LEARNING FAITHFUL MECHANISM SUBGRAPHS FROM PARTIALLY TRUSTED BIOLOGICAL PRIORS FOR GENETIC PERTURBATION PREDICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Predicting cellular responses to genetic perturbations is a central problem in virtual cell modeling. However, current methods face two key limitations: they typically treat biological priors as fixed ground truth, which are inherently incomplete, noisy, and only partially trustworthy; and they lack explicit mechanistic explanations for their predictions, instead relying on post hoc analyses rather than explicit, testable mediator entities. As a result, it remains unclear whether a model relies on biological mechanisms or merely exploits statistical correlations. These limitations motivate a broader machine learning question: how can we learn from partially trusted prior knowledge while ensuring that the learned explanation faithfully reflects the mechanism underlying the prediction? To address this question in genetic perturbation modeling, we propose **PerturbDigger**, a graph learning framework for perturbation prediction and mechanistic explanation. PerturbDigger formulates biological prior knowledge as a partially trusted prior graph, performs context-specific structural calibration to improve graph reliability, and learns perturbation responses with a biologically-guided graph transformer regularized by faithfulness constraints on mechanistic subgraphs. Extensive experiments show that PerturbDigger improves genetic perturbation prediction, generalizes better to unseen cellular contexts, and produces mechanistic subgraphs at both the sample and perturbation levels. Additional analyses show that the learned subgraphs are biologically meaningful, faithful, and stable, suggesting a promising step toward more interpretable and mechanism-aware virtual cell models. Our code is available at: https://anonymous.4open.science/r/PerturbDigger.

1 INTRODUCTION

Predicting cellular responses to genetic perturbations has become a central problem in virtual cell modeling (Ma et al., 2025; Roohani et al., 2025; Dimitrov et al., 2026). Recent learning methods have made substantial progress, particularly by incorporating biological priors such as gene-gene relationships, gene regulatory networks, and latent response programs (Roohani et al., 2023; Chi et al., 2025; Xu et al., 2025). However, the prior knowledge is inherently incomplete, noisy, and only partially trustworthy (Benedetti et al., 2020; Zhang et al., 2023). Moreover, in scientific applications, predictive accuracy alone is not enough (Wei et al., 2025; Roohani et al., 2025): we also want to understand whether a model captures underlying perturbation mechanisms and provides biologically meaningful explanations that help scientists discover biological knowledge and generate testable biological hypotheses. These concerns raise a broader machine learning question beyond biology: how can we learn from partially trusted prior knowledge while ensuring that the learned explanation faithfully reflects the mechanism underlying the prediction?

Existing methods do not answer this question well for two limitations: **(i) Typically treating biological priors as fixed ground truth:** regulatory edges are context-dependent (Mowery et al., 2024), and graph connectivity may contain both missing mechanisms and spurious shortcuts. Treating such knowledge as fixed ground truth can therefore limit both predictive performance and interpretability. **(ii) Lacking explicit mechanistic reasoning in the prediction process:** even when prior knowledge is incorporated into the predictor, explanations are typically obtained only post hoc, through feature attribution, latent factors, or response programs, rather than through explicit mediator entities aligned

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.

with testable biological mechanisms such as transcription factors (TFs) and pathways. Consequently, it remains unclear whether a model truly relies on biological mechanisms or merely exploits statistical correlations. There is still a lack of end-to-end models that explicitly align prediction with testable mediators and recover propagation paths from perturbation to downstream expression changes.

To address these challenges, we propose **PerturbDigger**, a graph learning framework for both prediction and mechanistic explanation, instantiated in genetic perturbation modeling. Our framework formulates biological prior knowledge as a **partially trusted prior graph** rather than a fixed ground truth. It then performs **context-specific structural calibration** using control data to improve graph reliability while respecting relation-specific biological priors (addressing Limitation (i)). On top of the calibrated graph, the model learns perturbation responses with a **biologically-guided graph transformer** regularized by **faithfulness constraints on mechanistic subgraphs**, encouraging the predictor to generate explanatory subgraphs that are truly responsible for the prediction (addressing Limitation (ii)). In this way, PerturbDigger is designed not only to predict perturbation responses, but also to recover interpretable mechanistic subgraphs that link perturbed genes to downstream expression changes through explicit mediator entities. Extensive experiments show that PerturbDigger improves perturbation prediction performance under both standard and out-of-distribution settings, while producing mechanistic subgraphs at both the sample and perturbation levels. Further analyses show that the learned subgraphs are biologically meaningful, faithful, and stable. Together, these results suggest a step toward more interpretable and mechanism-aware virtual cell models.

Our contributions can be summarized as follows:

- We formulate genetic perturbation modeling as learning from **partially trusted prior graphs**, shifting the focus from prediction alone toward mechanistic understanding with faithful explanations responsible for the prediction.
- We propose **PerturbDigger**, a graph learning framework for perturbation prediction and mechanistic explanation. PerturbDigger performs **context-specific structural calibration** to improve graph reliability while respecting relation-specific biological priors, and learns perturbation responses with a **biologically-guided graph transformer** regularized by **faithfulness constraints on mechanistic subgraphs**.
- Extensive experiments show that PerturbDigger improves genetic perturbation prediction with stronger transfer capability to unseen cellular contexts, while providing fine-grained interpretability through sample-level and perturbation-level mechanistic subgraphs. Additional analyses demonstrate that the learned explanations are biologically meaningful, faithful, and stable.

2 Related Work

2.1 Perturbation Response Prediction with Biological Priors

Predicting cellular responses to perturbations has become a central task in virtual cell modeling, enabling in silico screening, prioritization of experiments, and hypothesis generation under unobserved genetic or environmental conditions. Recent work spans multiple paradigms. Early methods such as scGen (Lotfollahi et al., 2019) and CellOT (Bunne et al., 2023) model perturbation-induced state transitions through latent-variable or transport-based formulations, while CPA (Lotfollahi et al., 2023) factorizes perturbation and covariate effects to support compositional generalization. More recent models, including Scouter (Zhu & Li, 2025), Squidiff (He et al., 2025), and State (Adduri et al., 2025), further improve prediction through gene semantic embeddings, diffusion-based modeling, or stronger generalization across diverse cellular contexts. A parallel line of work incorporates biological prior knowledge as structured inductive bias. GEARS (Roohani et al., 2023) uses a gene–gene knowledge graph to improve prediction for unseen perturbations, GRAPE (Chi et al., 2025) extends this direction with heterogeneous graph neural networks and graph structure learning, and CellOracle (Kamimoto et al., 2023) leverages gene regulatory networks for in silico TF perturbation analysis. These studies show that structured biological priors can substantially improve perturbation prediction and transferability.

However, interpretability remains a gap. Existing methods primarily use biological priors as fixed structure for prediction, while mechanistic interpretation is often obtained indirectly through post hoc statistics, latent factors, or response programs (Xu et al., 2025; Kamimoto et al., 2023; Pang et al.,

2025). Consequently, there is still a lack of end-to-end models that align prediction with testable mediators such as TFs and pathways, while recovering verifiable propagation paths from genetic perturbations to expression changes.

2.2 Faithful and Explainable Graph Learning

Graph learning offers a natural framework for mechanistic interpretation, as predictions can in principle be traced to informative nodes, edges, and subgraphs. A large body of work explains trained graph neural networks post hoc. For example, GNNExplainer (Ying et al., 2019) identifies compact explanatory subgraphs and feature subsets, GraphMask (Schlichtkrull et al., 2021) learns differentiable edge masks to determine which messages can be removed without changing predictions, and SubgraphX (Yuan et al., 2021) explains predictions by searching for important subgraphs. These methods establish the basic toolkit for subgraph-level explanation, but they do not guarantee that the recovered structures correspond to semantically valid mechanisms in scientific domains. More recent work has moved toward self-explainable or intrinsically interpretable graph learning. In particular, SUNNY-GNN (Deng & Shen, 2024) incorporates sufficient and necessary explanations directly into graph model training, while GraphFramEx (Amara et al., 2024) shows that explanation quality depends critically on sufficiency, necessity, and robustness to the distribution shift induced by subgraph masking. These studies suggest that explanation should not be treated as an afterthought, but tied to the predictor through explicit faithfulness criteria.

Our work connects these two lines of research. Different from prior perturbation models that use biological graphs mainly as fixed priors for prediction, PerturbDigger treats biological prior knowledge as a partially trusted graph prior whose reliability should be calibrated under the target cellular context. Different from general-purpose graph explanation methods, PerturbDigger is designed for mechanistically grounded perturbation modeling, where explanatory subgraphs are tied to explicit biological mediator entities and constrained to be faithful to the prediction process.

3 Methods

PerturbDigger is a graph learning framework for both genetic perturbation prediction and mechanistic explanation (Figure 1). Given a perturbation dataset, it first calibrates a partially trusted prior graph using control data to obtain context-specific edge reliabilities (Section 3.2). Then it predicts perturbation responses through ordered propagation from perturbed genes to transcription factor (TF) and pathway mediators, with a faithfulness constraint that ties the recovered mechanistic subgraphs to the prediction process (Section 3.3).

3.1 Problem Definition

The goal of genetic perturbation prediction is to predict single-cell transcriptional changes under a given perturbation condition. We consider a perturbation dataset $\mathcal{D} = \mathcal{D}_{\text{ctrl}} \cup \mathcal{D}_{\text{pert}}$ over N_g genes. The control subset $\mathcal{D}_{\text{ctrl}}$ contains unperturbed or non-targeting-control cells and is used to estimate the context-specific baseline and calibrate prior reliability. For each perturbed sample $i \in \mathcal{D}_{\text{pert}}$, $\boldsymbol{x}_i \in \mathbb{R}^{N_g}$ denotes a reference expression profile representing the cellular context, $\boldsymbol{z}_i \in \{0,1\}^{N_g}$ is the perturbation indicator over target genes, and $\boldsymbol{y}_i \in \mathbb{R}^{N_g}$ is the observed post-perturbation expression profile. When paired pre-perturbation cells are unavailable, $\boldsymbol{x}_i$ can be instantiated as a non-targeting control representation, which is uniformly sampled from the same dataset. We use $\boldsymbol{\Delta}_i = \boldsymbol{y}_i - \boldsymbol{x}_i$ as the primary prediction target.

Let $\mathcal{G}_{\mathcal{D}}$ be the partially trusted prior graph and $\widetilde{\mathcal{G}}_{\mathcal{D}} = (\mathcal{V}, \mathcal{E}, \mathcal{H}, \widetilde{\mathcal{A}})$ its context-calibrated version, where $\widetilde{\mathcal{A}} = \{\widetilde{a}_e\}_{e \in \mathcal{E}}$ are learnable edge reliabilities. The overall learning target is summarized as

$$\min_{\widetilde{\mathcal{A}}} \ \mathcal{L}_{\text{graph}}(\widetilde{\mathcal{A}}; \mathcal{D}_{\text{ctrl}}) \text{ and } \min_{\theta} \ \mathcal{L}_{\text{pert}}(\theta, \widetilde{\mathcal{A}}; \mathcal{D}_{\text{pert}}), \quad \hat{\boldsymbol{\Delta}}_i = f_\theta(\boldsymbol{x}_i, \boldsymbol{z}_i, \widetilde{\mathcal{G}}_{\mathcal{D}}), \quad \hat{\boldsymbol{y}}_i = \boldsymbol{x}_i + \hat{\boldsymbol{\Delta}}_i. \tag{1}$$

Beyond accurate prediction, we seek a mechanistic subgraph $\mathcal{S}_i \subseteq \widetilde{\mathcal{G}}_{\mathcal{D}}$ for each sample, such that the recovered subgraph links perturbed genes to downstream expression changes through explicit and testable mediator entities, particularly TFs and pathways. This subgraph is expected to be both biologically meaningful and faithful to the model prediction.

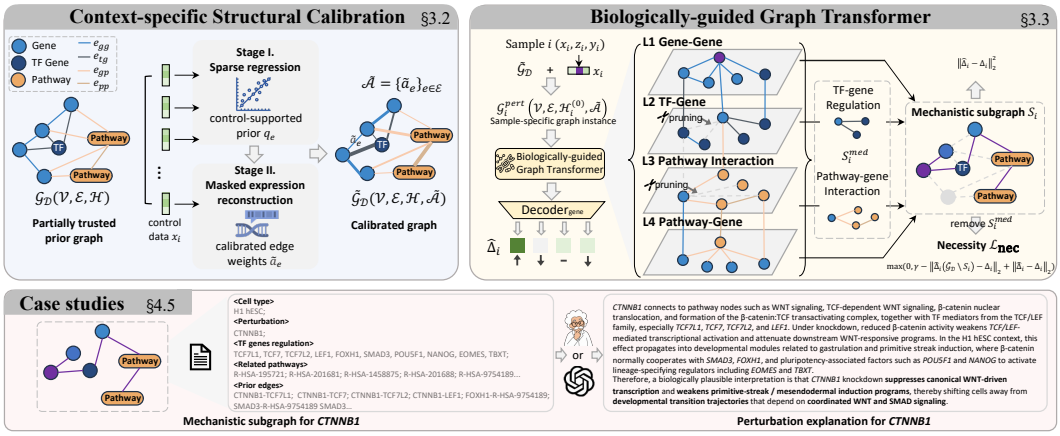


Figure 1: Framework of PerturbDigger. It consists of a Context-Specific Structural Calibration module and a Biologically-guided Graph Transformer regularized by faithfulness constraints on mechanistic subgraphs. The case study shows the mechanistic explanation for *CTNNB1* perturbation.

3.2 Context-Specific Structural Calibration on Partially Trusted Prior Graph

We begin by constructing a prior graph from biological knowledge and calibrating it with control data, so that downstream perturbation learning is performed on a context-reliable graph rather than on fixed prior structure.

Partially trusted prior graph from biological priors. For dataset $\mathcal{D}$, we construct a heterogeneous prior graph $\mathcal{G}_{\mathcal{D}} = (\mathcal{V}, \mathcal{E}, \mathcal{H})$, where $\mathcal{V} = \mathcal{V}_g \cup \mathcal{V}_p$ contains gene nodes $\mathcal{V}_g$ and pathway nodes $\mathcal{V}_p$. TFs are not modeled as a separate node type, but as gene nodes with outgoing regulatory edges.

The node attribute set $\mathcal{H}$ contains static features $\boldsymbol{h}_g = [\mathrm{ID}(g) \,\|\, \mathrm{M}(g)]$ for each gene $g \in \mathcal{V}_g$ and $\boldsymbol{h}_p = \mathrm{ID}(p)$ for each pathway $p \in \mathcal{V}_p$, where $\mathrm{ID}(\cdot)$ denotes a learnable identity embedding and $\mathrm{M}(g)$ denotes gene-level biological metadata.

The edge set $\mathcal{E}$ contains four relation types: gene–gene support edges $\mathcal{E}_{gg}$, TF–gene regulatory edges $\mathcal{E}_{tg}$, gene–pathway edges $\mathcal{E}_{gp}$, and pathway–pathway edges $\mathcal{E}_{pp}$. They are initialized from metadata-based gene similarity, regulatory priors from reference-genome information, and curated Reactome pathway membership/hierarchy information, detailed in Appendix A. Since regulatory interactions are context-dependent and support edges may introduce spurious shortcuts, we treat $\mathcal{G}_{\mathcal{D}}$ as a **partially trusted** graph to be calibrated rather than fixed ground truth.

Context-specific structural calibration. We calibrate edge reliabilities using the control subset $\mathcal{D}_{\mathrm{ctrl}} \subseteq \mathcal{D}$, which includes control cells from the perturbation dataset and, when available, additional external unperturbed scRNA-seq data. We learn edge weights $\widetilde{\mathcal{A}} = \{\widetilde{a}_e\}_{e\in\mathcal{E}}$ with $\widetilde{a}_e \in [0, 1]$, where $\widetilde{a}_e$ measures the reliability of prior edge e under the dataset context. Since heterogeneous sources do not provide comparable confidence scores, all edges are initialized with unit weight.

For each control sample i, we instantiate the shared prior graph by attaching the reference expression to gene nodes while keeping pathway nodes static: $\boldsymbol{h}_{i,g}^{(0)} = [x_{i,g} \,\|\, \boldsymbol{h}_g]$ for $g \in \mathcal{V}_g$ and $\boldsymbol{h}_{i,p}^{(0)} = \boldsymbol{h}_p$ for $p \in \mathcal{V}_p$. This yields a sample-specific graph $\mathcal{G}_i = (\mathcal{V}, \mathcal{E}, \mathcal{H}_i^0)$ with $\mathcal{H}_i^0 = \{\boldsymbol{h}_{i,v}^{(0)}\}_{v\in\mathcal{V}}$.

We calibrate $\widetilde{\mathcal{A}}$ in two stages. Stage I derives a control-supported calibration prior for TF–gene edges from sparse regression within the prior neighborhood. Specifically, it uses prior-constrained Bagging Ridge regression following the design of CellOracle (Kamimoto et al., 2023). For each target gene g, let $\mathcal{N}_{tg}(g)$ denote the TFs connected to g by prior TF–gene edges. Across control samples, we fit a sparse local regulatory model and convert the fitted coefficients into edge-level relevance priors:

$$x_{i,g} \approx b_g + \sum_{t\in\mathcal{N}_{tg}(g)} \beta_{t\to g} x_{i,t}, \qquad q_{(t,g)} = \frac{|\beta_{t\to g}|}{|\beta_{t\to g}| + c_\beta + \epsilon}, \quad (t,g) \in \mathcal{E}_{tg}, \tag{2}$$

where $\beta_{t\to g}$ denotes the mean coefficient across the base estimators containing t, b_g denotes the mean fitted intercept, c_β is the median of all fitted TF–gene edges, and ϵ is a small numerical constant.

Stage II learns $\widetilde{\mathcal{A}}$ by masked expression reconstruction, with a lightweight relation-aware regularizer anchoring each relation type to its intended role:

$$\mathcal{L}_{\text{graph}} = \underbrace{\sum_{i\in\mathcal{D}_{\text{ctrl}}}\sum_{g\in\mathcal{M}_i}(\hat{x}_{i,g}-x_{i,g})^2}_{\mathcal{L}_{\text{rec}}} + \underbrace{\frac{1}{|\mathcal{E}_{tg}|}\sum_{e\in\mathcal{E}_{tg}}(\widetilde{a}_e-q_e)^2}_{\text{TF–gene alignment}} + \underbrace{\frac{1}{|\mathcal{E}_{gp}|+|\mathcal{E}_{pp}|}\sum_{e\in\mathcal{E}_{gp}\cup\mathcal{E}_{pp}}(1-\widetilde{a}_e)^2}_{\text{pathway-edge retention}} + \underbrace{\frac{1}{|\mathcal{E}_{gg}|}\sum_{e\in\mathcal{E}_{gg}}|\widetilde{a}_e|}_{\text{support-edge sparsity}}. \tag{3}$$

Here $\mathcal{M}_i \subseteq \mathcal{V}_g$ is the masked gene set for i, and $\hat{x}_{i,g}$ is reconstructed with the same propagation operators used later for perturbation prediction. The first term is sample-level reconstruction on control graphs. The remaining terms act directly on the shared dataset-level reliabilities, which align TF–gene edges with the control-supported prior, retain pathway-related mechanistic edges unless contradicted by data, and suppress shortcut propagation on gene–gene support edges, respectively.

The final calibrated graph $\widetilde{\mathcal{G}}_{\mathcal{D}} = (\mathcal{V}, \mathcal{E}, \mathcal{H}, \widetilde{\mathcal{A}})$ is shared across samples and subsequently used for perturbation prediction and mechanistic explanation. Thus, context enters through sample-specific node features, while graph structure is calibrated at dataset level using the shared control distribution.

3.3 Biologically-guided Graph Transformer with Faithfulness Constraints on Mechanistic Subgraphs

Given the calibrated graph $\widetilde{\mathcal{G}}_{\mathcal{D}}$, PerturbDigger predicts perturbation responses by enforcing a biologically ordered propagation path. Perturbation signals are propagated through TF and pathway mediators, which form the basis of the mechanistic explanation.

Sample-specific graph instantiation. On top of the calibrated graph $\widetilde{\mathcal{G}}_{\mathcal{D}} = (\mathcal{V}, \mathcal{E}, \mathcal{H}, \widetilde{\mathcal{A}})$, we model perturbation response as a structured propagation process from perturbed genes to downstream genes through regulatory and pathway mediators.

For each sample $i \in \mathcal{D}_{\text{pert}}$, we build $\mathcal{G}_i^{\text{pert}} = (\mathcal{V}, \mathcal{E}, \mathcal{H}_i^{(0)}, \widetilde{\mathcal{A}})$. Gene node initialization is $\boldsymbol{h}_{i,g}^{(0)} = [\boldsymbol{h}_g \,\|\, \text{Enc}(\boldsymbol{x}_i)]$ if $z_{i,g} = 1$ and $\boldsymbol{h}_{i,g}^{(0)} = \mathbf{0}$ otherwise; pathway nodes use $\boldsymbol{h}_{i,p}^{(0)} = \boldsymbol{h}_p$. The $\text{Enc}(\cdot)$ denotes a lightweight feedforward network that captures the cellular context. We avoid injecting raw all-gene control expression directly into the perturbation graph, thereby reducing feature leakage and encouraging the predicted response to be generated through propagation from perturbed seeds.

Biologically-guided graph transformer. Effective perturbation modeling requires multi-hop information exchange among genes, TF regulators, and pathway mediators. However, applying a general multi-layer graph transformer over the full graph can lead to unconstrained message passing, where signals from different biological relations are repeatedly mixed, spurious shortcuts may be amplified, and node representations may become over-smoothed.

To avoid such unconstrained propagation while preserving mechanistic interpretability, PerturbDigger contains a biologically-guided graph transformer with four ordered propagation steps: L1 gene–gene support diffusion, L2 TF–gene regulation, L3 gene-to-pathway aggregation plus pathway-hierarchy propagation, and L4 pathway-to-gene readout. Specifically:

L1 diffuses perturbation signals over gene–gene support edges using calibrated weighted aggregation:

$$\boldsymbol{h}_{i,g}^{(1)} = \texttt{MLP}_{gg}(\boldsymbol{h}_{i,g}^{(0)}) + \sum_{g':(g',g)\in\mathcal{E}_{gg}} \widetilde{a}_{(g',g)}\, \boldsymbol{W}_{gg}\, \boldsymbol{h}_{i,g'}^{(0)}. \tag{4}$$

L2 explicitly propagates through TF-target regulation. We treat genes as queries and TFs as keys, reflecting the biological intuition that TFs regulate gene expression. Different from general Transformer modules, we use separate value projections for TF and gene nodes so that type-specific information can be preserved during propagation. Moreover, to mitigate noise from under-trained nodes and suppress uniform attention distributions, we employ a biological pruning strategy. Let $\texttt{Prune}_{K,\tau}$ denote thresholding at τ followed by top-K retention over incoming edges of each receiver and d

denote the hidden dimension of query/key projections, we compute

$$\begin{aligned}\alpha_{tg} &= \texttt{Prune}_{K_{tg},\tau_{tg}}\left(\sigma\left(\widetilde{a}_{(t,g)}\frac{(\boldsymbol{h}_{i,g}^{(1)}\boldsymbol{W}_Q^{tg})(\boldsymbol{h}_{i,t}^{(1)}\boldsymbol{W}_K^{tg})^\top}{\sqrt{d}}\right)\right),\\ \boldsymbol{h}_{i,g}^{(2)} &= \boldsymbol{h}_{i,g}^{(1)}\boldsymbol{W}_{V_g}^{tg} + \sum_{t:(t,g)\in\mathcal{E}_{tg}} \alpha_{tg}\,\boldsymbol{h}_{i,t}^{(1)}\boldsymbol{W}_{V_t}^{tg}.\end{aligned} \tag{5}$$

The active edges selected by α_{tg} define a sparse TF-gene regulation layer for sample i.

L3 reuses the same biological pruning strategy to aggregate selected gene responses into pathways and then propagate across the pathway hierarchy:

$$\begin{aligned}\alpha_{gp} &= \texttt{Prune}_{K_{gp},\tau_{gp}}\left(\sigma\left(\widetilde{a}_{(g,p)}\frac{(\boldsymbol{h}_p\boldsymbol{W}_Q^{gp})(\boldsymbol{h}_{i,g}^{(2)}\boldsymbol{W}_K^{gp})^\top}{\sqrt{d}}\right)\right),\\ \bar{\boldsymbol{h}}_{i,p}^{(3)} &= \texttt{MLP}_{gp}(\boldsymbol{h}_p) + \sum_{g:(g,p)\in\mathcal{E}_{gp}} \alpha_{gp}\,\boldsymbol{h}_{i,g}^{(2)}\boldsymbol{W}_V^{gp},\\ \boldsymbol{h}_{i,p}^{(3)} &= \texttt{MLP}_{pp}(\bar{\boldsymbol{h}}_{i,p}^{(3)}) + \sum_{p':(p',p)\in\mathcal{E}_{pp}} \widetilde{a}_{(p',p)}\,\boldsymbol{W}_{pp}\,\bar{\boldsymbol{h}}_{i,p'}^{(3)}.\end{aligned} \tag{6}$$

Compared with L2, the only conceptual change is that L3 selects pathway members instead of regulators, and then refines pathway states through the curated hierarchy in $\mathcal{E}_{pp}$. Edges with α_{gp} define a sparse pathway-gene interaction layer for sample i.

L4 projects refined pathway states back to genes and decodes perturbation responses:

$$\widetilde{\boldsymbol{h}}_{i,g} = \texttt{MLP}_{pg}(\boldsymbol{h}_{i,g}^{(2)}) + \sum_{p:(g,p)\in\mathcal{E}_{gp}} \widetilde{a}_{(g,p)}\,\boldsymbol{W}_{pg}\,\boldsymbol{h}_{i,p}^{(3)}, \qquad \hat{\Delta}_{i,g} = \texttt{Decoder}_{\text{gene}}(\widetilde{\boldsymbol{h}}_{i,g}),\ \ g\in\mathcal{V}_g, \tag{7}$$

where $\texttt{Decoder}_{\text{gene}}$ is a shared gene-wise feedforward decoder.

We define the sample-specific mechanistic subgraph $\mathcal{S}_i$ as the active propagation subgraph for sample i, where the sparse selections in L2 and L3 determine the mediator-specific core of $\mathcal{S}_i$ (denoted by $\mathcal{S}_i^{\text{med}}$), while L1 and L4 provide supporting diffusion and pathway-to-gene readout. Perturbation-level explanations are obtained by aggregating sample-specific subgraphs across the same perturbation.

Faithfulness constraints on mechanistic subgraphs. The model prediction is generated through forward propagation on the selected subgraph with mediator for regulatory and pathway interactions. Thus, an intrinsic form of sufficiency is already built into the model pipeline. However, the selected subgraph should be faithfully responsible for the prediction rather than merely co-occurring with it. We therefore impose a necessity constraint expecting the prediction quality to deteriorate once the selected subgraph is removed.

For each perturbation sample i, let $\hat{\Delta}_i$ denote the original prediction and let $\hat{\Delta}_i(\widetilde{\mathcal{G}}_{\mathcal{D}}\setminus\mathcal{S}_i^{\text{med}})$ denote the counterfactual prediction obtained by running the same model after masking out the selected mediator edges in $\mathcal{S}_i^{\text{med}}$ from the graph $\widetilde{\mathcal{G}}_{\mathcal{D}}$, while keeping all node features, calibrated edge reliabilities, and model parameters fixed. We define the necessity loss as

$$\mathcal{L}_{\text{nec}} = \max\Big(0,\ \gamma - \big\|\hat{\boldsymbol{\Delta}}_i(\widetilde{\mathcal{G}}_{\mathcal{D}}\setminus\mathcal{S}_i^{\text{med}}) - \boldsymbol{\Delta}_i\big\|_2 + \big\|\hat{\boldsymbol{\Delta}}_i - \boldsymbol{\Delta}_i\big\|_2\Big), \tag{8}$$

where $\gamma > 0$ is a margin, and optimize

$$\mathcal{L}_{\text{pert}} = \|\hat{\boldsymbol{\Delta}}_i - \boldsymbol{\Delta}_i\|_2^2 + \lambda_{\text{nec}}\mathcal{L}_{\text{nec}}. \tag{9}$$

Together, these constraints encourage PerturbDigger to produce explanation subgraphs that are compact, mechanistically meaningful, and faithful for the predicted perturbation response.

4 Experiments

We evaluate PerturbDigger on both predictive performance and mechanistic interpretability. We first describe the datasets and baselines in Section 4.1. Then we conduct extensive experiments to address

Table 1: Comparison across different models evaluated by RMSE(↓), Direction Accuracy Coefficient (DAC)(↑), and Pearson Correlation Coefficient (PCC)(↑). Boldface indicates the best performance.

Dataset	Adamson			Norman			Replogle		
Metric	RMSE	DAC	PCC	RMSE	DAC	PCC	RMSE	DAC	PCC
SAMS-VAE	$0.651_{\pm 0.011}$	$0.424_{\pm 0.022}$	$0.144_{\pm 0.020}$	$0.503_{\pm 0.017}$	$0.374_{\pm 0.021}$	$0.125_{\pm 0.012}$	$0.672_{\pm 0.015}$	$0.439_{\pm 0.020}$	$0.189_{\pm 0.030}$
GEARS	$0.081_{\pm 0.001}$	$0.656_{\pm 0.002}$	$0.551_{\pm 0.006}$	$0.070_{\pm 0.001}$	$0.520_{\pm 0.005}$	$0.428_{\pm 0.004}$	$0.092_{\pm 0.002}$	$0.627_{\pm 0.004}$	$0.613_{\pm 0.010}$
GRAPE	$0.075_{\pm 0.001}$	$0.695_{\pm 0.003}$	$0.746_{\pm 0.007}$	$0.048_{\pm 0.001}$	$0.758_{\pm 0.001}$	$0.474_{\pm 0.005}$	$0.083_{\pm 0.002}$	$0.662_{\pm 0.004}$	$0.728_{\pm 0.008}$
Squidiff	$0.080_{\pm 0.002}$	$0.660_{\pm 0.004}$	$0.698_{\pm 0.008}$	$0.049_{\pm 0.001}$	$0.722_{\pm 0.003}$	$0.477_{\pm 0.004}$	$0.082_{\pm 0.001}$	$0.689_{\pm 0.005}$	$0.708_{\pm 0.007}$
STATE	$0.079_{\pm 0.001}$	$0.672_{\pm 0.004}$	$0.748_{\pm 0.008}$	$0.051_{\pm 0.001}$	$0.703_{\pm 0.003}$	$0.452_{\pm 0.002}$	$0.081_{\pm 0.001}$	$0.701_{\pm 0.004}$	$0.728_{\pm 0.004}$
MLP	$0.232_{\pm 0.010}$	$0.514_{\pm 0.015}$	$0.467_{\pm 0.013}$	$0.198_{\pm 0.010}$	$0.418_{\pm 0.016}$	$0.395_{\pm 0.010}$	$0.250_{\pm 0.011}$	$0.562_{\pm 0.014}$	$0.436_{\pm 0.016}$
GCNII	$0.082_{\pm 0.003}$	$0.667_{\pm 0.004}$	$0.602_{\pm 0.008}$	$0.067_{\pm 0.001}$	$0.602_{\pm 0.006}$	$0.442_{\pm 0.008}$	$0.090_{\pm 0.004}$	$0.668_{\pm 0.003}$	$0.671_{\pm 0.009}$
GraphSAGE	$0.085_{\pm 0.003}$	$0.668_{\pm 0.006}$	$0.709_{\pm 0.008}$	$0.066_{\pm 0.001}$	$0.634_{\pm 0.003}$	$0.467_{\pm 0.003}$	$0.088_{\pm 0.002}$	$0.673_{\pm 0.002}$	$0.709_{\pm 0.009}$
GT	$0.080_{\pm 0.001}$	$0.684_{\pm 0.005}$	$0.750_{\pm 0.008}$	$0.050_{\pm 0.001}$	$0.719_{\pm 0.005}$	$0.484_{\pm 0.005}$	$0.081_{\pm 0.002}$	$0.721_{\pm 0.003}$	$0.721_{\pm 0.010}$
PerturbDigger	$\mathbf{0.067}_{\pm 0.001}$	$\mathbf{0.786}_{\pm 0.004}$	$\mathbf{0.801}_{\pm 0.008}$	$\mathbf{0.036}_{\pm 0.001}$	$\mathbf{0.836}_{\pm 0.005}$	$\mathbf{0.565}_{\pm 0.006}$	$\mathbf{0.072}_{\pm 0.001}$	$\mathbf{0.729}_{\pm 0.003}$	$\mathbf{0.745}_{\pm 0.003}$

the following research questions: **RQ1**: How well does PerturbDigger perform on benchmark datasets for genetic perturbation prediction? **RQ2**: How well does PerturbDigger generalize to unseen cellular contexts under out-of-distribution settings? **RQ3**: What is the contribution of each core component of our model to overall performance? **RQ4**: How does PerturbDigger provide faithful, stable, and biologically meaningful mechanistic explanations for perturbation responses?

4.1 Experimental Setup

Datasets. To ensure the reliability and comparability of our evaluation, we conduct experiments on four widely-used benchmark datasets for genetic perturbation: Adamson (Adamson et al., 2016), Norman (Norman et al., 2019), Replogle (K562 Cells), and VCC (H1 hESCs) (Roohani et al., 2025). Details of the datasets and metrics are in Appendix B.

Baselines. We compare PerturbDigger against two categories of baselines: (i) Previous advanced perturbation models, including SAMS-VAE (Bereket & Karaletsos, 2024), STATE (Adduri et al., 2025), GEARS (Roohani et al., 2023), GRAPE (Chi et al., 2025), and Squidiff (He et al., 2025). (ii) Methods based on our constructed graphs (using our graph learning framework but not our biologically-guided graph transformer), including a simple MLP, as well as advanced Graph Neural Networks (GNNs) such as GCNII (Chen et al., 2020), GraphSAGE (Hamilton et al., 2017), and Graph Transformer (GT) (Dwivedi & Bresson, 2020).

Other settings. We report experimental results using hyperparameter with a broad range detailed in Appendix B.3, selecting those that achieve the highest validation performance. Each experiment is repeated with 10 different random seeds, reporting the mean and standard deviation.

4.2 Perturbation Prediction on Unseen Perturbations (RQ1)

We first evaluate PerturbDigger under the standard benchmark setting. Under this protocol, test perturbation conditions are not observed during training, and the model must generalize to held-out perturbation conditions from the same dataset.

Results are reported in Table 1. PerturbDigger consistently achieves the best performance across datasets and evaluation metrics, demonstrating the effectiveness of combining partially trusted biological priors, context-specific structural calibration, and biologically-guided graph propagation. Compared with non-graph baselines, PerturbDigger benefits from structured prior knowledge that constrains perturbation effects to propagate along biologically plausible routes. Compared with GNN baselines built on the same graph corpus, it further improves performance by calibrating edge reliability in a context-specific manner and by enforcing biologically-guided mechanistic propagation. In addition to these standard metrics, Appendix B.4 provides a comprehensive DEG-focused evaluation, including DAC and PCC restricted to the most strongly responding genes, as well as DES and PDS for evaluating DEG-set recovery and perturbation-specific discrimination.

4.3 OOD Generalization to Unseen Cellular Contexts (RQ2)

A central goal of virtual cell modeling is to generalize perturbation responses across biological contexts. Following the setting of Virtual Cell Challenge which claimed that true zero-shot general-

Table 2: Ablation study of masking different components in PerturbDigger. The metric is RMSE(↓).

Dataset	Adamson	Norman	Replogle	Avg.Δ
PerturbDigger	0.0673	0.0364	0.0719	-
w/o $\mathcal{E}_{tg}$	0.1128	0.0641	0.1291	↑ 0.0435
w/o $\mathcal{E}_{gp}$	0.1073	0.0489	0.1302	↑ 0.0369
w/o Calibration	0.0764	0.0403	0.0834	↑ 0.0082
w/o BP	0.0801	0.0462	0.0872	↑ 0.0126
w/o $\mathcal{L}_{\text{nec}}$	0.0879	0.0445	0.0887	↑ 0.0152

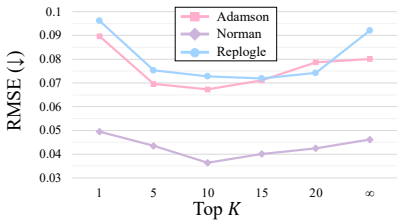


Figure 3: Hyperparameter analysis of pruning budget K in BP.

ization to a new cell state is still considered premature, we evaluate out-of-distribution generalization to unseen cellular contexts using the few-shot adaptation protocol: models are first trained on the source context (K562 cells), and then adapted using a small subset of perturbations from the target context(VCC dataset, H1 hESCs). The adapted models are evaluated on the remaining held-out target-context perturbations, which are never used during training or adaptation. Notably, for the target context, we recalibrate the edge reliabilities using its own non-targeting control cells. The source-calibrated graph is therefore not directly reused as the target graph. This design preserves target-context specificity while maintaining a common prior topology across domains.

Results are reported in Fig 2. PerturbDigger shows stronger transfer performance than competing methods, indicating that the learned model captures perturbation mechanisms that generalize more robustly across cellular contexts. We attribute this advantage to two factors. First, context-specific structural calibration reduces the mismatch between the prior graph and the target context. Second, the biologically-guided propagation process encourages the model to route perturbation effects through explicit mediator entities rather than relying on context-specific shortcuts. Together, these properties improve adaptation and prediction in unseen cellular contexts.

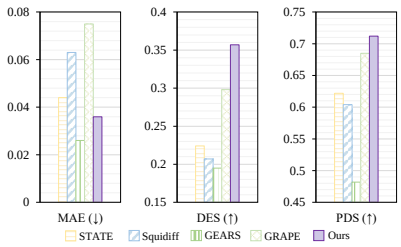


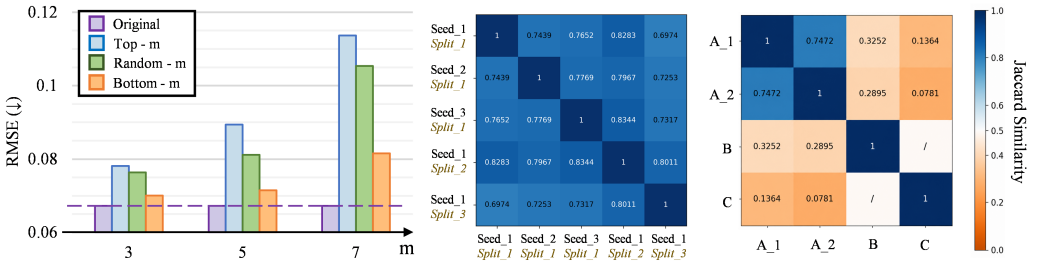
Figure 2: Results for OOD generalization on VCC dataset, evaluated by MAE, Differential Expression Score (DES), and Perturbation Discrimination Score (PDS).

4.4 Ablation Study (RQ3)

To disentangle the contributions of PerturbDigger's core components, we conduct an ablation study across both the graph-structural and model-architectural levels. Table 2 shows that removing any individual module degrades performance, confirming that each design plays a meaningful role.

Graph-level. We selectively mask (i) TF-gene edges, (ii) gene-pathway edges, and (iii) the context-specific structural calibration. In particular, removing TF-gene or gene-pathway edges leads to the largest drops. This confirms that explicit biological mediators provide critical inductive bias, allowing the model to reason through structured regulatory hierarchies rather than black-box correlations. Removing the context-specific structural calibration also harms accuracy, suggesting that biological priors are indeed partially trusted and require data-driven refinement to reduce noise and capture context-specific regulatory strengths.

Model-level. We further investigate the impact of (i) biological pruning (BP) in the biologically-guided graph transformer (replaces sparse mechanistic propagation with dense propagation) and (ii) the faithfulness constraints ($\mathcal{L}_{\text{nec}}$) during training. Removing BP also harms performance, indicating that our attention sparsification pruning is able to preserve compact mediator-centered routes and align model with biological priors. We further vary the pruning budget K_{tg} and K_{gp} in BP to examine the trade-off between predictive performance and explanation sparsity. To simplify the search space, we let $K = K_{tg} = K_{gp}$. As shown in Fig 3, performance first improves and then declines as K increases, suggesting that overly small K fails to capture sufficiently informative mechanistic subgraphs, whereas overly large K introduces redundant or noisy signals. Across most datasets, $K = 10$ performs best, consistent with the biological prior that gene regulation is typically sparse and governed by a small set of core TFs. Finally, removing faithfulness constraints causes a consistent degradation, indicating that explanation-aware training improves not only interpretability but also predictive quality. In addition, in Appendix B.5, we present fine-grained ablation experiments, including ablations disentangling the two calibration stages, ablations disentangling the regularizers in calibration Stage II (Eq. 3), and ablations of the mediator propagation order.



(a). Faithfulness Analysis **(b). Stability Analysis** **(c). Perturbation-level Comparison**

Figure 4: Biological interpretation and discovery analysis of PerturbDigger. (a) Faithfulness analysis via counterfactual edge deletion (on Adamson dataset). (b) Stability analysis of perturbation-level mechanistic subgraphs for *CTNNB1* across different training splits and random seeds (VCC dataset). (c) Perturbation-level mechanism comparison among *CTNNB1* (A), *EP300* (B), and *PCCB* (C).

4.5 Biological Interpretation and Discovery (RQ4)

Beyond predictive performance, the central goal of PerturbDigger is to provide faithful and biologically meaningful mechanistic explanations for perturbation responses. We evaluate this from four perspectives: faithfulness, stability, perturbation-level discovery, and sample-level case studies.

Faithfulness analysis. As the final prediction is produced by message passing on the selected sparse mechanistic subgraph, PerturbDigger provides intrinsic sufficiency. We therefore focus on necessity, i.e., whether the highest-ranked edges in the selected explanation are critical to the prediction. For each sample, we rank edges in the learned mechanistic subgraph by their gating scores and perform counterfactual deletion without re-running edge selection. Specifically, we remove the top-m, bottom-m, or random-m edges from the original explanation mask and recompute the prediction. As shown in Fig 4(a), removing top-ranked edges causes a much larger increase in prediction error than bottom or random edges, indicating that the learned ranking is aligned with the prediction process.

Stability analysis. A useful mechanistic explanation should be robust to training noise and moderate variation in the observed data. We therefore assess whether PerturbDigger recovers similar explanation subgraphs for the same perturbation across models trained with different random seeds and training splits. Quantitative results in Fig 4(b) shows consistently high Jaccard similarity across runs, supporting the stability of the learned mechanisms rather than artifacts of stochastic optimization.

Perturbation-level discovery. We aggregate sample-level explanations into perturbation-level mechanistic subgraphs to evaluate intra-perturbation reproducibility and inter-perturbation specificity. As shown in Fig 4(c), we split *CTNNB1* samples into disjoint subsets *CTNNB1-1* and *CTNNB1-2*, aggregate explanations independently and compare the resulting subgraphs. Their high similarity (average Jaccard similarity 0.74) indicates that PerturbDigger recovers a stable perturbation-level mechanism. For biological specificity, we compare *CTNNB1* with the biologically related perturbation *EP300* and functionally distinct perturbation *PCCB*. The *CTNNB1* subgraph is more similar to *EP300* (0.32) than to *PCCB* (0.13), consistent with prior knowledge that *CTNNB1* and *EP300* participate in a shared Wnt/β-catenin transcriptional regulatory axis, whereas *PCCB* functions in mitochondrial propionyl-CoA metabolism. Together, these results suggest that PerturbDigger captures mechanisms stable within a perturbation while sensitive to meaningful biological differences across perturbations.

Sample-level case studies. To examine how PerturbDigger explains individual cellular responses, we inspect sample-specific mechanistic subgraphs recovered for representative perturbations. As shown in Fig 1, 5, 6 and discussion in Appendix C, the subgraphs exhibit biologically plausible sparsity and highlight mediator entities. In particular, the *CTNNB1* perturbation is centered on canonical WNT/β-catenin signaling and extends to developmental modules related to primitive-streak induction and germ-layer formation, consistent with prior studies (Mukherjee et al., 2020; Blassberg et al., 2022), suggesting that PerturbDigger provides interpretable mechanistic hypotheses at sample level.

5 Conclusion

We propose PerturbDigger, a framework that bridges the gap between black-box perturbation prediction and mechanistic explanation through structural calibration and biologically-guided graph transformer with faithfulness-constrained reasoning. PerturbDigger achieves state-of-the-art predictive accuracy and recovers biologically meaningful, faithful, and stable mechanistic subgraphs, suggesting a promising step toward more interpretable and mechanism-aware virtual cell models.

AI USE STATEMENT

We used generative AI tools for language polishing and manuscript review, including feedback on methodological descriptions. As part of the study, LLMs were also used to interpret the mechanistic subgraphs generated by PerturbDigger. The outputs were reviewed by human biological experts with the learned mechanistic subgraphs, relevant biological knowledge, and supporting literature before being incorporated into the case studies in Appendix C. The prompts used for LLM-based interpretation are provided in the anonymized code repository. The authors retained full control over all research decisions, interpretation of results, and the final manuscript.

REFERENCES

Britt Adamson, Thomas Norman, Marco Jost, Min Cho, James Nuñez, Yuwen Chen, Jacqueline Villalta, Luke Gilbert, Max Horlbeck, Marco Hein, Ryan Pak, Andrew Gray, Carol Gross, Atray Dixit, Oren Parnas, Aviv Regev, and Jonathan Weissman. A multiplexed single-cell crispr screening platform enables systematic dissection of the unfolded protein response. *Cell*, 2016. URL https://www.cell.com/cell/fulltext/S0092-8674(16)31660-9.

Abhinav K. Adduri, Dhruv Gautam, Beatrice Bevilacqua, Alishba Imran, Rohan Shah, Mohsen Naghipourfar, Noam Teyssier, Rajesh Ilango, Sanjay Nagaraj, Mingze Dong, Chiara Ricci-Tam, Christopher Carpenter, Vishvak Subramanyam, Aidan Winters, Sravya Tirukkovular, Jeremy Sullivan, Brian S. Plosky, Basak Eraslan, Nicholas D. Youngblut, Jure Leskovec, Luke A. Gilbert, Silvana Konermann, Patrick D. Hsu, Alexander Dobin, Dave P. Burke, Hani Goodarzi, and Yusuf H. Roohani. Predicting cellular responses to perturbation across diverse contexts with state. *bioRxiv*, 2025. URL https://www.biorxiv.org/content/early/2025/06/27/2025.06.26.661135.

Kenza Amara, Rex Ying, Zitao Zhang, Zhihao Han, Yinan Shan, Ulrik Brandes, Sebastian Schemm, and Ce Zhang. Graphframex: Towards systematic evaluation of explainability methods for graph neural networks, 2024. URL https://arxiv.org/abs/2206.09677.

Elisa Benedetti, Maja Pučić-Baković, Toma Keser, Nathalie Gerstner, Mustafa Büyüközkan, Tamara Štambuk, Maurice H. J. Selman, Igor Rudan, Ozren Polašek, Caroline Hayward, Hassen Al-Amin, Karsten Suhre, Gabi Kastenmüller, Gordan Lauc, and Jan Krumsiek. A strategy to incorporate prior knowledge into correlation network cutoff selection. *Nature Communications*, 2020. URL http://dx.doi.org/10.1038/s41467-020-18675-3.

Michael Bereket and Theofanis Karaletsos. Modelling cellular perturbations with the sparse additive mechanism shift variational autoencoder, 2024. URL https://arxiv.org/abs/2311.02794.

Robert Blassberg, Harshil Patel, Thomas Watson, Mina Gouti, Vicki Metzis, M. Joaquina Delás, and James Briscoe. Sox2 levels regulate the chromatin occupancy of wnt mediators in epiblast progenitors responsible for vertebrate body formation. *Nature Cell Biology*, 2022. URL http://dx.doi.org/10.1038/s41556-022-00910-2.

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*, (11), 2023. URL http://dx.doi.org/10.1038/s41592-023-01969-x.

Yun-Shen Chan, Jonathan Göke, Jia-Hui Ng, Xinyi Lu, Kevin Andrew Uy Gonzales, Cheng-Peow Tan, Wei-Quan Tng, Zhong-Zhi Hong, Yee-Siang Lim, and Huck-Hui Ng. Induction of a human pluripotent state with distinct regulatory circuitry that resembles preimplantation epiblast. *Cell Stem Cell*, 13(6):663–675, dec 2013. doi: 10.1016/j.stem.2013.11.015. URL http://dx.doi.org/10.1016/j.stem.2013.11.015.

Ming Chen, Zhewei Wei, Zengfeng Huang, Bolin Ding, and Yaliang Li. Simple and deep graph convolutional networks. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pp. 1725–1735. PMLR, 2020. URL http://proceedings.mlr.press/v119/chen20v.html.

Changxi Chi, Jun Xia, Jingbo Zhou, Jiabei Cheng, Chang Yu, and Stan Z. Li. Grape: Heterogeneous graph representation learning for genetic perturbation with coding and non-coding biotype, 2025. URL https://arxiv.org/abs/2505.03853.

Jiale Deng and Yanyan Shen. Self-interpretable graph learning with sufficient and necessary explanations. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024. URL https://ojs.aaai.org/index.php/AAAI/article/view/29059.

Daniel Dimitrov, Stefan Schrod, Martin Rohbeck, and Oliver Stegle. Interpretation, extrapolation and perturbation of single cells. *Nature Reviews Genetics*, 2026. URL http://dx.doi.org/10.1038/s41576-025-00920-4.

Vijay Prakash Dwivedi and Xavier Bresson. A generalization of transformer networks to graphs, 2020. URL https://arxiv.org/abs/2012.09699.

Fang Fang, Yifeng Xu, Kai-Khen Chew, Xi Chen, Huck-Hui Ng, and Paul Matsudaira. Coactivators p300 and cbp maintain the identity of mouse embryonic stem cells by mediating long-range chromatin structure. *Stem Cells*, 32(7):1805–1816, june 2014. doi: 10.1002/stem.1705. URL http://dx.doi.org/10.1002/stem.1705.

Nina S. Funa, Karen A. Schachter, Mads Lerdrup, Jenny Ekberg, Katja Hess, Nikolaj Dietrich, Christian Honoré, Klaus Hansen, and Henrik Semb. β-catenin regulates primitive streak induction through collaborative interactions with smad2/smad3 and oct4. *Cell Stem Cell*, 16(6):639–652, june 2015. ISSN 1934-5909. doi: 10.1016/j.stem.2015.03.008. URL http://dx.doi.org/10.1016/j.stem.2015.03.008.

Jonathan Göke, Marc Jung, Sarah Behrens, Lukas Chavez, Sean O'Keeffe, Bernd Timmermann, Hans Lehrach, James Adjaye, and Martin Vingron. Combinatorial binding in human and mouse embryonic stem cells identifies conserved enhancers active in early embryonic development. *PLoS Computational Biology*, 7(12):e1002304, dec 2011. doi: 10.1371/journal.pcbi.1002304. URL http://dx.doi.org/10.1371/journal.pcbi.1002304.

William L. Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pp. 1024–1034, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/5dd9db5e033da9c6fb5ba83c7a7ebea9-Abstract.html.

Siyu He, Yuefei Zhu, Daniel Naveed Tavakol, Haotian Ye, Yeh-Hsing Lao, Zixian Zhu, Cong Xu, Shradha Chauhan, Guy Garty, Raju Tomer, Gordana Vunjak-Novakovic, James Zou, Elham Azizi, and Kam W. Leong. Squidiff: predicting cellular development and responses to perturbations using a diffusion model. *Nature Methods*, (1), 2025. URL http://dx.doi.org/10.1038/s41592-025-02877-y.

Weihua Hu, Matthias Fey, Marinka Zitnik, Yuxiao Dong, Hongyu Ren, Bowen Liu, Michele Catasta, and Jure Leskovec. Open graph benchmark: Datasets for machine learning on graphs. In *Proc. of NeurIPS*, 2020.

Kenji Kamimoto, Blerta Stringa, Christy M. Hoffmann, Kunal Jindal, Lilianna Solnica-Krezel, and Samantha A. Morris. Dissecting cell identity via network inference and in silico gene perturbation. *Nature*, (7949), 2023. URL http://dx.doi.org/10.1038/s41586-022-05688-9.

Mohammad Lotfollahi, F. Alexander Wolf, and Fabian J. Theis. scgen predicts single-cell perturbation responses. *Nature Methods*, 2019. URL http://dx.doi.org/10.1038/s41592-019-0494-8.

Mohammad Lotfollahi, Anna Klimovskaia Susmelj, Carlo De Donno, Leon Hetzel, Yuge Ji, Ignacio L Ibarra, Sanjay R Srivatsan, Mohsen Naghipourfar, Riza M Daza, Beth Martin, Jay Shendure, Jose L McFaline-Figueroa, Pierre Boyeau, F Alexander Wolf, Nafissa Yakubova, Stephan Günnemann, Cole Trapnell, David Lopez-Paz, and Fabian J Theis. Predicting cellular responses to complex perturbations in high-throughput screens. *Molecular Systems Biology*, 2023. URL http://dx.doi.org/10.15252/msb.202211517.

Chunyu Ma, Han Zhang, Yiwei Rao, Xinyu Jiang, Boheng Liu, Zhikang Sun, Zhenyu Song, Yuan Gao, Yuhao Cui, Xinyu Liu, and Zedong Li. Ai-driven virtual cell models in preclinical research: technical pathways, validation mechanisms, and clinical translation potential. *npj Digital Medicine*, 9(1):25, dec 2025. ISSN 2398-6352. doi: 10.1038/s41746-025-02198-6. URL http://dx.doi.org/10.1038/s41746-025-02198-6.

Cody T. Mowery, Jacob W. Freimer, Zeyu Chen, Salvador Casaní-Galdón, Jennifer M. Umhoefer, Maya M. Arce, Ketrin Gjoni, Bence Daniel, Katalin Sandor, Benjamin G. Gowen, Vinh Nguyen, Dimitre R. Simeonov, Christian M. Garrido, Gemma L. Curie, Ralf Schmidt, Zachary Steinhart, Ansuman T. Satpathy, Katherine S. Pollard, Jacob E. Corn, Bradley E. Bernstein, Chun Jimmie Ye, and Alexander Marson. Systematic decoding of cis gene regulation defines context-dependent control of the multi-gene costimulatory receptor locus in human t cells. *Nature Genetics*, 2024. URL http://dx.doi.org/10.1038/s41588-024-01743-5.

Shreyasi Mukherjee, Praneet Chaturvedi, Scott A Rankin, Margaret B Fish, Marcin Wlizla, Kitt D Paraiso, Melissa MacDonald, Xiaoting Chen, Matthew T Weirauch, Ira L Blitz, Ken WY Cho, and Aaron M Zorn. Sox17 and β-catenin co-occupy wnt-responsive enhancers to govern the endoderm gene regulatory network. *eLife*, 2020. URL http://dx.doi.org/10.7554/elife.58029.

Shreyasi Mukherjee, David M. Luedeke, Leslie McCoy, Makiko Iwafuchi, and Aaron M. Zorn. Sox transcription factors direct tcf-independent wnt/β-catenin responsive transcription to govern cell fate in human pluripotent stem cells. *Cell Reports*, 2022. ISSN 2211-1247. URL https://www.sciencedirect.com/science/article/pii/S2211124722010658.

Thomas M. Norman, Max A. Horlbeck, Joseph M. Replogle, Alex Y. Ge, Albert Xu, Marco Jost, Luke A. Gilbert, and Jonathan S. Weissman. Exploring genetic interaction manifolds constructed from rich single-cell phenotypes. *Science*, 2019. URL https://www.science.org/doi/10.1126/science.aax4438.

Kuan Pang, Yanay Rosen, Kasia Kedzierska, Ziyuan He, Abhejit Rajagopal, Claire E. Gustafson, Grace Huynh, and Jure Leskovec. Pulsar: a foundation model for multi-scale and multicellular biology. *bioRxiv*, 2025. URL https://www.biorxiv.org/content/early/2025/11/26/2025.11.24.685470.

Lei S. Qi, Matthew H. Larson, Luke A. Gilbert, Jennifer A. Doudna, Jonathan S. Weissman, Adam P. Arkin, and Wendell A. Lim. Repurposing crispr as an rna-guided platform for sequence-specific control of gene expression. *Cell*, 2013. URL https://www.sciencedirect.com/science/article/pii/S0092867413002110.

Eliot Ragueneau, Chuqiao Gong, Pierre Sinquin, Cristoffer Sevilla, Deidre Beavers, Alexander Grentner, Johannes Griss, Gregory F J Hogue, Nancy T Li, Lisa Matthews, Bruce May, Marija Milacic, Helia Mohammadi, Robert Petryszak, Karen Rothfels, Veronica Shamovsky, Ralf Stephan, Krishna Tiwari, Joel Weiser, Adam Wright, Marc Gillespie, Guanming Wu, Lincoln Stein, Henning Hermjakob, and Peter D'Eustachio. The reactome knowledgebase 2026. *Nucleic Acids Research*, 2025. URL http://dx.doi.org/10.1093/nar/gkaf1223.

Joseph M Replogle, Reuben A Saunders, Angela N Pogson, et al. Mapping information-rich genotype-phenotype landscapes with genome-scale perturb-seq. *Cell*, 185(14):2559–2575.e28, 2022. doi: 10.1016/j.cell.2022.05.013.

Yusuf Roohani, Kexin Huang, and Jure Leskovec. Predicting transcriptional outcomes of novel multigene perturbations with gears. *Nature Biotechnology*, 2023. URL http://dx.doi.org/10.1038/s41587-023-01905-6.

Yusuf H. Roohani, Tony J. Hua, Po-Yuan Tung, Lexi R. Bounds, Feiqiao B. Yu, Alexander Dobin, Noam Teyssier, Abhinav Adduri, Alden Woodrow, Brian S. Plosky, Reshma Mehta, Benjamin Hsu, Jeremy Sullivan, Chiara Ricci-Tam, Nianzhen Li, Julia Kazaks, Luke A. Gilbert, Silvana Konermann, Patrick D. Hsu, Hani Goodarzi, and Dave P. Burke. Virtual cell challenge: Toward a turing test for the virtual cell. *Cell*, 2025. URL http://dx.doi.org/10.1016/j.cell.2025.06.008.

Michael Sejr Schlichtkrull, Nicola De Cao, and Ivan Titov. Interpreting graph neural networks for {nlp} with differentiable edge masking. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=WznmQa42ZAx.

Robert A. Sierra, Nathan P. Hoverter, Ricardo N. Ramirez, Linh M. Vuong, Ali Mortazavi, Bradley J. Merrill, Marian L. Waterman, and Peter J. Donovan. Tcf7l1 suppresses primitive streak gene expression to support human embryonic stem cell pluripotency. *Development*, 145(4), feb 2018. doi: 10.1242/dev.161075. URL http://dx.doi.org/10.1242/dev.161075.

Xiaotang Wang, Yun Zhu, Haizhou Shi, Yongchao Liu, and Chuntao Hong. Graph triple attention network: A decoupled perspective, 2024. URL https://arxiv.org/abs/2408.07654.

Xiaotang Wang, Xuanwei Lin, Yun Zhu, Hao Li, and Yongqi Zhang. Learning crossmodal interaction patterns via attributed bipartite graphs for single-cell omics. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026a. URL https://openreview.net/forum?id=swf3Hbc3Qe.

Xiaotang Wang, Yun Zhu, Haizhou Shi, Yongchao Liu, and Yongqi Zhang. Unigap: A universal and adaptive graph upsampling approach to mitigate over-smoothing in node classification tasks, 2026b. URL https://arxiv.org/abs/2407.19420.

Zhiting Wei, Yiheng Wang, Yicheng Gao, Shuguang Wang, Ping Li, Duanmiao Si, Yuli Gao, Siqi Wu, Danlu Li, Kejing Dong, Xingbo Yang, Chen Tang, Shaliu Fu, Xiaohan Chen, Wannian Li, Yuzhou You, Chen Zhang, Aibin Liang, Guohui Chuai, and Qi Liu. Benchmarking algorithms for generalizable single-cell perturbation response prediction. *Nature Methods*, 2025. URL http://dx.doi.org/10.1038/s41592-025-02980-0.

Yang Xu, Stephen Fleming, Matthew Tegtmeyer, Steven A. McCarroll, and Mehrtash Babadi. Explainable modeling of single-cell perturbation data using attention and sparse dictionary learning. *Cell Systems*, 2025. URL http://dx.doi.org/10.1016/j.cels.2025.101245.

Xiaodong Yang, Guole Liu, Guihai Feng, Dechao Bu, and Pengfei et al. Wang. Genecompass: deciphering universal gene regulatory mechanisms with a knowledge-informed cross-species foundation model. *Cell Research*, 2024. URL http://dx.doi.org/10.1038/s41422-024-01034-y.

Rex Ying, Dylan Bourgeois, Jiaxuan You, Marinka Zitnik, and Jure Leskovec. Gnnexplainer: Generating explanations for graph neural networks, 2019. URL https://arxiv.org/abs/1903.03894.

Hao Yuan, Haiyang Yu, Jie Wang, Kang Li, and Shuiwang Ji. On explainability of graph neural networks via subgraph explorations, 2021. URL https://arxiv.org/abs/2102.05152.

Shilu Zhang, Saptarshi Pyne, Stefan Pietrzak, Spencer Halberg, Sunnie Grace McCalla, Alireza Fotuhi Siahpirani, Rupa Sridharan, and Sushmita Roy. Inference of cell type-specific gene regulatory networks on cell lineages from single cell omic datasets. *Nature Communications*, 2023. URL http://dx.doi.org/10.1038/s41467-023-38637-9.

Shasha Zhou, Mingyu Huang, and Ke Li. Position: Genomic model research must move beyond anecdotal evaluation of interpretability methods, 2026. URL https://arxiv.org/abs/2606.07607.

Ouyang Zhu and Jun Li. Scouter predicts transcriptional responses to genetic perturbations with large language model embeddings. *Nature Computational Science*, 2025. URL http://dx.doi.org/10.1038/s43588-025-00912-8.

SUPPLEMENTARY MATERIAL

A Details of Partially Trusted Prior Graph Construction **14**

A.1 Gene–gene Support Edges from Metadata-based Gene Similarity 14

A.2 TF–gene Regulatory Edges from CellOracle-based Regulatory Priors 15

A.3 Pathway-related Edges from Reactome Pathway Membership and Hierarchy 15

B Experimental Details **15**

B.1 Details of the Datasets 15

B.2 Evaluation Metrics 16

B.3 Hyperparameters 17

B.4 Extended DEG-focused Evaluation 18

B.5 Fine-grained Ablations of Calibration and Mediator Propagation 18

C Case of Biological Interpretation and Discovery **19**

C.1 Case Study: Biological Interpretation of *CTNNB1* Knockdown 19

C.2 Case Study: Biological Interpretation of *EP300* Knockdown 20

D Complexity Analysis **21**

D.1 Time Complexity 22

D.2 Space Complexity 22

E Limitations **22**

F Broader impact **23**

A DETAILS OF PARTIALLY TRUSTED PRIOR GRAPH CONSTRUCTION

A.1 GENE–GENE SUPPORT EDGES FROM METADATA-BASED GENE SIMILARITY

Gene–gene support edges e_{gg} are constructed from gene-level biological metadata via pretrained gene representations. Specifically, each gene $g \in \mathcal{V}_g$ is associated with an embedding produced by the pretrained gene foundation model GeneCompass (Yang et al., 2024), which incorporates multiple sources of biological prior knowledge during pretraining, including promoter information, gene co-expression relationships, gene family information, and TF–target regulatory relationships.

Formally, for each gene g, we obtain

$$\mathbf{u}_g = \mathrm{GeneCompass}(g) \in \mathbb{R}^{d_u},$$

where $\mathrm{GeneCompass}(\cdot)$ denotes the pretrained embedding function.

We measure pairwise gene similarity using cosine similarity:

$$s(g, g') = \frac{\mathbf{u}_g^\top \mathbf{u}_{g'}}{\|\mathbf{u}_g\|_2 \|\mathbf{u}_{g'}\|_2}.$$

We then construct a sparse directed support graph by connecting each target gene g to its top-k_{gg} most similar neighbors whose similarity exceeds a threshold τ_{gg}:

$$(g', g) \in \mathcal{E}_{gg} \quad \text{if} \quad g' \in \mathrm{TopK}_{k_{gg}}(g) \text{ and } s(g, g') \geq \tau_{gg}.$$

Here, $\mathrm{TopK}_{k_{gg}}(g)$ denotes the set of the k_{gg} genes with highest similarity to g among the genes in $\mathcal{V}_g$, excluding g itself.

These edges do not necessarily represent direct regulatory interactions. Instead, they provide auxiliary support for information propagation by capturing functional similarity and shared biological context. Consequently, they may introduce spurious shortcut connections and are therefore treated as weak, partially trusted edges, which are later regularized toward sparsity during graph calibration.

A.2 TF–gene Regulatory Edges from CellOracle-based Regulatory Priors

TF–gene regulatory edges e_{tg} are derived from reference-genome-based regulatory priors provided by CellOracle (Kamimoto et al., 2023). In particular, we use the CellOracle base GRN prior, which encodes candidate TF–target relationships obtained from TF-binding motif information together with genome annotation, rather than the downstream cell-state-specific GRN inferred from expression data.

Concretely, for each transcription factor t and gene g, CellOracle identifies putative regulatory interactions by scanning regulatory DNA regions associated with g (e.g., promoter and, when available, enhancer regions) for TF-binding motifs, together with genome annotation and motif databases. This yields a directed prior edge set

$$(t, g) \in \mathcal{E}_{tg},$$

where $t \in \mathcal{V}_g$ is a gene node acting as a transcription factor, and both t and g are restricted to genes present in the dataset vocabulary.

These edges encode biologically plausible regulatory relationships, but they are not context-specific and may contain both false positives and missing interactions. We therefore treat $\mathcal{E}_{tg}$ as a partially trusted regulatory prior rather than ground truth. Their reliability is subsequently refined by the context-specific calibration procedure in Section 3, which uses control data to adjust edge weights and align the prior graph with observed expression patterns.

A.3 Pathway-related Edges from Reactome Pathway Membership and Hierarchy

Pathway-related edges are constructed from the Reactome pathway database (Ragueneau et al., 2025), which provides curated pathway definitions together with a hierarchical organization of biological processes. We consider two types of pathway-related edges.

Gene–pathway membership edges. For each pathway $p \in \mathcal{V}_p$, Reactome provides a set of annotated member genes. We construct bipartite edges

$$(g, p) \in \mathcal{E}_{gp} \quad \text{if gene } g \text{ is annotated as a member of pathway } p,$$

where $g \in \mathcal{V}_g$ is restricted to genes present in the dataset vocabulary. These edges allow the model to aggregate gene-level perturbation signals into pathway-level representations.

Pathway–pathway hierarchy edges. Reactome organizes pathways hierarchically through parent–child relationships. We encode this hierarchy as directed edges

$$(p', p) \in \mathcal{E}_{pp} \quad \text{if } p' \text{ is a parent pathway of } p \text{ in the Reactome hierarchy.}$$

These edges support information propagation across different levels of pathway granularity, allowing the model to reason over both broad functional programs and more specific subpathways.

Together, $\mathcal{E}_{gp}$ and $\mathcal{E}_{pp}$ define a structured pathway subgraph that supports multi-level reasoning from genes to pathways and back to genes. Compared with gene-level support edges, these pathway-related edges encode curated mechanistic knowledge and are therefore encouraged to be preserved during graph calibration unless contradicted by data.

B Experimental Details

B.1 Details of the Datasets

We conduct experiments on four widely-used benchmark datasets for genetic perturbation: Adamson (Adamson et al., 2016), Norman (Norman et al., 2019), Replogle (Replogle et al., 2022), and VCC (Roohani et al., 2025). Summary statistics are provided in Table 3.

Table 3: Statistics of the original datasets.

Dataset	#Genes	#Control cells	#Perturbed cells	#Perturb types	Split	Cell line
Adamson (Adamson et al., 2016)	5,060	24,263	44,340	86	57/7/22	K562
Norman (Norman et al., 2019)	5,045	7,353	83,852	236	106/24/96	K562
Replogle (Replogle et al., 2022)	8,563	10,691	299,694	2,057	1,542/309/206	K562
VCC (Roohani et al., 2025)	18,080	38,176	183,097	300	150/50/100	h1 hESCs

Adamson. The Adamson dataset originates from a Perturb-seq study of the mammalian unfolded protein response (UPR), where CRISPR interference (CRISPRi (Qi et al., 2013)) perturbations were combined with droplet-based single-cell RNA-seq to dissect UPR regulatory programs. In the original study, two genome-scale CRISPRi screens were used to identify genes whose repression perturbs ER homeostasis, and then approximately 100 hits were profiled by Perturb-seq for fine-grained transcriptional characterization. For machine learning evaluation, the commonly used benchmark is the GEARS-processed version, which is log-normalized and filtered to the top 5,000 highly variable genes.

Norman. The Norman dataset was introduced to study genetic interaction manifolds from rich single-cell phenotypes. Biologically, it is a gain-of-function Perturb-seq dataset based on CRISPR activation (CRISPRa) in K562 cells, designed to characterize both single-gene and combinatorial genetic perturbations at single-cell resolution. In current perturbation-prediction benchmarks, the most commonly used release is also the GEARS-processed version. Since it contains both single and double perturbations, some methods treat these two settings separately. In contrast, our framework naturally scales to perturbations involving different numbers of target genes.

Replogle. The Replogle dataset family comes from a genome-scale Perturb-seq study that targeted all expressed genes with CRISPRi across more than 2.5 million human cells. The benchmark dataset most commonly used in perturbation-prediction work is the K562 Essential release, which contains 310,385 single-cell RNA-seq profiles from K562 cells, including 10,691 control cells and 299,694 perturbed cells, spanning 2,057 distinct gene-knockdown conditions with matched controls.

VCC (Virtual Cell Challenge). The VCC dataset is the inaugural benchmark dataset released by Arc Institute for the Virtual Cell Challenge. Unlike earlier perturbation-prediction benchmarks that are dominated by immortalized lines such as K562, VCC is built in H1 human embryonic stem cells (H1 hESCs) to explicitly test context generalization. The challenge dataset contains 300 target genes, and is organized into train, validation, and held-out test perturbation sets with 150/50/100.

B.2 Evaluation Metrics

For the Adamson, Norman, and Replogle datasets, following prior works such as GEARS and GRAPE, we evaluate model performance using Root Mean Square Error (RMSE), Direction Accuracy Coefficient (DAC), and Pearson Correlation Coefficient (PCC). For the VCC benchmark, we follow the evaluation protocol of the Virtual Cell Challenge (OOD Generalization to Unseen Cellular Contexts) and adopt the official metrics, including Mean Absolute Error (MAE), Differential Expression Score (DES), and Perturbation Discrimination Score (PDS).

RMSE. Root Mean Square Error measures the overall discrepancy between predicted and true expression vectors, for each perturbation k:

$$\mathrm{RMSE}_k = \frac{1}{\sqrt{N_g}} \|\hat{y}^k - y^k\|_2$$

where N_g is the total number of genes.

DAC. Direction Accuracy Coefficient measures whether the predicted perturbation effect has the correct sign for each gene:

$$\mathrm{DAC}_k = \frac{1}{N_g} \sum_{g=1}^{N_g} \mathbf{1}\Big[\mathrm{sign}(\hat{\Delta}_{k,g}) = \mathrm{sign}(\Delta_{k,g})\Big] .$$

where $\mathbf{1}[\cdot]$ denotes the indicator function.

PCC. Pearson Correlation Coefficient on perturbation effects quantifies the linear correlation between predicted and true perturbation-induced changes:

$$\text{PCC-}\Delta_k = \frac{(\hat{\Delta}_k - \overline{\hat{\Delta}_k})^\top (\Delta_k - \overline{\Delta_k})}{\|\hat{\Delta}_k - \overline{\hat{\Delta}_k}\|_2 \, \|\Delta_k - \overline{\Delta_k}\|_2},$$

where $\overline{\hat{\Delta}_k}$ and $\overline{\Delta_k}$ are the mean values of the corresponding vectors.

MAE. Mean Absolute Error measures the average absolute deviation between predicted and true post-perturbation expression:

$$\text{MAE}_k = \frac{1}{N_g} \|\hat{y}_k - y_k\|_1.$$

DES. Differential Expression Score evaluates whether the model recovers the correct set of differentially expressed (DE) genes. For each perturbation condition k, DE genes are identified separately from predicted and true data by a Wilcoxon rank-sum test between perturbed and control cells, with Benjamini–Hochberg correction at FDR ≤ 0.05. Let $G_{k,\text{true}}$ and $G_{k,\text{pred}}$ denote the resulting DE gene sets, and let $n_{k,\text{true}} = |G_{k,\text{true}}|$. If $|G_{k,\text{pred}}| \leq n_{k,\text{true}}$, we define

$$\text{DES}_k = \frac{|G_{k,\text{pred}} \cap G_{k,\text{true}}|}{n_{k,\text{true}}}.$$

Otherwise, we truncate the predicted DE list to the top $n_{k,\text{true}}$ genes ranked by absolute log fold change, denoted by $\tilde{G}_{k,\text{pred}}$, and compute

$$\text{DES}_k = \frac{|\tilde{G}_{k,\text{pred}} \cap G_{k,\text{true}}|}{n_{k,\text{true}}}.$$

PDS. Perturbation Discrimination Score evaluates whether the predicted perturbation effect is closest to its true counterpart among all perturbations. Let x^{ntc} denote the non-targeting control profile. For each perturbation condition k, define

$$\delta^k = y^k - x^{\text{ntc}}, \qquad \hat{\delta}^k = \hat{y}^k - x^{\text{ntc}}.$$

For the evaluated perturbation k, we compare its predicted effect $\hat{\delta}^k$ against the true effects of all perturbation conditions k', and compute

$$d_{k,k'} = \|\hat{\delta}^k - \delta^{k'}\|_1.$$

We rank all k' by ascending $d_{k,k'}$. Let rank_k denote the rank of the correct perturbation $k' = k$. The score is

$$\text{PDS}_k = 1 - \frac{\text{rank}_k - 1}{N_{\text{pert}}},$$

where N_{pert} is the total number of perturbation conditions.

B.3 Hyperparameters

Experimental results are reported on the hyperparameter settings below, where we choose the settings that achieve the highest performance on the validation set. We choose hyperparameter grids that do not necessarily give optimal performance, but hopefully cover enough regimes so that each model is reasonably evaluated on each dataset.

- learning_rate $\in \{1e-3, 5e-4, 1e-4, 5e-5, 1e-5\}$
- weight_decay $\in \{1e-4, 5e-5, 1e-5, 5e-6, 1e-6\}$
- dropout $\in \{0, 0.1, 0.3, 0.5, 0.8\}$

For PerturbDigger,

- ID embedding dims $\in \{64, 128, 256, 512\}$
- hidden dims $\in \{64, 128, 256, 512\}$
- $\lambda_{\text{nec}} \in \{0, 0.5, 1\}$
- layer_num $\in \{1, 2\}$

Table 4: DEG-focused evaluation of PerturbDigger across the Adamson, Norman, Replogle, and VCC datasets. DE20 and DE40 denote evaluation over the top-20 and top-40 observed differentially expressed genes identified independently for each perturbation relative to its matched controls.

Dataset	RMSE↓	DAC↑	DAC-DE20↑	DAC-DE40↑	PCC↑	PCC-DE20↑	PCC-DE40↑	DES↑	PDS↑
Adamson	0.067	0.786	0.927	0.979	0.801	0.854	0.885	0.329	0.766
Norman	0.036	0.836	0.864	0.882	0.565	0.674	0.691	0.368	0.804
Replogle	0.072	0.729	0.772	0.834	0.745	0.821	0.840	0.319	0.703
VCC	0.058	0.791	0.833	0.861	0.710	0.772	0.811	0.357	0.712

Table 5: Fine-grained ablations of the two calibration stages and the relation-specific regularizers in Stage II on the Adamson dataset.

	Variant	Stage I	Stage II	$\mathcal{R}_{tg}$	$\mathcal{R}_{\text{path}}$	$\mathcal{R}_{gg}$	RMSE↓	DAC↑	PCC↑
(a)	Full model (PerturbDigger)	✓	✓	✓	✓	✓	**0.067**	**0.786**	**0.801**
(b)	Stage I only	✓	–	–	–	–	0.075	0.651	0.704
(c)	Stage II only	–	✓	–	✓	✓	0.074	0.689	0.711
(d)	w/o pathway-edge retention	✓	✓	✓	–	✓	0.069	0.751	0.772
(e)	w/o gene–gene sparsity	✓	✓	✓	✓	–	0.071	0.734	0.759
(f)	No calibration	–	–	–	–	–	0.077	0.643	0.692

B.4 Extended DEG-focused Evaluation

To provide a more comprehensive evaluation of perturbation-response prediction, we further apply the same Differential Expression Score (DES) and Perturbation Discrimination Score (PDS) evaluation protocol used for VCC to the Adamson, Norman, and Replogle datasets. DES evaluates whether the model recovers the observed set of differentially expressed genes, whereas PDS evaluates whether the predicted response retains perturbation-specific signals that distinguish the target perturbation from other conditions.

We additionally evaluate DAC and PCC on genes exhibiting the strongest observed perturbation responses. Specifically, for each perturbation, we independently identify the top-20 and top-40 differentially expressed genes relative to its matched controls and compute DAC-DE20, DAC-DE40, PCC-DE20, and PCC-DE40 over the corresponding gene subsets. Unlike evaluations over all genes, these metrics place greater emphasis on genes most strongly affected by each perturbation and reduce the influence of genes with negligible observed changes.

As shown in Table 4, PerturbDigger achieves strong performance across all datasets and evaluation metrics. The performance in DAC-DE20, DAC-DE40, PCC-DE20, and PCC-DE40 demonstrate that its predictive advantages are preserved when evaluation is restricted to genes with the strongest observed perturbation responses. Meanwhile, the strong performance in DES and PDS indicate that PerturbDigger accurately recovers DEG sets and preserves perturbation-specific response patterns.

B.5 Fine-grained Ablations of Calibration and Mediator Propagation

Calibration stage. We first conduct fine-grained ablations to disentangle the contributions of the two calibration stages and the relation-specific regularizers in Stage II. Let $\mathcal{R}_{tg}$, $\mathcal{R}_{\text{path}}$, and $\mathcal{R}_{gg}$ denote the TF–gene alignment, pathway-edge retention, and gene–gene sparsity regularizers in Equation 3, respectively. The results on the Adamson dataset are reported in Table 5.

Comparing variants (a), (c), (d), and (e), removing any of the three regularizers consistently degraded performance across RMSE, DAC, and PCC. In particular, removing TF–gene alignment produced the largest degradation, while removing either of the other two regularizers also resulted in consistent performance degradation. These results indicate that the three regularizers impose complementary rather than redundant structural constraints.

Variants (b) and (c) isolate the contributions of the two calibration stages. Both single-stage variants outperformed the no calibration (f) but remained consistently below the full model (a). These results support their complementary roles: Stage I provides a control-supported TF–gene relevance prior,

Table 6: Ablations of different mediator propagation orders on the Adamson dataset.

Variant	Propagation order	RMSE↓	DAC↑	PCC↑
Proposed sequential order (PerturbDigger)	L1→L2→L3→L4	**0.067**	**0.786**	**0.801**
Parallel mediators	L1→(L2, L3)→L4	0.074	0.699	0.725
Reversed sequential order	L1→L3→L2→L4	0.076	0.672	0.703
Unconstrained propagation	Four-layer Graph Transformer	0.081	0.624	0.691

whereas Stage II integrates this prior with context-specific masked reconstruction and relation-specific structural constraints.

Mediator propagation order. We further investigate whether the proposed mediator propagation order provides an effective inductive bias. Our sequential design follows the intended information dependency: gene representations are first updated through gene–gene and TF–gene regulation, subsequently aggregated into pathway-level responses, and finally propagated back to genes. We compare this design with a parallel mediator architecture, a reversed sequential order, and an unconstrained 4-layer graph transformer. The results on the Adamson dataset are shown in Table 6.

The proposed TF-before-pathway ordering achieved the best performance, outperforming both the reversed order and the parallel variant. The comparison with the reversed variant supports the benefit of allowing TF-regulated gene representations to serve as inputs to subsequent pathway-level aggregation. The comparison with the parallel variant further shows that sequential dependency is beneficial beyond independently processing and combining TF- and pathway-mediated signals. Together, these results provide direct empirical evidence for the proposed propagation order as an effective inductive bias for perturbation-response prediction. Moreover, all three biologically structured order consistently outperformed the unconstrained four-layer graph transformer, consistent with our hypothesis that repeatedly mixing heterogeneous biological relations can weaken relation-specific information and introduce less informative propagation paths.

C Case of Biological Interpretation and Discovery

C.1 Case Study: Biological Interpretation of *CTNNB1* Knockdown

Here, we provide a detailed analysis of the mechanistic subgraph recovered for *CTNNB1* knockdown.

As shown in Fig 5, at the TF level, the recovered mediator nodes are well aligned with the known β-catenin transcriptional machinery. In canonical WNT signaling, nuclear β-catenin primarily functions by forming activating complexes with TCF/LEF-family transcription factors. Therefore, the appearance of mediators such as *TCF7L1*, *TCF7*, *TCF7L2*, and *LEF1* in the inferred subgraph is not incidental, but directly consistent with curated pathway knowledge. In particular, *TCF7L1* is especially relevant in hESCs, where prior studies have shown that it suppresses primitive-streak gene expression and helps maintain the primed pluripotent state. From this perspective, the PerturbDigger-recovered subgraph is biologically coherent: reduced *CTNNB1* activity is expected to weaken TCF/LEF-mediated transcriptional activation and alter the balance between pluripotency maintenance and developmental induction (Funa et al., 2015; Mukherjee et al., 2020).

Importantly, the recovered subgraph also extends beyond the core WNT axis into developmental pathway modules that are highly relevant to the H1 hESC context. Prior studies in human embryonic stem cells have shown that β-catenin does not act alone during primitive-streak induction, but instead collaborates with SMAD2/3 and OCT4 to activate developmental programs. Reactome further organizes these processes into pathway/event nodes such as Germ layer formation at gastrulation, Expression of TBXT (T, Brachyury) in the primitive streak, and Formation of definitive endoderm. Accordingly, when PerturbDigger recovers mediator TFs and pathway nodes involving *SMAD3*, *FOXH1*, *POU5F1*, *NANOG*, *EOMES*, and *TBXT*, the resulting subgraph matches a well-supported literature: *CTNNB1* knockdown dampens not only canonical WNT-responsive transcription, but also primitive-streak / mesendodermal induction programs that depend on coordinated WNT and SMAD signaling (Sierra et al., 2018; Blassberg et al., 2022; Mukherjee et al., 2022).

Taken together, this case study supports the biological plausibility of the recovered mechanistic subgraph from three complementary perspectives. First, the central pathway nodes correspond to

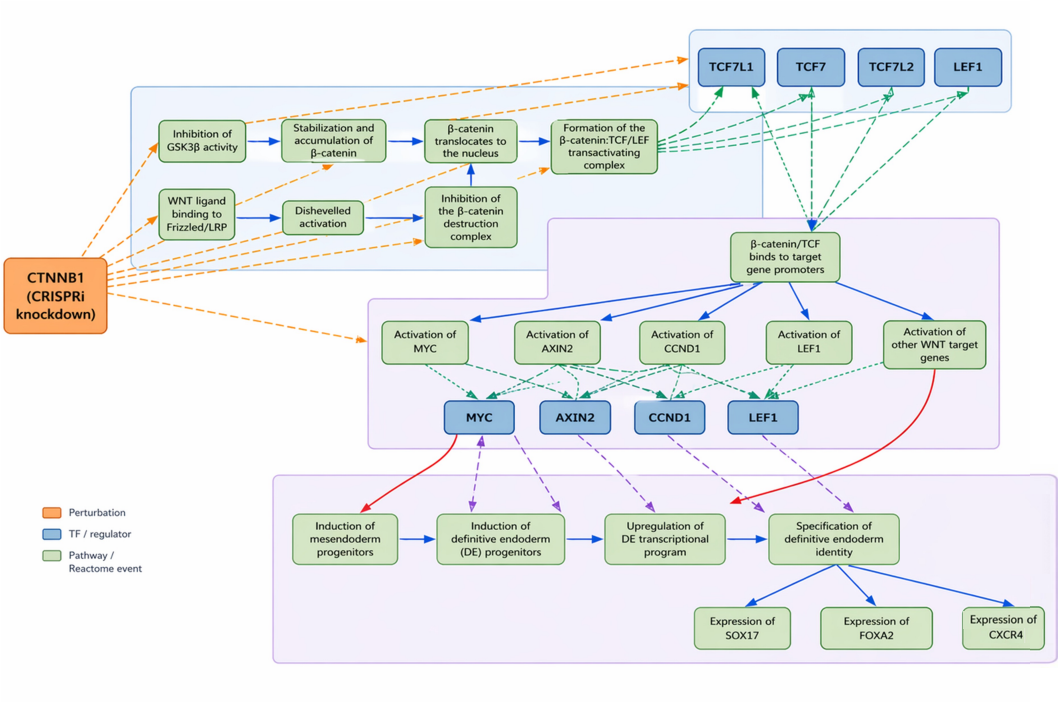


Figure 5: Detailed mechanism subgraph of *CTNNB1* knockdown perturbation.

curated Reactome events along the canonical WNT/β-catenin axis. Second, the mediator TFs are consistent with established β-catenin-associated transcriptional regulators in pluripotent and early developmental settings. Third, the downstream developmental modules align with prior studies showing that β-catenin contributes to primitive-streak and germ-layer induction in human ESC models. Therefore, the PerturbDigger explanation is strongly supported as a literature-consistent and biologically grounded mechanistic hypothesis.

C.2 Case Study: Biological Interpretation of *EP300* Knockdown

Here, we provide a detailed analysis of the mechanistic subgraph recovered for *EP300* knockdown.

As shown in Fig 6, the recovered subgraph highlights a regulatory layer that is clearly distinct from the *CTNNB1* case. At the TF level, the dominant mediator nodes are centered on pluripotency-associated regulators, including *POU5F1*, *SOX2*, and *NANOG*, rather than being driven primarily by the core WNT effector machinery itself. This pattern is highly consistent with the known role of p300 as a transcriptional coactivator and histone acetyltransferase that supports enhancer activity and pluripotency-associated transcriptional regulation. Correspondingly, the central pathway nodes in the recovered subgraph are enriched for chromatin and transcription-related modules, including HATs acetylate histones, EP300 acetylates histone H2A, H2B, H3, H4, and Transcriptional regulation of pluripotent stem cells. Taken together, these observations support a coherent interpretation in which *EP300* knockdown weakens enhancer-associated chromatin activation and reduces transcriptional support for the pluripotency network (Chan et al., 2013; Göke et al., 2011).

At the same time, the recovered subgraph also overlaps with the canonical WNT transcriptional branch, which explains the biological relatedness between *EP300* and *CTNNB1*. Importantly, this overlap occurs at the level of nuclear transcriptional coactivation rather than upstream WNT signal transduction. Reactome explicitly places p300/CBP downstream of nuclear β-catenin through the event Beta-catenin recruits CBP/p300, within TCF dependent signaling in response to WNT. Consistent with this pathway structure, the recovered mediator layer includes TCF/LEF-family factors such as *TCF7L1*, *TCF7*, *TCF7L2*, and *LEF1*, together with pathway nodes related to WNT-responsive transcription. This part of the recovered subgraph therefore supports a specific interpretation: *EP300* knockdown does not primarily perturb the upstream WNT signaling axis, but instead weakens the

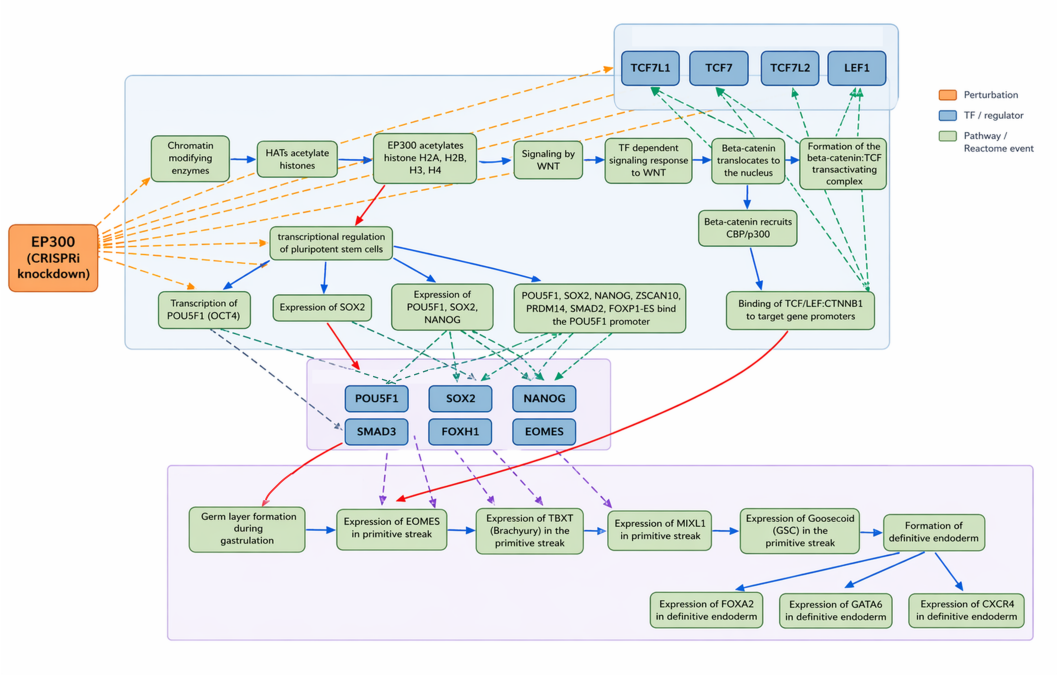


Figure 6: Detailed mechanism subgraph of *EP300* knockdown perturbation.

coactivator layer required for efficient TCF/LEF-mediated transcriptional output (Ragueneau et al., 2025; Fang et al., 2014).

The recovered subgraph further extends into developmental modules that are relevant to the H1 hESC context. In particular, PerturbDigger recovers mediator TFs and pathway nodes involving *SMAD3*, *FOXH1*, and *EOMES*, together with pathway/event nodes such as Germ layer formation at gastrulation, Expression of EOMES in the primitive streak, and Formation of definitive endoderm. Compared with the *CTNNB1* case, this developmental branch is better interpreted as a downstream extension of EP300-dependent chromatin coactivation rather than as a primary signaling axis. In other words, the recovered subgraph indicates that reduced p300 activity propagates from enhancer and pluripotency-associated transcriptional regulation into early developmental programs, thereby attenuating transcriptional programs associated with lineage induction in hESCs (Funa et al., 2015).

Taken together, this case study supports the biological plausibility of the recovered *EP300* mechanistic subgraph from three complementary perspectives. First, its central pathway nodes are concentrated on chromatin acetylation, enhancer activation, and pluripotency-associated transcription, clearly distinguishing it from the more WNT-core-centered *CTNNB1* case. Second, the recovered TCF/LEF-related branch explains why *EP300* remains mechanistically related to *CTNNB1*, while locating this relationship at the level of downstream transcriptional coactivation. Third, the developmental modules extend this interpretation into the H1 hESC context, linking EP300-dependent chromatin regulation to early lineage-associated transcriptional programs. Therefore, the PerturbDigger explanation is strongly supported as a literature-consistent and biologically grounded mechanistic hypothesis rather than an arbitrary post hoc attribution.

D Complexity Analysis

In this section, we analyze the time and space complexity of PerturbDigger. For simplicity, we assume that the hidden feature dimension remains unchanged across layers. Let $N_g = |\mathcal{V}_g|$ and $N_p = |\mathcal{V}_p|$ denote the numbers of gene and pathway nodes, respectively, and $N = N_g + N_p$. We denote by $|\mathcal{E}_{gg}|$, $|\mathcal{E}_{tg}|$, $|\mathcal{E}_{gp}|$, and $|\mathcal{E}_{pp}|$ the numbers of gene–gene, TF–gene, gene–pathway, and pathway–pathway edges, respectively. Let $E = |\mathcal{E}_{gg}| + |\mathcal{E}_{tg}| + |\mathcal{E}_{gp}| + |\mathcal{E}_{pp}|$ be the total number of

prior edges, and d_h is the hidden dimension. In the calibration stage, the sparse-regression step used to construct the TF–gene calibration prior is a one-time pre-processing cost and does not affect the asymptotic complexity of model training or inference. The reconstruction step reuses the propagation operators as the perturbation model, its per-epoch complexity has the same asymptotic order as the main model, up to constant factors. Therefore, we focus on the complexity of one forward pass of PerturbDigger.

D.1 Time Complexity

The primary time overhead of PerturbDigger arises from four ordered propagation modules: Gene–Gene support diffusion (L1), TF–Gene regulatory propagation (L2), Pathway aggregation and interaction (L3), and Pathway–Gene readout (L4).

The L1 consists of a node-wise linear transformation and sparse aggregation over gene–gene support edges. Its time complexity is $\mathcal{O}(N_g d_h^2 + |\mathcal{E}_{gg}|d_h)$. For L2, PerturbDigger computes query/key/value projections for gene and TF nodes, evaluates edge scores on the sparse prior graph, and aggregates messages along the retained sparse edges. The projection cost is $\mathcal{O}(N_g d_h^2)$, and the edge-wise score computation and aggregation cost is $\mathcal{O}(|\mathcal{E}_{tg}|d_h)$. Thus, the time complexity of L2 is $\mathcal{O}(N_g d_h^2 + |\mathcal{E}_{tg}|d_h)$. Although biologically guided pruning retains at most top-k_{tg} edges per target gene, the asymptotic complexity remains dominated by the sparse prior-edge traversal. L3 first aggregates gene-level responses into pathway states over gene–pathway edges, and then propagates across pathway–pathway hierarchy edges. The projection and update cost over gene and pathway nodes is $\mathcal{O}((N_g + N_p)d_h^2)$, while sparse edge traversal contributes $\mathcal{O}((|\mathcal{E}_{gp}| + |\mathcal{E}_{pp}|)d_h)$. Therefore, the time complexity of L3 is $\mathcal{O}((N_g + N_p)d_h^2 + (|\mathcal{E}_{gp}| + |\mathcal{E}_{pp}|)d_h)$. Finally, in L4, pathway states are projected back to genes through sparse gene–pathway membership edges, with complexity $\mathcal{O}(N_g d_h^2 + |\mathcal{E}_{gp}|d_h)$. Combining the four modules, the total time complexity of one forward pass of PerturbDigger is $\mathcal{O}\Big((N_g + N_p)d_h^2 + Ed_h\Big)$, where constant factors from the four propagation steps are omitted.

Compared with dense graph transformers that apply all-pairs attention over all $N = N_g + N_p$ nodes, whose time complexity is typically $\mathcal{O}(N^2 d_h)$, PerturbDigger is significantly more efficient because it restricts message passing to sparse biologically motivated prior edges and further prunes mediator interactions during propagation.

D.2 Space Complexity

For each graph instance, PerturbDigger stores hidden node embeddings, relation-specific projected features, sparse edge scores, and intermediate message buffers. The node-level hidden representations require $\mathcal{O}((N_g + N_p)d_h)$ memory. The sparse edge scores and message buffers over the prior graph require $\mathcal{O}(Ed_h)$. Hence, the overall space complexity of PerturbDigger is $\mathcal{O}\Big((N_g + N_p)d_h + Ed_h\Big)$.

Compared with dense attention models that require storing pairwise attention maps over all node pairs, with space complexity $\mathcal{O}(N^2)$ or higher when feature-dependent buffers are included, PerturbDigger achieves substantially better memory efficiency by operating only on sparse prior edges and relation-specific message passing.

E Limitations

While PerturbDigger achieves strong performance and interpretability, it does not currently incorporate rich edge attributes, which can play a crucial role in capturing fine-grained interactions in graph-based tasks (Hu et al., 2020; Wang et al., 2024; 2026a;b). Modeling such edge-level information requires the integration of more detailed biological priors and suitable encoding strategies.

Moreover, model faithfulness and biological validity are complementary but distinct properties (Zhou et al., 2026). Establishing biological validity generally requires independent perturbation evidence and carefully controlled wet-lab experiments. Nevertheless, PerturbDigger makes this relationship directly testable. Unlike explanations generated only after prediction, its propagation is constrained to a calibrated, biologically grounded mediator graph whose selected edges directly participate in forward computation. This couples biological priors with model decisions, jointly promoting faithfulness to model decisions and alignment with biological evidence.

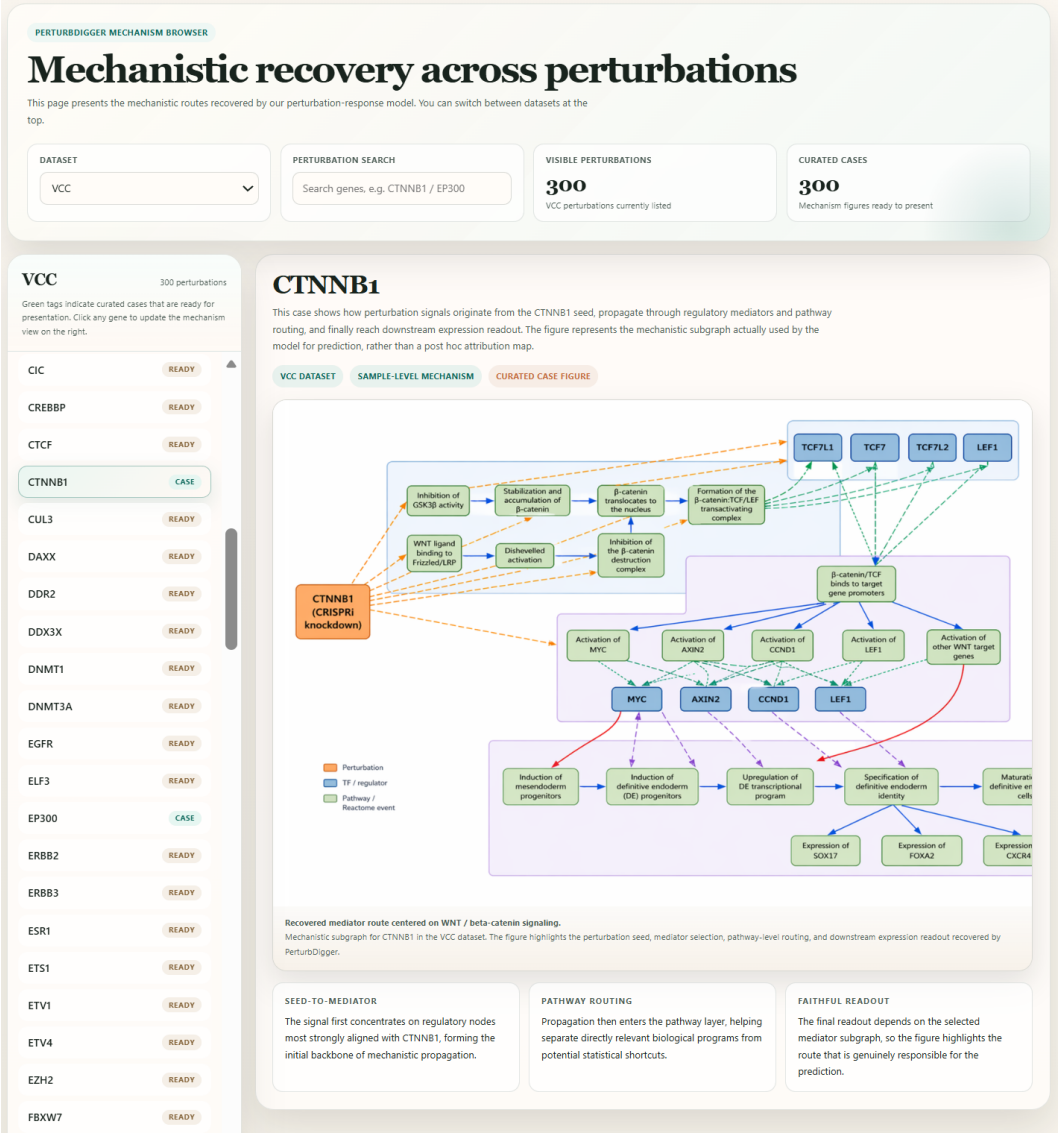


Figure 7: Our **Mechanism Explanation Webpage**. We are committed to maintaining a webpage that records the mechanistic subgraphs we discover, thereby providing references and testable hypotheses for interpretable and mechanism-aware virtual cell models.

F Broader Impact

The potential positive impact is twofold. First, more accurate perturbation response prediction may help prioritize costly or time-consuming wet-lab experiments, especially when exploring large perturbation spaces or cellular contexts with limited experimental data. Second, mediator-centered mechanistic subgraphs may provide interpretable hypotheses about transcription factors, pathways, and regulatory routes involved in cellular responses, supporting more transparent virtual cell modeling and downstream biological investigation. However, if models for mechanistic explanation are applied in biomedical or therapeutic settings, careful validation, uncertainty assessment, and domain-specific oversight will be necessary. We encourage future work to evaluate explanation reliability across diverse biological contexts, quantify uncertainty in recovered mechanisms, and integrate experimental feedback to further improve the trustworthiness of mechanism-aware virtual cell models. Looking ahead, as the webpage shown in Fig 7, we are committed to developing a comprehensive knowledge base, advancing more interpretable and mechanism-aware virtual cell models.