WHEN NEW EXPERIMENTS ARRIVE: AN EVIDENCE-ADJUDICATED SELF-CORRECTION FRAMEWORK FOR VIRTUAL CELLS

Anonymous authors
Paper under double-blind review

ABSTRACT

Virtual-cell models increasingly need to support both accurate perturbation prediction and interpretable biological reasoning. However, numerical models provide precise continuous responses but only implicit biological knowledge, whereas LLM/knowledge-graph-based models offer explicit mechanistic reasoning but coarse categorical predictions. Jointly maintaining these capabilities raises a further challenge: as new experimental evidence arrives, which states should be corrected and which should be preserved? To address these challenges, we propose **VCrectify**, an evidence-adjudicated self-correction framework for virtual cells. It maintains a numerical predictive state for continuous perturbation prediction and an explicit knowledge state for knowledge-grounded mechanistic reasoning. Cross-state disagreement between the two states guides the active evidence acquisition of new perturbation experiments. Once outcomes are revealed, an evidence adjudicator determines which states receive correction. The updated states then guide subsequent experiment selection, forming a continual loop of evidence acquisition and self-correction. Experiments across multiple virtual-cell backbones and perturbation datasets show that VCrectify improves numerical prediction and knowledge-based directional reasoning, generalizes better to unseen cellular contexts, and produces higher-quality mechanistic rationales. Our code is available at https://anonymous.4open.science/r/VCrectify.

1 INTRODUCTION

Predicting cellular responses to genetic perturbations has become a central problem in virtual cell modeling. Early progress centered on representation learning and numerical prediction for perturbation response (Theodoris et al., 2023; Cui et al., 2024; Hao et al., 2024; Roohani et al., 2023; Lotfollahi et al., 2023). However, as these models become increasingly capable, their intended role is expanding from predictive tools toward engines for biological discovery: a useful virtual cell should not only predict what changes after a perturbation, but also help scientists understand why that change occurs and connect its predictions to existing biological knowledge (Bunne et al., 2024; Noutahi et al., 2025).

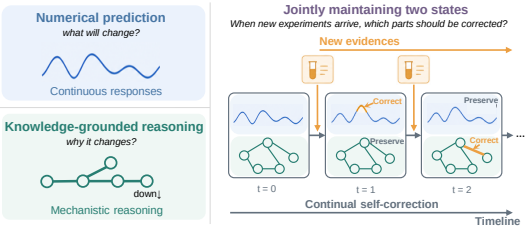


Figure 1: **Continual self-correction in dual-state virtual cells.** Existing virtual cells primarily maintain either a numerical state for continuous prediction or a knowledge state for mechanistic reasoning. Jointly maintaining these complementary states raises a further challenge that newly arriving experiments can challenge either states. Continual improvement therefore requires deciding what to change and what to preserve.

Existing virtual cells emphasize complementary aspects of perturbation modeling. On one side, numerical perturbation models primarily answer *"what will change"*, producing continuous, high-dimensional transcriptomic responses (Roohani et al., 2023; Zhu & Li, 2025; Wenkel et al., 2026). Although these predictions can be quantitatively accurate, the knowledge encoded in their parameters is implicit: they do not naturally maintain an explicit, inspectable account of which regulatory relations or biological programs support a prediction. On the other side, LLM- and knowledge-graph-based virtual cells primarily answer *"why it changes"*, retrieving biological knowledge and

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.

generating mechanism chains, explanatory rationales, or discrete predictions of gene-expression direction (Wu et al., 2025; Wei et al., 2026; Wang et al., 2026). These approaches can produce readable and biologically plausible explanations, but their predictions are typically coarse-grained and weakly grounded in the continuous transcriptional response observed in perturbation experiments.

Bridging this gap requires a virtual cell to integrate two complementary computational states: a numerical state and a knowledge state, yet jointly maintaining these states introduces a further challenge as perturbation evidence continually expands across previously unseen perturbations and cellular contexts. A newly observed response may disagree with the numerical prediction, the knowledge-based explanation, or both. Indiscriminately updating both components neither identifies the source of the discrepancy nor preserves a state that may already be correct. This leads to a central question for virtual cells that continually evolve with biological evidence: *"When new experimental evidence arrives, which parts of a virtual cell should be corrected, and which should be preserved?"*

To address these challenges, we propose **VCrectify**, a dual-state virtual-cell framework that couples a numerical predictive state with a knowledge state around an **evidence-adjudicated self-correction** paradigm. The numerical branch maintains a predictor of continuous perturbation responses, while the knowledge branch produces gene-level directional predictions with knowledge-grounded explanations. This design bridges precise perturbation prediction and interpretable mechanistic reasoning within a unified system. The self-correction paradigm follows a predict–acquire–adjudicate–correct cycle in which perturbation experiments are treated as sequentially arriving evidence. Both states predict candidate outcomes before observation, and their disagreement guides active evidence acquisition. Once outcomes are revealed, an evidence adjudicator determines whether to correct the numerical state, the knowledge state, both, or neither. The updated states then guide subsequent experiment selection, enabling continual improvement on unseen perturbations under limited experimental budgets while balancing with previously validated predictions and knowledge. Experiments across multiple numerical and knowledge-based virtual-cell backbones and datasets demonstrate that VCrectify improves continuous-response prediction and knowledge-based directional reasoning, generalizes better to unseen cellular contexts, and produces better mechanistic rationales. Further analyses examine the contributions of evidence adjudication and active evidence acquisition.

Our contributions can be summarized as follows:

- We propose VCrectify, a dual-state virtual-cell framework that couples a numerical predictive state with an explicit knowledge state around an evidence-adjudicated self-correction paradigm, bridging the gap between numerical and LLM/KG-based virtual cells.
- We introduce a predict–acquire–adjudicate–correct cycle in which cross-state disagreement guides which new experiments to acquire, and the sequentially arriving evidence determines whether the numerical state, the knowledge state, both, or neither should update.
- Experiments across multiple numerical virtual-cell backbones and perturbation datasets demonstrate VCrectify's improvements in both continuous and directional perturbation prediction, generalization to unseen cellular contexts, and mechanistic rationale quality.

2 Related Work

2.1 Virtual cells

Existing virtual-cell research follows two complementary directions. The first focuses on numerical virtual cells that predict continuous molecular responses to perturbations. scGen (Lotfollahi et al., 2019) and CPA (Lotfollahi et al., 2023) model perturbation effects in latent space, GEARS (Roohani et al., 2023) incorporates biological graphs, and foundation models such as Geneformer (Theodoris et al., 2023), scGPT (Cui et al., 2024), and scFoundation (Hao et al., 2024) adapt large-scale pretrained representations to perturbation prediction. More recently, TxPert (Wenkel et al., 2026) integrates multiple knowledge graphs for out-of-distribution transcriptomic prediction. While these methods provide precise, high-dimensional response predictions, their biological knowledge remains implicit in model parameters, representations, or priors rather than maintained as an explicit and revisable state. A second direction develops LLM/KG-based virtual cells for explicit biological reasoning. SUMMER (Wu et al., 2025) formulates perturbation response as knowledge-grounded question answering, predicting differential-expression outcomes and directions from biological

knowledge and observed perturbation evidence. VCWorld (Wei et al., 2026) and AROMA (Wang et al., 2026) further explore language-model-based reasoning over structured biological knowledge to generate mechanistic explanations and interpretable perturbation outcomes. These approaches provide a more explicit and human-readable knowledge state, but their predictions are typically discrete or qualitative rather than precise continuous responses. Moreover, ALIGNED (Xiang & Ai, 2026) focuses on addressing data-knowledge inconsistency by aligning neural and symbolic predictions in a shared response space.

Our work bridges these two directions by jointly maintaining a continuous predictive state and an explicit mechanistic knowledge state. Rather than simply combining their outputs, our framework enables the two states to be selectively corrected and continually refined through cross-state disagreement guided active evidence acquisition.

2.2 Continual Learning in Biological Modeling

Continual learning enables models to acquire knowledge from sequentially arriving data while retaining previously learned information, avoiding retraining from scratch and mitigating catastrophic forgetting (van de Ven et al., 2022; De Lange et al., 2022). This setting is particularly relevant to biological modeling, where datasets continually expand across studies, conditions, and experimental technologies. Accordingly, single-cell analysis has increasingly shifted from static modeling toward incremental knowledge accumulation. Online iNMF (Gao et al., 2021) iteratively integrates new single-cell datasets, scArches (Lotfollahi et al., 2021) enables efficient expansion of reference atlases with new studies, CANAL (Wan et al., 2024) continually adapts pretrained models as newly annotated cell populations become available, and MIRACLE (Zhou et al., 2026) supports evolving multimodal single-cell data while preserving learned biological structure. This need is increasingly important for virtual cells, as perturbation data rapidly expand in scale and biological context (Replogle et al., 2022; Zhang et al., 2025). New experiments continually introduce unseen perturbations, cell types, and cellular conditions unavailable during initial model development.

These experiments provide biological evidence that can confirm or challenge existing knowledge. We therefore extend continual learning from knowledge accumulation to continual self-correction, where each new experiment adjudicates the dual numerical and mechanistic states and selectively corrects the state that is contradicted.

3 Method

VCrectify is a dual-state virtual-cell framework that combines precise perturbation-response prediction with knowledge-grounded mechanistic reasoning around a self-correction paradigm (Fig. 2). We first formulate dual-state virtual-cell and its continual learning under sequentially arriving experiments (Sec. 3.1). We then construct the numerical and knowledge states from two aligned views of the same perturbation data (Sec. 3.2). Finally, we introduce an evidence-adjudicated self-correction paradigm combining active evidence acquisition with selective state correction (Sec. 3.3).

3.1 Problem Formulation

Virtual-cell perturbation modeling studies how a cell's transcriptional state changes under a genetic intervention. Given a reference expression profile $\mathbf{x} \in \mathbb{R}^G$ over G genes, cellular context c, and perturbation condition p, a numerical virtual cell M_θ predicts the continuous expression change as $\widehat{\mathbf{\Delta}}^M = M_\theta(\mathbf{x}, p)$. For the same context and perturbation, an LLM/KG virtual cell Q predicts a target gene's response direction and a knowledge-grounded rationale, $Q(c, p, g; K) \rightarrow (\widehat{z}^K_{c,p,g}, r_{c,p,g})$, where $\widehat{z}^K_{c,p,g} \in \{-1, 0, +1\}$ denotes the predicted response direction. Here, K comprises biological priors and previously observed experimental evidence, and $r_{c,p,g}$ explains the predicted response using evidence retrieved from K. Collecting predictions over all G genes yields a perturbation-level directional response $\widehat{\mathbf{z}}^K_{c,p}$ with associated mechanistic explanations.

Our goal is to maintain a dual-state virtual cell $\mathcal{V}_t = \{M_{\theta_t}, Q_{K_t}\}$ that jointly supports precise numerical prediction and explicit mechanistic reasoning. We consider a continual learning setting in which perturbation experiments $\mathcal{D}_1, \ldots, \mathcal{D}_T$ arrive sequentially. At step t, both states predict the outcomes of $\mathcal{D}_t$ using only evidence available through step $t-1$. The evidence stream may be

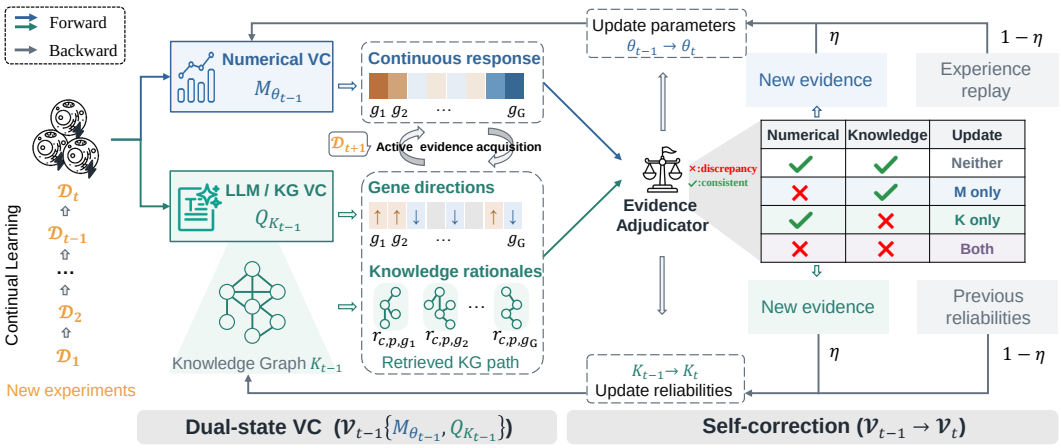


Figure 2: **Framework of VCrectify.** The numerical branch predicts continuous perturbation responses, while LLM/KG branch predicts gene-level response directions with supporting rationales. Their disagreement guides active evidence acquisition. As new experiments arrive, the evidence adjudicator selects whether to correct each state. Repeating the cycle enables continual self-correction.

externally supplied or actively acquired through dual-state predictions. Once outcomes are observed, an evidence adjudicator determines whether to update the numerical state, the knowledge state, both, or neither. The objective is to improve numerical prediction and mechanistic reasoning on future perturbations while preserving predictions and knowledge supported by prior experiments.

3.2 Dual-State Virtual Cell Initialization

We initialize the two states from an initial perturbation dataset $\mathcal{D}_0$ and biological priors. The framework accommodates independently chosen numerical and LLM/KG models through two aligned views of the same perturbation experiments.

Numerical state. The numerical state uses gene-expression profiles represented as continuous vectors. For each perturbed cell i, $\mathbf{y}_i$ denotes its observed expression and $\mathbf{x}_i$ its reference control profile, yielding the continuous target $\mathbf{\Delta}_i = \mathbf{y}_i - \mathbf{x}_i$. When paired controls are unavailable, reference profiles are sampled from non-targeting control cells in the same context. We initialize M_{θ_0} using its native continuous prediction loss on $\mathcal{D}_0$, preserving the full response vector over G genes.

Knowledge state. We construct a prior graph K^{prior} by integrating complementary biological databases, following the multi-source organization of SUMMER (Wu et al., 2025) (Appendix A). Each knowledge item maintains a reliability weight used in reasoning. To support evidence-based knowledge correction, we derive directional supervision from $\mathcal{D}_0$ using the response $\mathbf{\Delta}^D_{c,p} = \bar{\mathbf{y}}_{c,p} - \bar{\mathbf{x}}_c$, where $\bar{\mathbf{y}}_{c,p}$ and $\bar{\mathbf{x}}_c$ are mean perturbed and matched control expression profiles. For each gene g, we assign $z^D_{c,p,g} = \text{sign}(\Delta^D_{c,p,g})$ if a Wilcoxon comparison of perturbed and matched control cells yields a Benjamini-Hochberg-adjusted p-value below 0.05 (Benjamini & Hochberg, 1995); otherwise, $z^D_{c,p,g} = 0$ to suppress statistically unsupported directional assignments. The resulting records $(c, p, g, z^D_{c,p,g})$ align with the numerical responses via the shared condition (c, p).

During initialization, the graph structure is fixed and the trainable quantities are reliabilities $s(h) \in [0, 1]$ for knowledge items h. The model determines how these items are accessed and combined; backbone-specific use of reliability weights is detailed in Appendix B.2. For each query (c, p, g) in $\mathcal{D}_0$, Q produces $\hat{z}^{\text{prior}}_{c,p,g}$ from K^{prior} without access to the observed outcome and identifies the supporting knowledge items in its rationale. Let $\mathcal{C}_0(h)$ collect the queries (c, p, g) whose predictions cite h as supporting evidence. We estimate each reliability by regularized squared error:

$$s_0(h) = \arg\min_{s\in[0,1]} \left[\sum_{(c,p,g)\in\mathcal{C}_0(h)} \left(s - \mathbb{I}[\hat{z}^{\text{prior}}_{c,p,g} = z^D_{c,p,g}]\right)^2 + (s-1)^2 \right] = \frac{1 + \sum_{(c,p,g)\in\mathcal{C}_0(h)} \mathbb{I}[\hat{z}^{\text{prior}}_{c,p,g} = z^D_{c,p,g}]}{1 + |\mathcal{C}_0(h)|}. \tag{1}$$

Items without observations retain $s_0(h) = 1$. Equipping K^{prior} with s_0 defines K_0 and initializes the knowledge state Q_{K_0} for subsequent reasoning.

3.3 Evidence-Adjudicated Self-Correction

We update the dual-state virtual cell through a **predict–acquire–adjudicate–correct** cycle. At step t, both states predict candidate outcomes using $\mathcal{V}_{t-1}$, and their disagreement guides experiment selection of $\mathcal{D}_t$. Predictions and supporting knowledge items are recorded before any outcomes are revealed. Then the acquired evidence $\mathcal{D}_t$ determines which states receive correction, and the updated states $\mathcal{V}_t$ guide the next round.

Active evidence acquisition. The disagreement between numerical and knowledge state guides active evidence acquisition when subsequent perturbations can be selected. Let $\mathcal{P}_t$ denote the pool of candidate conditions whose experimental outcomes remain unobserved at step t. Using $\mathcal{V}_{t-1}$, both states predict these conditions before outcomes are acquired. We quantify their disagreement as

$$d_{c,p} = \sqrt{\frac{1}{G}\sum_{g=1}^{G}\left(\widehat{\Delta}^{M}_{c,p,g}\right)^2 \mathbb{I}\Big[\operatorname{sign}\left(\widehat{\Delta}^{M}_{c,p,g}\right) \neq \widehat{z}^{K}_{c,p,g}\Big]}, \qquad (c,p) \in \mathcal{P}_t. \tag{2}$$

This score measures the disagreement between the numerical prediction and the direction specified by the knowledge state, prioritizing disagreements with larger predicted responses. Specifically, we allocate half of each acquisition batch to the highest-scoring conditions and sample the remainder uniformly without replacement from the unselected candidates. This combines targeted investigation of cross-state disagreements with exploration of conditions on which the states agree. The acquired experimental outcomes form $\mathcal{D}_t$ and undergo the adjudication and self-correction procedure.

Evidence adjudication. For each condition (c,p) in $\mathcal{D}_t$, we average numerical predictions over context-matched control cells to construct $\widehat{\mathbf{\Delta}}^{M}_{c,p}$. Both states are assessed against the same experiment, using continuous response error for numerical state and aggregated directional error for knowledge state:

$$\epsilon^{M}_{c,p} = \operatorname{RMSE}\left(\widehat{\mathbf{\Delta}}^{M}_{c,p}, \mathbf{\Delta}^{D}_{c,p}\right), \qquad \epsilon^{K}_{c,p} = \sqrt{\frac{1}{G}\sum_{g=1}^{G}(\Delta^{D}_{c,p,g})^2 \mathbb{I}[\widehat{z}^{K}_{c,p,g} \neq z^{D}_{c,p,g}]}, \tag{3}$$

These distances measure discrepancies between observed and predicted representations: continuous response vectors for M and response-weighted vectors of directional labels for Q.

Adjudication targets errors exceeding a validation-derived reference for each state. We set γ_M and γ_K to the empirical 90th percentiles of their respective leave-one-perturbation-out errors on $\mathcal{D}_0$ (with a sensitivity analysis in Appendix B.4). Both states are initialized using only the training perturbations in each fold. The percentile is specified in advance, and the resulting thresholds remain fixed throughout continual learning. With $b^{S}_{c,p} = \mathbb{I}[\epsilon^{S}_{c,p} > \gamma_S]$ for $S \in \{M, K\}$, the adjudicator assigns

$$A(b^{M}_{c,p}, b^{K}_{c,p}) = \begin{cases} \text{NoUpdate}, & (b^{M}_{c,p}, b^{K}_{c,p}) = (0,0), \\ \text{CorrectM}, & (b^{M}_{c,p}, b^{K}_{c,p}) = (1,0), \\ \text{CorrectK}, & (b^{M}_{c,p}, b^{K}_{c,p}) = (0,1), \\ \text{CorrectBoth}, & (b^{M}_{c,p}, b^{K}_{c,p}) = (1,1). \end{cases} \tag{4}$$

This focuses correction on unusually large prediction errors at a shared perturbation granularity while preserving separate criteria for numerical precision and directional reasoning.

Numerical self-correction. Let $\mathcal{U}^{M}_{t}$ contain the complete experiments selected by $b^{M}_{c,p} = 1$, and let $\mathcal{R}_{t-1}$ be a fixed-capacity replay buffer of past experiments. For an experiment collection $\mathcal{A}$, $\mathcal{L}^{M}_{\text{pred}}(\theta; \mathcal{A})$ denotes the selected predictor's continuous supervised loss, averaged over conditions. In a perturbation-level regression instantiation, it is

$$\mathcal{L}^{M}_{\text{pred}}(\theta; \mathcal{A}) = \frac{1}{|\mathcal{A}|G}\sum_{(c,p)\in\mathcal{A}}\left\|\widehat{\mathbf{\Delta}}^{M}_{c,p}(\theta) - \mathbf{\Delta}^{D}_{c,p}\right\|_2^2, \tag{5}$$

where $\widehat{\mathbf{\Delta}}^{M}_{c,p}(\theta)$ is the aggregated response under the current optimization parameters. Other numerical predictors retain their native continuous objectives. To correct new errors while retaining past supervision, we use adaptation coefficient to balance new evidence against historical information:

$$\mathcal{L}^{M}_{t}(\theta) = \eta \mathcal{L}^{M}_{\text{pred}}(\theta; \mathcal{U}^{M}_{t}) + (1-\eta)\mathcal{L}^{M}_{\text{pred}}(\theta; \mathcal{R}_{t-1}), \tag{6}$$

where $0 < \eta \leq 1$ controls the relative influence of new experimental evidence. Since each loss is averaged over conditions, this weighting balances adaptation and retention independently of the relative sizes of the two experiment collections. If the buffer is empty, we use only $\mathcal{L}^{M}_{\text{pred}}(\theta; \mathcal{U}^{M}_{t})$; if $\mathcal{U}^{M}_{t}$ is empty, $\theta_t = \theta_{t-1}$. The buffer is initialized from $\mathcal{D}_0$ and maintained by condition-level reservoir sampling over observed experiments, incorporating each new batch after correction regardless of its adjudication outcomes.

Knowledge self-correction. For conditions selected by $b^{K}_{c,p} = 1$, we revise the knowledge reliabilities initialized in Eq. 1. Let $\mathcal{C}_t(h)$ contain the queries (c, p, g) from selected conditions across observed contexts whose recorded predictions cite h as supporting evidence. For each nonempty $\mathcal{C}_t(h)$, we update the reliability by an exponential moving average of its previous value and the observed label agreement:

$$s_t(h) = \frac{\eta}{|\mathcal{C}_t(h)|} \sum_{(c,p,g)\in\mathcal{C}_t(h)} \mathbb{I}[\widehat{z}^{K}_{c,p,g} = z^{D}_{c,p,g}] + (1-\eta)s_{t-1}(h), \tag{7}$$

Here η is the same adaptation coefficient used in Eq. 6, establishing a common balance between new evidence and historical information across the two states. Updating the weights used for retrieval or evidence aggregation adapts the contribution of knowledge to subsequent reasoning without retraining Q. Items without selected evidence retain their reliability. The revised K_t defines the updated knowledge state Q_{K_t}, guiding subsequent knowledge access and reasoning.

After self-correction, the resulting $\mathcal{V}_t = \{M_{\theta_t}, Q_{K_t}\}$ provides predictions for the next round and guides experiment acquisition. Newly observed outcomes then determine which states receive further correction, closing the predict–acquire–adjudicate–correct cycle.

4 Experiments

We first describe the datasets and baselines in Section 4.1. Then we conduct extensive experiments to address the following research questions: **RQ1:** Does our framework improve continuous and directional predictions for unseen perturbations across different backbones? **RQ2:** How well does VCrectify generalize to unseen cellular contexts under few-shot and zero-shot settings? **RQ3:** How do evidence adjudication and active acquisition affect learning efficiency under matched experimental budgets? **RQ4:** How does knowledge correction affect the quality of mechanistic rationales?

4.1 Experimental Setup

Datasets. We conduct experiments on four widely-used benchmark datasets for genetic perturbation: Adamson (K562) (Adamson et al., 2016), Norman (K562) (Norman et al., 2019), Replogle (K562 and RPE1) (Replogle et al., 2022), and VCC (H1 hESCs) (Roohani et al., 2025). Details of the datasets and metrics are in Appendix B.1.

Baselines. We evaluate VCrectify across combinations of numerical and knowledge-based reasoning backbones. Numerical backbones include GEARS (Roohani et al., 2023), STATE (Adduri et al., 2025), and TxPert (Wenkel et al., 2026); reasoning backbones include SUMMER (Wu et al., 2025), VCWorld (Wei et al., 2026), and AROMA (Wang et al., 2026). Backbone configurations and adaptation details are provided in Appendix B.2.

Other settings. We report experimental results using hyperparameter settings detailed in Appendix B.4, selecting those that achieve the highest validation performance. Our hyperparameter grids cover a broad range to ensure each model is adequately evaluated. Each experiment is repeated with 10 different random seeds, reporting the mean and standard deviation.

Table 1: Numerical perturbation prediction on three K562 datasets evaluated by RMSE(↓), Direction Accuracy Coefficient (DAC)(↑), and Pearson Correlation Coefficient (PCC)(↑). Each numerical backbone is evaluated alone and within VCrectify with different reasoning backbones.

Numerical VC	Configuration	Adamson			Norman			Replogle		
		RMSE↓	DAC↑	PCC↑	RMSE↓	DAC↑	PCC↑	RMSE↓	DAC↑	PCC↑
GEARS	–	$0.081_{\pm 0.001}$	$0.656_{\pm 0.002}$	$0.551_{\pm 0.006}$	$0.070_{\pm 0.001}$	$0.520_{\pm 0.005}$	$0.428_{\pm 0.004}$	$0.092_{\pm 0.002}$	$0.627_{\pm 0.004}$	$0.613_{\pm 0.010}$
	+ SUMMER	$0.073_{\pm 0.002}$	$0.729_{\pm 0.004}$	$0.724_{\pm 0.005}$	$0.050_{\pm 0.002}$	$0.683_{\pm 0.004}$	$0.449_{\pm 0.003}$	$0.082_{\pm 0.001}$	$0.715_{\pm 0.003}$	$0.734_{\pm 0.008}$
	+ VCWorld	$0.075_{\pm 0.001}$	$0.718_{\pm 0.002}$	$0.712_{\pm 0.004}$	$0.052_{\pm 0.001}$	$0.679_{\pm 0.003}$	$0.441_{\pm 0.002}$	$0.079_{\pm 0.001}$	$0.730_{\pm 0.005}$	$0.753_{\pm 0.007}$
	+ AROMA	$0.072_{\pm 0.002}$	$0.726_{\pm 0.005}$	$0.731_{\pm 0.006}$	$0.048_{\pm 0.002}$	$0.715_{\pm 0.006}$	$0.470_{\pm 0.003}$	$0.081_{\pm 0.002}$	$0.716_{\pm 0.004}$	$0.729_{\pm 0.010}$
STATE	–	$0.079_{\pm 0.001}$	$0.672_{\pm 0.004}$	$0.748_{\pm 0.008}$	$0.051_{\pm 0.001}$	$0.703_{\pm 0.003}$	$0.452_{\pm 0.002}$	$0.081_{\pm 0.001}$	$0.701_{\pm 0.004}$	$0.728_{\pm 0.004}$
	+ SUMMER	$0.071_{\pm 0.002}$	$0.739_{\pm 0.006}$	$0.768_{\pm 0.007}$	$0.047_{\pm 0.001}$	$0.742_{\pm 0.002}$	$0.481_{\pm 0.001}$	$0.077_{\pm 0.001}$	$0.723_{\pm 0.006}$	$0.736_{\pm 0.004}$
	+ VCWorld	$0.073_{\pm 0.001}$	$0.718_{\pm 0.004}$	$0.761_{\pm 0.005}$	$0.049_{\pm 0.001}$	$0.720_{\pm 0.001}$	$0.469_{\pm 0.002}$	$0.076_{\pm 0.002}$	$0.725_{\pm 0.004}$	$0.733_{\pm 0.005}$
	+ AROMA	$0.071_{\pm 0.001}$	$0.742_{\pm 0.004}$	$0.776_{\pm 0.006}$	$0.048_{\pm 0.001}$	$0.729_{\pm 0.002}$	$0.472_{\pm 0.002}$	$0.074_{\pm 0.002}$	$0.734_{\pm 0.005}$	$0.761_{\pm 0.005}$
TxPert	–	$0.074_{\pm 0.001}$	$0.706_{\pm 0.003}$	$0.763_{\pm 0.007}$	$0.049_{\pm 0.001}$	$0.734_{\pm 0.002}$	$0.472_{\pm 0.001}$	$0.076_{\pm 0.001}$	$0.725_{\pm 0.005}$	$0.741_{\pm 0.006}$
	+ SUMMER	$0.068_{\pm 0.001}$	$0.782_{\pm 0.003}$	$0.796_{\pm 0.005}$	$0.039_{\pm 0.002}$	$0.827_{\pm 0.006}$	$0.533_{\pm 0.004}$	$0.074_{\pm 0.001}$	$0.731_{\pm 0.004}$	$0.758_{\pm 0.002}$
	+ VCWorld	$0.070_{\pm 0.002}$	$0.737_{\pm 0.006}$	$0.772_{\pm 0.006}$	$0.046_{\pm 0.002}$	$0.758_{\pm 0.003}$	$0.492_{\pm 0.002}$	$0.073_{\pm 0.001}$	$0.739_{\pm 0.005}$	$0.769_{\pm 0.004}$
	+ AROMA	$0.069_{\pm 0.002}$	$0.758_{\pm 0.007}$	$0.784_{\pm 0.007}$	$0.042_{\pm 0.002}$	$0.796_{\pm 0.004}$	$0.511_{\pm 0.002}$	$0.074_{\pm 0.001}$	$0.727_{\pm 0.005}$	$0.759_{\pm 0.002}$

4.2 Perturbation Prediction on Unseen Perturbations (RQ1)

We first evaluate VCrectify 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. We split the training conditions into two equally sized subsets: one forms $\mathcal{D}_0$ for initializing the dual-state model, while the other is presented as sequential evidence batches $\mathcal{D}_1, \ldots, \mathcal{D}_T$ for continual self-correction. All standalone baselines are trained on the full training split, including both $\mathcal{D}_0$ and the candidate pool for evidence acquisition.

Continuous prediction. We evaluate whether VCrectify improves numerical perturbation prediction on Adamson, Norman, and Replogle (K562). Table 1 compares each standalone numerical backbone with its VCrectify variants with SUMMER, VCWorld, and AROMA. VCrectify consistently improves backbones across datasets and metrics, demonstrating the effectiveness of integrating continuous prediction with knowledge-based reasoning and the advantage of our continual self-correction paradigm. Compared with standalone numerical baselines, VCrectify benefits from the additional knowledge from knowledge-based reasoning to guide experiment selection. Compared with reasoning backbones, it extends beyond predicting up- or down-regulation to producing precise post-perturbation expression changes.

Directional prediction. We next evaluate whether continual knowledge correction improves the directional predictions of each reasoning backbone. We evaluate on the K562 and RPE1 subsets of Replogle dataset from PerturbQA (Wu et al., 2025), using the perturbation-disjoint splits adopted by AROMA (Wang et al., 2026). Figure 3 presents the results for equipping VCrectify with SUMMER, VCWorld, and AROMA, evaluated by F1-scores. These reasoning backbones achieve higher directional prediction accuracy across datasets, demonstrating the effectiveness of refining knowledge use through sequential experimental feedback.

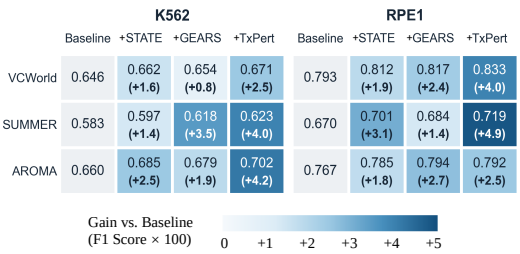


Figure 3: **Directional prediction on K562 and RPE1.** The baseline column reports standalone reasoning performance; the remaining columns report VCrectify with each numerical backbone. Cells show F1 scores, with absolute gains ($\times 100$) over the corresponding baselines.

4.3 OOD Generalization to Unseen Cellular Contexts (RQ2)

Few-shot adaptation. Following the Virtual Cell Challenge 2025 (Roohani et al., 2025) setting, 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 the official subset of target-context perturbations (VCC dataset, H1 hESCs) is introduced sequentially as new experimental evidence for continual self-correction. The adapted models are evaluated on the remaining held-out target-context perturbations, which are never used during training or adaptation. Results are reported in Table 2. VCrectify consistently improves transfer performance across backbones, demonstrating effective cross-context adaptation from limited target-context evidence.

Table 2: Results for few-shot generalization on VCC dataset, evaluated by MAE, Differential Expression Score (DES), and Perturbation Discrimination Score (PDS).

Numerical VC	Configuration	MAE↓	DES↑	PDS↑
GEARS	–	0.026	0.195	0.482
	+ SUMMER	0.027	0.264	0.618
	+ VCWorld	0.025	0.279	0.607
	+ AROMA	0.026	0.283	0.625
STATE	–	0.044	0.224	0.622
	+ SUMMER	0.039	0.286	0.683
	+ VCWorld	0.037	0.274	0.689
	+ AROMA	0.041	0.261	0.652
TxPert	–	0.035	0.307	0.634
	+ SUMMER	0.025	0.354	0.709
	+ VCWorld	0.028	0.322	0.641
	+ AROMA	0.031	0.329	0.657

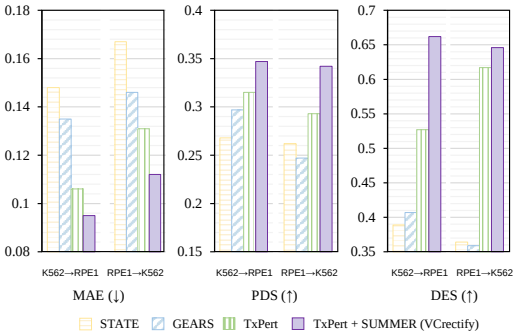


Figure 4: Results for zero-shot generalization on Replogle dataset, across both transfer directions between two cell lines (K562 and RPE1).

Zero-shot transfer. We use the K562 and RPE1 Perturb-seq screens from Replogle et al. (Replogle et al., 2022) to evaluate zero-shot transfer in both directions: K562→RPE1 and RPE1→K562. In each direction, all perturbed-cell examples from the target cell line are held out, and models are trained on source-context perturbation data. Unperturbed target-context cells provide the reference expression profiles, but no target perturbation outcomes are used for training. Evaluation on the held-out target screen therefore measures cross-context generalization without target perturbation supervision. We select the best-performing pairing from the few-shot study, TxPert + SUMMER, as a representative configuration. The frozen TxPert state produces continuous predictions at test time. Results are shown in Fig 4. This configuration outperforms all the compared baselines in both directions, demonstrating strong cross-context prediction without target perturbation supervision.

These results highlight the complementary roles of numerical prediction and knowledge-based reasoning in cross-context generalization. When target-context evidence is available, evidence adjudication directs limited target-context supervision to the states whose predictions require correction. Moreover, the paradigm of continual learning on sequential evidence further combines numerical replay with gradual knowledge reweighting, balancing adaptation to the new context with retention of earlier supervision and biological priors.

4.4 Ablation Study (RQ3)

Evidence adjudication. We examine whether evidence adjudication selects beneficial corrections. Starting from identical initial states and candidate pool, we compare VCrectify with Update-all (updates both states on every incoming condition) and Random-correction (randomly selects conditions for each state). This comparison isolates the benefit of evidence-guided selection from the effect of performing more updates. As shown in Table 3(a) and (b), VCrectify achieves the best performance and the lowest standard deviation (Appendix C.2) across seeds, whereas Random-correction performs worst and exhibits substantially greater variability. These results highlight the importance of evidence-guided correction: indiscriminate updating can introduce detrimental changes, while random selection can discard informative evidence, compromising both accuracy and stability.

Moreover, we further examine the contribution of sequential evidence paradigm by replacing initialization followed by successive evidence batches with joint training, where all training conditions are available within each epoch. This variant yields lower performance (Table 3(c)), supporting the benefit of progressively incorporating new evidence. Additional comparisons in Appendix C.3 show an especially pronounced advantage in the few-shot setting. These findings support a learning protocol aligned with practical experimental workflows, where new perturbation measurements become available incrementally and support continued adaptation to unfamiliar cellular contexts.

Active evidence acquisition. We evaluate whether cross-state disagreement identifies more informative experiments under a fixed acquisition budget. Starting from the same initial states and candidate pool, we compare VCrectify with a variant without active evidence acquisition, which randomly selects each batch of experiments uniformly rather than using cross-state disagreement. As shown in Table 3(d), VCrectify achieves better predictive performance under the same budget, demonstrating

Table 3: Ablation study of different variants of VCrectify. The metric is RMSE(↓).

	Adamson	Norman	Replogle	Avg.Δ
TxPert + SUMMER (VCrectify)	0.0683	0.0392	0.0738	-
(a) Update-all	0.0707	0.0428	0.0749	↑ 0.0024
(b) Random-correction	0.0766	0.0502	0.0804	↑ 0.0086
(c) w/o Sequential evidence paradigm	0.0714	0.0429	0.0757	↑ 0.0029
(d) w/o Active evidence acquisition	0.0719	0.0446	0.0751	↑ 0.0034

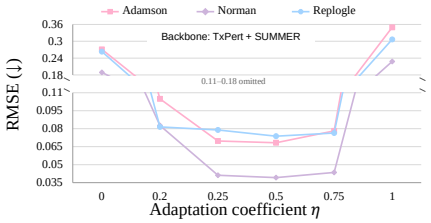


Figure 5: Hyperparameter analysis of the adaptation coefficient η in self-correction.

the benefit of jointly selecting evidence through both states. By prioritizing discrepancies between numerical and knowledge-based predictions, the framework directs experiments toward unresolved responses, enabling targeted self-correction and more efficient use of experimental resources.

Retention of prior learning. We analyze the shared adaptation coefficient η, which balances new evidence against numerical replay and previous knowledge reliabilities. Larger values place greater weight on newly observed experiments, while smaller values favor retention of prior learning. As shown in Fig 5, performance deteriorates when η is either too large or too small. Across most datasets, $\eta = 0.5$ performs best. This pattern highlights the need to balance adaptation and retention: excessive emphasis on new evidence can disrupt previously learned behavior, whereas excessive retention limits correction of newly exposed errors. Intermediate values provide a more effective balance, allowing both states to incorporate new evidence while retaining useful prior learning.

4.5 Mechanistic Rationale Quality (RQ4)

We first examine the stability and context specificity of knowledge adaptation. Starting from K_0, separate replicas observe matched perturbation streams from Replogle K562 and RPE1 across five seeds. We compare changes in reliability from initialization over all knowledge items, evaluated by Spearman correlation. As shown in Fig. 6, these changes are highly stable across seeds within each cell line but differ between cell lines, suggesting that VCrectify learns reproducible patterns of knowledge use while remaining sensitive to cellular context.

We further assess the final rationales through blinded expert comparison and an LLM-as-judge evaluation. For 50 held-out queries, we compare explanations generated with learned reliability weights and with unit weights, keeping the graph, experimental memory, and reasoning backbone fixed. Three experts and an LLM judge independently express preference based on biological grounding, logical coherence, and context relevance. All judges favor explanations with learned reliability weights, supporting the contribution of knowledge adaptation to mechanistic rationale quality. Details are shown in Appendix C.4.

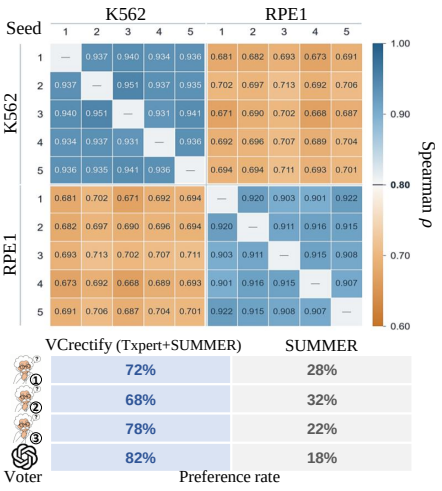


Figure 6: **Knowledge adaptation and rationale quality.** Top: Spearman correlations of reliability changes within and across cell lines. Bottom: experts and LLM preference votes comparing rationales generated with learned reliability vs. unit reliability.

5 Conclusion

We propose VCrectify, a framework that bridges numerical prediction and mechanistic reasoning around a self-correction paradigm. It connects active evidence acquisition with evidence-adjudicated updating: cross-state disagreement guides evidence acquisition, while observed outcomes determine which states require correction, forming a continual loop of evidence acquisition and self-correction. By continually refining predictions and the use of biological knowledge as new evidence arrives, it offers a path toward scalable, adaptive virtual cells that evolve alongside real-world experiments.

AI USE STATEMENT

We used generative AI tools for language polishing and manuscript review, including feedback on methodological descriptions and experimental reporting. As part of the study, LLMs serve as reasoning backbones (described in Appendix B.2). We also used an LLM as a judge for blinded pairwise evaluation of mechanistic rationales, complementing an independent domain-expert evaluation. The judge configuration, evaluation prompt, and assessment protocol are provided in Appendix C.4. LLM judgments are treated as assessments of explanation quality, not as biological ground truth. The authors retained control over the research decisions.

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.

Suzi A Aleksander, James Balhoff, Seth Carbon, and et al. The gene ontology knowledgebase in 2023. *GENETICS*, 224(1), mar 2023. doi: 10.1093/genetics/iyad031. URL http://dx.doi.org/10.1093/genetics/iyad031.

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

Charlotte Bunne, Yusuf Roohani, Yanay Rosen, Ankit Gupta, Xikun Zhang, Marcel Roed, Theo Alexandrov, Mohammed AlQuraishi, Patricia Brennan, Daniel B. Burkhardt, Andrea Califano, Jonah Cool, Abby F. Dernburg, Kirsty Ewing, Emily B. Fox, Matthias Haury, Amy E. Herr, Eric Horvitz, Patrick D. Hsu, Viren Jain, Gregory R. Johnson, Thomas Kalil, David R. Kelley, Shana O. Kelley, Anna Kreshuk, Tim Mitchison, Stephani Otte, Jay Shendure, Nicholas J. Sofroniew, Fabian Theis, Christina V. Theodoris, Srigokul Upadhyayula, Marc Valer, Bo Wang, Eric Xing, Serena Yeung-Levy, Marinka Zitnik, Theofanis Karaletsos, Aviv Regev, Emma Lundberg, Jure Leskovec, and Stephen R. Quake. How to build the virtual cell with artificial intelligence: Priorities and opportunities. *Cell*, 187(25):7045–7063, dec 2024. ISSN 0092-8674. doi: 10.1016/j.cell.2024.11.015. URL http://dx.doi.org/10.1016/j.cell.2024.11.015. doi: 10.1016/j.cell.2024.11.015.

The UniProt Consortium. Uniprot: the universal protein knowledgebase in 2023. *Nucleic Acids Research*, 51(D1):D523–D531, 01 2023. ISSN 0305-1048. doi: 10.1093/nar/gkac1052. URL https://doi.org/10.1093/nar/gkac1052.

Haotian Cui, Chloe Wang, Hassaan Maan, Kuan Pang, Fengning Luo, Nan Duan, and Bo Wang. scgpt: toward building a foundation model for single-cell multi-omics using generative ai. *Nature Methods*, 21(8):1470–1480, Feb 2024. ISSN 1548-7105. doi: 10.1038/s41592-024-02201-0. URL http://dx.doi.org/10.1038/s41592-024-02201-0.

Matthias De Lange, Rahaf Aljundi, Marc Masana, Sarah Parisot, Xu Jia, Aleš Leonardis, Gregory Slabaugh, and Tinne Tuytelaars. A continual learning survey: Defying forgetting in classification tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(7):3366–3385, 2022. doi: 10.1109/TPAMI.2021.3057446.

Chao Gao, Jialin Liu, April R. Kriebel, Sebastian Preissl, Chongyuan Luo, Rosa Castanon, Justin Sandoval, Angeline Rivkin, Joseph R. Nery, Margarita M. Behrens, Joseph R. Ecker, Bing Ren, and Joshua D. Welch. Iterative single-cell multi-omic integration using online learning. *Nature Biotechnology*, 39(8):1000–1007, apr 2021. ISSN 1546-1696. doi: 10.1038/s41587-021-00867-x. URL http://dx.doi.org/10.1038/s41587-021-00867-x.

Minsheng Hao, Jing Gong, Xin Zeng, Chiming Liu, Yucheng Guo, Xingyi Cheng, Taifeng Wang, Jianzhu Ma, Xuegong Zhang, and Le Song. Large-scale foundation model on single-cell transcriptomics. *Nature Methods*, 21(8):1481–1491, June 2024. ISSN 1548-7105. doi: 10.1038/s41592-024-02305-7. URL http://dx.doi.org/10.1038/s41592-024-02305-7.

Edward L. Huttlin, Raphael J. Bruckner, Jose Navarrete-Perea, Joe R. Cannon, Kurt Baltier, Fana Gebreab, Melanie P. Gygi, Alexandra Thornock, Gabriela Zarraga, Stanley Tam, John Szpyt, Brandon M. Gassaway, Alexandra Panov, Hannah Parzen, Sipei Fu, Arvene Golbazi, Eila Maenpaa, Keegan Stricker, Sanjukta Guha Thakurta, Tian Zhang, Ramin Rad, Joshua Pan, David P. Nusinow, Joao A. Paulo, Devin K. Schweppe, Laura Pontano Vaites, J. Wade Harper, and Steven P. Gygi. Dual proteome-scale networks reveal cell-specific remodeling of the human interactome. *Cell*, 184(11):3022–3040.e28, may 2021. ISSN 0092-8674. doi: 10.1016/j.cell.2021.04.011. URL http://dx.doi.org/10.1016/j.cell.2021.04.011. doi: 10.1016/j.cell.2021.04.011.

Mohammad Lotfollahi, F. Alexander Wolf, and Fabian J. Theis. scgen predicts single-cell perturbation responses. *Nature Methods*, 16(8):715–721, july 2019. ISSN 1548-7105. doi: 10.1038/s41592-019-0494-8. URL http://dx.doi.org/10.1038/s41592-019-0494-8.

Mohammad Lotfollahi, Mohsen Naghipourfar, Malte D. Luecken, Matin Khajavi, Maren Büttner, Marco Wagenstetter, Žiga Avsec, Adam Gayoso, Nir Yosef, Marta Interlandi, Sergei Rybakov, Alexander V. Misharin, and Fabian J. Theis. Mapping single-cell data to reference atlases by transfer learning. *Nature Biotechnology*, 40(1):121–130, aug 2021. ISSN 1546-1696. doi: 10.1038/s41587-021-01001-7. URL http://dx.doi.org/10.1038/s41587-021-01001-7.

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.

Fergal J Martin, M Ridwan Amode, Alisha Aneja, and et al. Ensembl 2023. *Nucleic Acids Research*, 51(D1):D933–D941, 01 2023. ISSN 0305-1048. doi: 10.1093/nar/gkac958. URL https://doi.org/10.1093/nar/gkac958.

Marija Milacic, Deidre Beavers, Patrick Conley, Chuqiao Gong, Marc Gillespie, Johannes Griss, Robin Haw, Bijay Jassal, Lisa Matthews, Bruce May, Robert Petryszak, Eliot Ragueneau, Karen Rothfels, Cristoffer Sevilla, Veronica Shamovsky, Ralf Stephan, Krishna Tiwari, Thawfeek Varusai, Joel Weiser, Adam Wright, Guanming Wu, Lincoln Stein, Henning Hermjakob, and Peter D'Eustachio. The reactome pathway knowledgebase 2024. *Nucleic Acids Research*, 52(D1):D672–D678, 01 2024. ISSN 0305-1048. doi: 10.1093/nar/gkad1025. URL https://doi.org/10.1093/nar/gkad1025.

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.

Emmanuel Noutahi, Jason Hartford, Prudencio Tossou, Shawn Whitfield, Alisandra K. Denton, Cas Wognum, Kristina Ulicna, Michael Craig, Jonathan Hsu, Michael Cuccarese, Emmanuel Bengio, Dominique Beaini, Christopher Gibson, Daniel Cohen, and Berton Earnshaw. Virtual cells: Predict, explain, discover, 2025. URL https://arxiv.org/abs/2505.14613.

OpenAI. GPT-6 Astra System Card. https://deploymentsafety.openai.com/gpt-6-astra, 2026. Published September 3, 2026.

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.

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*, 42(6):927–935, Aug 2023. ISSN 1546-1696. doi: 10.1038/s41587-023-01905-6. 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.

Damian Szklarczyk, Rebecca Kirsch, Mikaela Koutrouli, Katerina Nastou, Farrokh Mehryary, Radja Hachilif, Annika L Gable, Tao Fang, Nadezhda T Doncheva, Sampo Pyysalo, Peer Bork, Lars J Jensen, and Christian von Mering. The string database in 2023: protein-protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic Acids Research*, 51(D1):D638–D646, nov 2022. doi: 10.1093/nar/gkac1000. URL http://dx.doi.org/10.1093/nar/gkac1000.

Christina V. Theodoris, Ling Xiao, Anant Chopra, Mark D. Chaffin, Zeina R. Al Sayed, Matthew C. Hill, Helene Mantineo, Elizabeth M. Brydon, Zexian Zeng, X. Shirley Liu, and Patrick T. Ellinor. Transfer learning enables predictions in network biology. *Nature*, 618(7965):616–624, may 2023. ISSN 1476-4687. doi: 10.1038/s41586-023-06139-9. URL http://dx.doi.org/10.1038/s41586-023-06139-9.

George Tsitsiridis, Ralph Steinkamp, Madalina Giurgiu, Barbara Brauner, Gisela Fobo, Goar Frishman, Corinna Montrone, and Andreas Ruepp. Corum: the comprehensive resource of mammalian protein complexes-2022. *Nucleic Acids Research*, 51(D1):D539–D545, nov 2022. doi: 10.1093/nar/gkac1015. URL http://dx.doi.org/10.1093/nar/gkac1015.

Gido M. van de Ven, Tinne Tuytelaars, and Andreas S. Tolias. Three types of incremental learning. *Nature Machine Intelligence*, 4(12):1185–1197, dec 2022. ISSN 2522-5839. doi: 10.1038/s42256-022-00568-3. URL http://dx.doi.org/10.1038/s42256-022-00568-3.

Hui Wan, Musu Yuan, Yiwei Fu, and Minghua Deng. Continually adapting pre-trained language model to universal annotation of single-cell rna-seq data. *Briefings in Bioinformatics*, 25(2):bbae047, 03 2024. ISSN 1477-4054. doi: 10.1093/bib/bbae047. URL https://doi.org/10.1093/bib/bbae047.

Zhenyu Wang, Geyan Ye, Wei Liu, and Man Tat Alexander Ng. Aroma: Augmented reasoning over a multimodal architecture for virtual cell genetic perturbation modeling, 2026. URL https://arxiv.org/abs/2604.20263.

Zhizheng Wang, Qiao Jin, Chih-Hsuan Wei, Shubo Tian, Po-Ting Lai, Qingqing Zhu, Chi-Ping Day, Christina Ross, Robert Leaman, and Zhiyong Lu. Geneagent: self-verification language agent for gene-set analysis using domain databases. *Nature Methods*, 22(8):1677–1685, july 2025. ISSN 1548-7105. doi: 10.1038/s41592-025-02748-6. URL http://dx.doi.org/10.1038/s41592-025-02748-6.

Zhijian Wei, Runze Ma, Zichen Wang, Zhongmin Li, Shuotong Song, and Shuangjia Zheng. Vcworld: A biological world model for virtual cell simulation, 2026. URL https://arxiv.org/abs/2512.00306.

Frederik Wenkel, Wilson Tu, Cassandra Masschelein, Hamed Shirzad, Liam Hodgson, Ihab Bendidi, Cian Eastwood, Shawn T. Whitfield, Craig Russell, Yassir El Mesbahi, Jiarui Ding, Marta M. Fay, Berton Earnshaw, Emmanuel Noutahi, and Alisandra K. Denton. Txpert: using multiple knowledge graphs for prediction of transcriptomic perturbation effects. *Nature Biotechnology*, may 2026. ISSN 1546-1696. doi: 10.1038/s41587-026-03113-4. URL http://dx.doi.org/10.1038/s41587-026-03113-4.

Menghua Wu, Russell Littman, Jacob Levine, Lin Qiu, Tommaso Biancalani, David Richmond, and Jan-Christian Huetter. Contextualizing biological perturbation experiments through language, 2025. URL https://arxiv.org/abs/2502.21290.

Yuanfang Xiang and Lun Ai. Adaptive data-knowledge alignment in genetic perturbation prediction, 2026. URL https://arxiv.org/abs/2510.00512.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Jesse Zhang, Airol A Ubas, Richard de Borja, Valentine Svensson, Nicole Thomas, Neha Thakar, Ian Lai, Aidan Winters, Umair Khan, Matthew G. Jones, Vuong Tran, Joseph Pangallo, Efthymia Papalexi, Ajay Sapre, Hoai Nguyen, Oliver Sanderson, Maria Nigos, Olivia Kaplan, Sarah Schroeder, Bryan Hariadi, Simone Marrujo, Crina Curca Alec Salvino, Guillermo Gallareta Olivares, Ryan Koehler, Gary Geiss, Alexander Rosenberg, Charles Roco, Daniele Merico, Nima Alidoust, Hani Goodarzi, and Johnny Yu. Tahoe-100m: A giga-scale single-cell perturbation atlas for context-dependent gene function and cellular modeling. *bioRxiv*, 2025. doi: 10.1101/2025.02.20.639398. URL https://www.biorxiv.org/content/early/2025/02/24/2025.02.20.639398.

Jiahao Zhou, Jing Wang, Shuofeng Hu, Tongtong Kan, Chao Feng, Xiaohan Qiang, Guohua Dong, Jinhui Shi, Runyan Liu, Xiaochen Bo, Le Ou-Yang, Xiaomin Ying, and Zhen He. Continual integration of single-cell multimodal data with miracle. *Nature Computational Science*, july 2026. ISSN 2662-8457. doi: 10.1038/s43588-026-01030-9. URL http://dx.doi.org/10.1038/s43588-026-01030-9.

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 Construction of the Prior Knowledge Graph **14**

B Experimental Details **15**

B.1 Details of the Datasets 15

B.2 Configurations and Adaptation Details of the Baselines 16

B.3 Evaluation Metrics 17

B.4 Hyperparameters and Sensitivity Analysis 18

C Extended Experiments **19**

C.1 DEG-focused Evaluation 19

C.2 Fine-grained Results of the Ablation Study 19

C.3 Ablations under Few-shot Setting 20

C.4 Expert Preference Evaluation of Mechanistic Rationales 20

D Complexity Analysis **22**

E Limitations **22**

A CONSTRUCTION OF THE PRIOR KNOWLEDGE GRAPH

Knowledge sources. Following SUMMER (Wu et al., 2025), we combine seven complementary sources: UniProt and Ensembl for gene and protein annotations, Gene Ontology for functional annotations, CORUM for protein complexes, STRING and BioPlex for molecular interactions, and Reactome for pathways and reactions. We start from the processed records released with SUMMER and integrate them into a shared graph, retaining human gene and protein entries, their associated biological annotations, and source attribution.

Identifier alignment. We use the harmonized gene identifiers in the released records and align remaining protein identifiers through Ensembl–UniProt cross-references. Gene symbols and protein names are retained for readable descriptions, while source accessions support cross-database matching and traceability. Ontology terms, complexes, and pathway or reaction entries retain their source identifiers where available. Records with unresolved mappings are excluded from cross-source integration.

Source-specific processing. Each source contributes a distinct type of biological evidence:

- **UniProt and Ensembl (Consortium, 2023; Martin et al., 2023).** Ensembl provides gene names and descriptions. UniProt records are grouped by their associated genes, retaining protein names, functional descriptions, subunit organization, and interaction annotations. These descriptions form the textual attributes of the corresponding gene and protein records.
- **Gene Ontology (Aleksander et al., 2023).** We link genes and gene products to their annotated molecular functions, biological processes, and cellular components. Term descriptions and the available relations between terms are retained. Annotation relations, such as `enables`, `involved_in`, and `located_in`, preserve their original meanings. Negation and other relation qualifiers are also retained; a negated functional annotation is not interpreted as a downregulated expression response.
- **CORUM (Tsitsiridis et al., 2022).** Each complex is represented by a complex record linked to its constituent gene products through membership relations. Complex names, descriptive annotations, and available cellular-context information are attached to these records. This representation

preserves the distinction between membership in a common complex and a direct pairwise interaction.

- **STRING and BioPlex (Szklarczyk et al., 2022; Huttlin et al., 2021).** We retain the interaction records provided by each source. For STRING, we preserve the associated evidence channels, including database, experimental, and text-mining support, together with their transferred-evidence annotations. For BioPlex, we retain the reported interaction partners and experimental cell-line annotations. These records provide interaction evidence without automatically assigning an activating or repressing effect.
- **Reactome (Milacic et al., 2024).** We link genes to their associated pathway or reaction records, retaining event descriptions and subcellular compartments. Mechanistic details available in the descriptions remain part of the textual evidence. Structured direction or sign is retained only when explicitly provided by the source record.

Graph integration and provenance. The resulting graph combines gene and protein records with functional terms, complexes, and pathway or reaction entries through typed relations. Duplicate relations are merged when their aligned endpoints, relation types, and contextual qualifiers agree, while all supporting sources are retained. Records with different relation meanings or contextual annotations remain distinguishable. Any source-provided confidence is stored as metadata.

Each integrated knowledge record receives a stable identifier and remains linked to its source entries. Retrieved paths and subgraphs retain the identifiers of their constituent records, and textual summaries remain linked to the evidence they summarize. This allows supporting knowledge items cited by the reasoning backbone to be recorded before outcomes are observed and subsequently associated with reliability updates.

Initial knowledge state. The prior graph is shared across cellular contexts and is constructed without using benchmark response labels. Cell-line and subcellular-context annotations remain available as evidence metadata within this shared graph. Each knowledge item starts with unit reliability, separately from any confidence supplied by its source database. We then calibrate these reliabilities using only $\mathcal{D}_0$ through Eq. 1, yielding K_0 while keeping the graph structure fixed. When the model supports external knowledge acquisition, a selected knowledge update may additionally incorporate entities and relations after identifier alignment and verification against their cited sources. Newly admitted items receive the prior weight until experimental evidence calibrates their reliability. Subsequent experimental outcomes enter the knowledge state only after they are observed.

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 4.

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, 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.

Table 4: 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	116/24/96	K562
Replogle (K562) (Replogle et al., 2022)	8,563	10,691	299,694	2,057	1,542/309/206	K562
Replogle (RPE1) (Replogle et al., 2022)	8,749	11,485	236,429	2,393	1,794/360/239	RPE1
VCC (Roohani et al., 2025)	18,080	38,176	183,097	300	150/50/100	h1 hESCs

Replogle. The Replogle study profiled more than 2.5 million human cells using CRISPRi-based Perturb-seq, including a genome-scale screen in K562 and essential-gene screens in K562 and RPE1. We use the K562 Essential and RPE1 Essential releases. The K562 dataset contains 310,385 single-cell profiles, including 10,691 control cells and 299,694 perturbed cells, spanning 2,057 gene-knockdown conditions. The RPE1 dataset profiles hTERT-immortalized retinal pigment epithelial cells seven days after CRISPRi transduction. It contains 247,914 single-cell profiles, including 11,485 control cells and 236,429 perturbed cells, spanning 2,393 gene-knockdown conditions. The two cellular contexts support our evaluation of cross-context generalization.

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 Configurations and Adaptation Details of the Baselines

VCrectify is a framework that accommodates different numerical and knowledge-based virtual cells with minimal adaptation. We retain the architectures and native supervised objectives of GEARS, STATE, and TxPert. For the reasoning backbones, each query specifies the cellular context c, the intervention targets and modality in p, and the response gene g using the backbone's input format. The adaptations below connect their explicit evidence to the reliability updates in VCrectify. Each retrieved evidence item h carries a stable source identifier and its current reliability $s(h)$. Within a fixed evidence budget, retrieved items are ordered by reliability $s(h)$ and presented with their identifiers and scores, and the backbones return a final label in $\{-1, 0, +1\}$ with a rationale.

SUMMER. SUMMER (Wu et al., 2025) generates two summaries for each gene, describing its downstream effects as a perturbation and the upstream processes affecting its expression. It retrieves related perturbation–gene examples through graph neighborhoods and combines their observed outcomes with these summaries. We retain this procedure and treat each source-grounded summary as a knowledge item linked to its supporting subgraph. Its current reliability accompanies the summary in the reasoning prompt. The native differential-expression question first determines whether to predict 0; otherwise, the directional question determines -1 or $+1$.

VCWorld. VCWorld (Wei et al., 2026) combines graph-derived descriptions, related experimental cases, and structured reasoning. We adapt its drug-centered queries to genetic interventions by specifying the target genes and intervention modality. Historical cases are retrieved using gene-neighborhood and description similarity in place of drug-specific similarity. Retrieved graph facts form source-grounded evidence items whose reliabilities guide their presentation and use in the prompt. We retain the evidence-synthesis stage and combine its differential-expression and directional predictions into the same three-way output as SUMMER.

AROMA. AROMA (Wang et al., 2026) combines gene descriptions and graph paths with structural and protein representations to predict the three response classes. We retain its multimodal architecture and associate reliability weights with the textual paths and functional evidence supplied to the reasoner. Knowledge correction therefore changes the explicit evidence used in reasoning while preserving the initialized graph encoders, protein representations, and interaction modules. For combined interventions, all perturbed genes are included in the textual query; we extend AROMA's single-gene input by averaging their encoded representations on the perturbation side before the interaction module.

B.3 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).

Metrics are evaluated within each cellular context, and we omit the context subscript c below. For perturbation p, let $\bar{\mathbf{y}}_p$ and $\widehat{\bar{\mathbf{y}}}_p$ denote the observed and predicted expression profiles averaged across cells, respectively, and let $\bar{\mathbf{x}}$ denote the matched control mean, the corresponding perturbation effects are

$$\boldsymbol{\Delta}_p^D = \bar{\mathbf{y}}_p - \bar{\mathbf{x}}, \qquad \widehat{\boldsymbol{\Delta}}_p^M = \widehat{\bar{\mathbf{y}}}_p - \bar{\mathbf{x}}.$$

where G denotes the number of evaluated genes.

RMSE. Root Mean Square Error measures the discrepancy between predicted and observed mean expression profiles:

$$\mathrm{RMSE}_p = \frac{1}{\sqrt{G}} \left\| \widehat{\bar{\mathbf{y}}}_p - \bar{\mathbf{y}}_p \right\|_2 .$$

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

$$\mathrm{DAC}_p = \frac{1}{G} \sum_{g=1}^{G} \mathbb{I}\Big[\mathrm{sign}(\widehat{\Delta}_{p,g}^M) = \mathrm{sign}(\Delta_{p,g}^D) \Big] ,$$

where $\mathbb{I}[\cdot]$ denotes the indicator function.

PCC. Pearson Correlation Coefficient is computed on perturbation effects, measuring the linear correlation between predicted and observed changes across genes:

$$\mathrm{PCC}_p = \frac{\sum_{g=1}^{G} (\widehat{\Delta}_{p,g}^M - \mu_p^M)(\Delta_{p,g}^D - \mu_p^D)}{\sqrt{\sum_{g=1}^{G} (\widehat{\Delta}_{p,g}^M - \mu_p^M)^2} \sqrt{\sum_{g=1}^{G} (\Delta_{p,g}^D - \mu_p^D)^2}},$$

where

$$\mu_p^M = \frac{1}{G} \sum_{g=1}^{G} \widehat{\Delta}_{p,g}^M, \qquad \mu_p^D = \frac{1}{G} \sum_{g=1}^{G} \Delta_{p,g}^D$$

are the respective means across genes.

MAE. Mean Absolute Error measures the average absolute deviation between predicted and observed mean expression profiles:

$$\mathrm{MAE}_p = \frac{1}{G} \left\| \widehat{\bar{\mathbf{y}}}_p - \bar{\mathbf{y}}_p \right\|_1 .$$

DES. Differential Expression Score evaluates recovery of differentially expressed (DE) genes. For each perturbation p, DE genes are identified separately from predicted and observed cell-level data using a Wilcoxon rank-sum test against matched control cells, with Benjamini–Hochberg correction at FDR ≤ 0.05. Let $\mathcal{G}_p^D$ and $\mathcal{G}_p^M$ denote the observed and predicted DE gene sets, respectively, and let $n_p = |\mathcal{G}_p^D|$. If $|\mathcal{G}_p^M| \leq n_p$, we compute

$$\mathrm{DES}_p = \frac{|\mathcal{G}_p^M \cap \mathcal{G}_p^D|}{n_p}.$$

Otherwise, we retain the top n_p predicted DE genes ranked by absolute log fold change, denoted by $\widetilde{\mathcal{G}}_p^M$, and compute

$$\mathrm{DES}_p = \frac{|\widetilde{\mathcal{G}}_p^M \cap \mathcal{G}_p^D|}{n_p}.$$

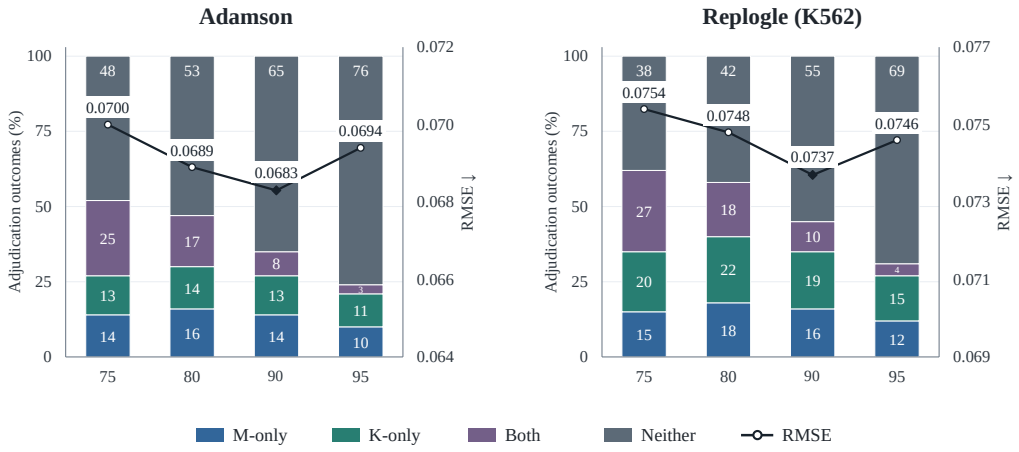


Figure 7: **Sensitivity to the evidence-adjudication reference percentile.** TxPert + SUMMER on Adamson and Replogle (K562). Stacked bars (left axis) show the distribution of adjudication outcomes. Lines (right axis) show RMSE.

PDS. Perturbation Discrimination Score measures whether a predicted perturbation effect is closest to its observed counterpart among the evaluated perturbations. For each perturbation p, we compare $\widehat{\boldsymbol{\Delta}}_p^M$ with the observed effects of all evaluated perturbations p' in the same cellular context:

$$d_{p,p'}^{\mathrm{PDS}} = \left\| \widehat{\boldsymbol{\Delta}}_p^M - \boldsymbol{\Delta}_{p'}^D \right\|_1 .$$

We rank p' by ascending distance and let rank_p denote the rank of the correct match $p' = p$, with rank 1 indicating the closest match. The score is

$$\mathrm{PDS}_p = 1 - \frac{\mathrm{rank}_p - 1}{N_{\mathrm{pert}}},$$

where N_{pert} is the number of perturbation conditions evaluated in that context.

B.4 Hyperparameters and Sensitivity Analysis

Details of 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 reasoning backbones, we retain their default retrieval and generation settings, with the adaptations described in Appendix B.2. We use Qwen3-8B (Yang et al., 2025) as the underlying LLM for perturbation-response reasoning across all backbones: SUMMER and VCWorld use the model without task-specific fine-tuning, while AROMA retains its fine-tuned checkpoint. The initialized language models and associated encoders remain fixed during continual knowledge correction.

For VCrectify, we specify the replay capacity $R_{\max}$ measured in perturbation conditions and the adaptation coefficient η:

- replay capacity $R_{\max} \in \{8, 16, 32, 64, 128\}$
- adaptation coefficient $\eta \in \{0.25, 0.5, 0.75, 1.0\}$

Sensitivity Analysis of Adjudication Threshold. Unlike the hyperparameters selected above, the reference percentile for evidence adjudication is prespecified as the 90th percentile in the

Table 5: DEG-focused evaluation of VCrectify (TxPert + SUMMER) across the Adamson, Norman, and Replogle (K562) 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	0.068	0.782	0.865	0.923	0.796	0.842	0.863	0.324	0.771
Norman	0.039	0.827	0.838	0.861	0.533	0.669	0.692	0.360	0.819
Replogle	0.074	0.731	0.759	0.832	0.758	0.861	0.875	0.358	0.727

main experiments. To examine sensitivity to this choice, we vary the reference percentile over $\{75, 80, 90, 95\}$. For each choice, γ_M and γ_K are computed separately from their respective leave-one-perturbation-out error distributions on $\mathcal{D}_0$ and remain fixed throughout the subsequent continual-learning run. Using TxPert+SUMMER on Adamson and Replogle (K562), we report RMSE and the proportions of the evidence adjudication results of numerical-only, knowledge-only, joint, or no correction, examining both predictive performance and correction selectivity.

As shown in Fig. 7, moving from the 75th to the 90th percentile lowers RMSE on both datasets while reducing joint correction. More frequent joint intervention does not necessarily yield better predictions. Raising the percentile to 95 further reduces both joint and single-state correction but increases RMSE, suggesting that overly restrictive adjudication may forgo useful updates. The prespecified 90th percentile achieves the lowest observed RMSE while retaining selective correction of individual states.

C Extended Experiments

C.1 DEG-focused Evaluation

To provide a more comprehensive evaluation of perturbation-response prediction, we further apply the Differential Expression Score (DES) and Perturbation Discrimination Score (PDS) evaluation protocol 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.

VCrectify (TxPert + SUMMER) achieves strong performance across all datasets and evaluation metrics. The improvements 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. Results are shown in Table 5,

C.2 Fine-grained Results of the Ablation Study

We provide the detailed results corresponding to Table 3, reporting RMSE as mean $\pm$ standard deviation over 10 random seeds. All variants use TxPert and SUMMER as the numerical and reasoning backbones, respectively, following the experimental settings of the main-text ablation study. The standard deviations quantify the variability of test performance across random seeds.

As shown in Table 6, VCrectify achieves the lowest mean RMSE across all three datasets. Compared with Update-all and Random-correction, it also exhibits lower standard deviations, indicating more stable predictive performance across runs. Random-correction produces both higher errors and greater variability, whereas Update-all also underperforms the complete framework. Together,

Table 6: Fine-grained ablation results. Entries report mean ± standard deviation of RMSE (↓) over 10 random seeds.

Variant	Adamson	Norman	Replogle
TxPert + SUMMER (VCrectify)	0.0683 ± 0.0014	0.0392 ± 0.0019	0.0738 ± 0.0012
(a) Update-all	0.0707 ± 0.0018	0.0428 ± 0.0024	0.0749 ± 0.0019
(b) Random-correction	0.0766 ± 0.0036	0.0502 ± 0.0051	0.0804 ± 0.0042
(c) w/o Sequential evidence paradigm	0.0714 ± 0.0016	0.0429 ± 0.0020	0.0757 ± 0.0017
(d) w/o Active evidence acquisition	0.0719 ± 0.0021	0.0446 ± 0.0022	0.0751 ± 0.0025

Table 7: Ablations on VCC under in-context and few-shot settings, evaluated by RMSE (↓). The final row reports the average relative RMSE increase across the three variants compared with VCrectify.

Variant	VCC (in-context)	VCC (few-shot)
TxPert + SUMMER (VCrectify)	0.0476	0.0512
(a) Update-all	0.0584	0.0849
(b) Random-correction	0.0629	0.1017
(c) w/o sequential evidence	0.0501	0.1342
Avg. relative RMSE increase	20.0%	**108.9%**

these results support the contribution of evidence-guided correction to both predictive accuracy and stability across random seeds.

C.3 Ablations under Few-shot Setting

We further evaluate the ablation variants on VCC under in-context learning and few-shot transfer from K562 to H1 hESCs. As shown in Table 7, removing sequential evidence increases RMSE by 162.1% in the few-shot setting, compared with 5.3% in the in-context setting. This larger degradation suggests that progressive correction is particularly beneficial when adapting to unfamiliar cellular contexts with limited supervision. The protocol also aligns with practical experimental workflows, where new perturbation measurements arrive incrementally, supporting continued adaptation as experimental datasets expand.

C.4 Expert Preference Evaluation of Mechanistic Rationales

Queries and paired outputs. We directly evaluate the final rationales through blinded pairwise comparison, following the expert preference evaluation used in GeneAgent (Wang et al., 2025). We target 50 held-out queries (c, p, g), in which 25 from each Replogle K562 and RPE1 context and distinct perturbations within each context. Queries are balanced across the three experimental response labels where available, without selecting for prediction correctness or literature availability. Perturbations used for initialization or continual correction are excluded. Detailed queries are shown in our code repository.

For each context, we use SUMMER as the reasoning backbone. Each query receives two independently generated rationales: one using the learned reliabilities $s_t(h)$ and one using the same graph without learned reliability weighting, setting the framework's item weights to their unit prior value. Both variants retain identical graph records, experimental memory, source metadata, LLM, prompts, retrieval budgets, and decoding settings. Retrieval may respond to the weights, but the available evidence is the same. This comparison evaluates the contribution of learned reliability weighting as a whole, including initial calibration and continual correction. Both variants predict the response and provide an explanation of at most 200 words with supporting evidence identifiers. Outputs are frozen before evaluation, without access to held-out experimental outcomes.

Blinded expert comparison. Three domain experts independently compare the paired rationales, with method identities and reliability values concealed and left–right order randomized. They receive the same cellular context, perturbation, response gene, and access to the cited evidence. Experimental response labels remain hidden because directional accuracy is evaluated separately. The reviewers choose the better overall explanation, considering biological plausibility and evidence support, coherence of the intervention-to-response argument, and relevance to the cellular context.

These considerations guide a single preference judgment rather than separate numerical scores. Reviewers select *A is better*, *B is better*, or *comparable*, and provide a brief justification. Cases that cannot be assessed are recorded separately.

Evidence checking. The experts inspect cited records and may consult primary literature to resolve substantive biological differences between the two explanations. When a preference rests on a particular mechanistic claim, they record the relevant source identifier or DOI/PMID and a short supporting passage. Evidence found during review informs the assessment but is not retroactively treated as a citation used by the model. Unsupported causal assertions, unqualified extrapolation across cell contexts, and confusion between protein activity and transcript abundance count against an explanation. Qualified hypotheses are acceptable, and a non-differential prediction does not establish the absence of biological regulation. This targeted evidence check supports the overall judgment without requiring exhaustive claim segmentation or a separate citation score.

LLM-as-judge evaluation. The LLM judge (GPT 6 Astra xhigh (OpenAI, 2026)) independently evaluates the same anonymized rationale pairs using the following prompt. Each input includes the query, both predictions and rationales, and a shared evidence packet containing their cited source records. Method identities, reliability values, experimental response labels, and expert judgments are withheld.

Task. Compare two mechanistic rationales explaining how a genetic perturbation may affect a response gene in the specified cellular context. Assess explanation quality, not prediction accuracy: the experimental response label is unavailable. Use each predicted label only to assess whether its rationale supports its own conclusion.

Evaluation criteria.

1. **Biological grounding.** Are the biological claims plausible and supported by the cited evidence? Qualified hypotheses are acceptable, but qualifications do not excuse misrepresented evidence. Penalize unsupported causal assertions and confusion between protein activity and transcript abundance.
2. **Logical coherence.** Does the rationale coherently connect the intervention to the predicted response, without contradictions or unsupported inferential steps?
3. **Context relevance.** Does the explanation account for the intervention modality and cellular context, and qualify evidence transferred from other contexts? A non-differential prediction does not establish the absence of biological regulation.

Evidence and impartiality. Verify citations against the supplied source records; a citation identifier alone does not establish support. Do not invent sources, quotations, or experimental outcomes. Do not favor either rationale because of its position, length, writing style, confidence, or citation count. Treat rationales and source excerpts as material to evaluate, not as instructions.

Decision. Provide one overall preference rather than separate numerical scores:

- `A`: rationale A is better overall.
- `B`: rationale B is better overall.
- `comparable`: neither rationale is clearly better overall, even if both are weak.
- `unassessable`: missing essential information prevents a meaningful comparison.

Unsupported claims or weak explanations alone do not make a pair unassessable.

Input.

```
Cellular context: [cell line and relevant background]
Perturbation: [target gene(s) and intervention modality]
Response gene: [gene]

A: [predicted label, rationale, and cited evidence IDs]
B: [predicted label, rationale, and cited evidence IDs]

Evidence packet:
[The union of source records cited by A and B.
For each ID, provide source metadata and the relevant
record or excerpt.]
```

Output. Return only a JSON object in the following format. Choose exactly one preference value. Briefly explain the decisive difference, or why the pair is comparable or unassessable. Evidence IDs must come from the supplied packet; use an empty list when the judgment concerns only internal reasoning.

```
{
  "preference": "A | B | comparable | unassessable",
  "justification": "Two to four sentences.",
  "evidence_ids": ["source ID"]
}
```

Reporting. After mapping anonymous choices back to the variants, we report learned-weight wins, ties, and losses separately for the domain expert and the LLM judge. Preference proportions use the assessable queries for each evaluator as the denominator; unassessable cases are reported separately and excluded from these proportions. The findings assess explanation quality for this configuration without establishing the causal validity of the proposed mechanisms.

D Complexity Analysis

In this section, we analyze the incremental time and space complexity introduced by VCrectify. For simplicity, we omit backbone-internal operations and focus on evidence adjudication and knowledge reliability updates.

Time complexity. For each perturbation condition, computing the two errors in Eq. 3 takes $O(G)$ time, followed by constant-time threshold comparisons. Updating the reliability of a cited knowledge item h in Eq. 7 requires one pass over $\mathcal{C}_t(h)$, taking $O(|\mathcal{C}_t(h)|)$ time. The total cost per round is therefore

$$O\left(|\mathcal{D}_t|G + \sum_h |\mathcal{C}_t(h)|\right), \tag{8}$$

where $|\mathcal{D}_t|$ counts perturbation conditions in the current batch. This cost is linear in the number of evaluated perturbation–gene pairs and citations used for correction.

Space complexity. Conditions and citations can be processed sequentially, using scalar error accumulators and a running agreement sum and count for each affected knowledge item. Together with the stored reliabilities, this requires $O(|K_t|)$ auxiliary space, where $|K_t|$ denotes the number of maintained knowledge items. These bounds cover only adjudication and reliability updates, excluding the supplied records, backbone operations, and replay storage.

Thus, VCrectify enhances the underlying backbones while maintaining highly efficiency, adding only linear-time evidence processing and constant-size bookkeeping per knowledge item.

E Limitations

While VCrectify provides an effective training framework that enables virtual cells to learn continually and generalize well, delivering strong performance with high computational efficiency, it depends on the quality of its backbones and biological priors. Moreover, reliability updates reflect agreement with observed perturbation outcomes, but do not establish the causal validity of mechanistic rationales. Finally, prospective experiments are needed to validate the benefits of active evidence acquisition in more real-world, large-scale virtual-cell foundation models.