LNC} into each frozen child mechanism and one scalar that aligns the predicted LINCS-scale response with the microarray contrast. Because the new node is a root, adding its outgoing edges cannot introduce a directed cycle into a frozen DAG.

For child j, the new neural path strength follows the same definition as the original graph,

$$s_{\mathrm{LNC}\to j} = \sum_{h} \left|w_{jh}^{\mathrm{LNC}}\right| |v_{jh}|\,, \tag{125}$$

where v_{jh} is the frozen output weight of child mechanism j. Training minimizes centered response MSE plus a sparsity coefficient times the mean new path strength. The coefficient is selected from $\{0, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}\}$ by leaving out each knockdown replicate in turn, fitting the other two, and maximizing held-out correlation. The model is then refitted to the mean of all three knockdown responses. New edges use the same $q_{\mathrm{eval}} = 0.09$ threshold as the A549 backbone. The evaluation quantities and permutation test are defined in Appendix F.5.

H.3 Held-out generalization and new connections

Across the five frozen A549 models, mean leave-one-replicate-out correlation is 0.454 ± 0.043. Across 25 gene-label permutations, the corresponding correlation ranges from -0.014 to 0.046, so the observed value exceeds every null realization ($p_{\mathrm{emp}} = 1/26 = 0.0385$).

At the declared threshold, 17 outgoing connections recur in all five augmented models and 20 recur in at least four. The 5/5 targets are *BBC3*, *CLEC4E*, *DDIT3*, *DUSP5*, *EGR1*, *ETS1*, *GDF15*, *HMOX1*, *IER3*, *IL11*, *ITGA2*, *JUN*, *KLF6*, *MAFF*, *NFE2L3*, *PLSCR1*, and *PPP1R15A*. Each permuted control shares at most one 5/5 target with the observed set. The null controls produce 16-59 consensus connections, so the evidence does not come from obtaining an unusually large target set.

The 17 consensus-predicted targets yield 12 Gene Ontology Biological Process terms at Benjamini-Hochberg $q \le 0.05$, compared with at most four across the 25 permuted target sets ($p_{\mathrm{emp}} = 0.0385$). Table 30 reports the principal stress, apoptosis, and DNA-damage annotations; transcriptional terms are retained in the complete analysis but omitted from this compact summary.

Table 30: Selected GO Biological Process enrichments among the 17 consensus-predicted LINC00525 targets.

Process	Overlap	BH q
Integrated stress response signaling	3/17	2.10×10^{-3}
Response to endoplasmic-reticulum stress	4/17	2.10×10^{-3}
Intrinsic apoptosis induced by ER stress	3/17	4.68×10^{-3}
Apoptotic process	5/17	6.19×10^{-3}
DNA-damage response	4/17	2.73×10^{-2}

H.4 Propagated programs and experimental concordance

We next summarize the complete propagated response rather than only the new direct connections. The four tested programs were fixed before inspecting these summaries. G1/S transition, cell cycle, and mitotic cell cycle all decrease significantly (Table 31); the broader cell population proliferation category decreases slightly but is not significant. The first three results are therefore the supported program-level predictions.

Table 31: Predicted program response after LINC00525 knockdown. Effect is the mean of the median predicted gene response; consistency is the number of A549 models with a negative program mean.

Program	Genes	Effect	Consistency / BH q
G1/S transition	64	-0.0382	4/5 / 6.67×10^{-5}
Cell cycle	176	-0.0222	5/5 / 6.67×10^{-5}
Mitotic cell cycle	141	-0.0232	5/5 / 6.67×10^{-5}
Cell population proliferation	33	-0.0043	5/5 / 0.152

These program-level predictions agree with orthogonal assays in the source A549 study: LINC00525 silencing reduced real-time proliferation, EdU incorporation, and clonogenicity and altered cell-cycle progression (Fang et al., 2021). A separate colorectal-cancer study likewise reported reduced proliferation, increased apoptosis, and G0/G1 arrest after LINC00525 knockdown (Zheng & Peng, 2019). This cross-context result is supportive rather than A549-specific validation.

Overall, this single-regulator, three-replicate case study supports model extensibility, held-out response generalization, and program-level biological coherence. Its fitted new connections remain model-derived regulatory hypotheses rather than independently validated causal edges.