Escaping Local Optima in the Waddington Landscape: A Multi-Stage TRPO–PPO Approach for Single-Cell Perturbation Analysis

Francis Boabang and Samuel Asante Gyamerah

***Abstract*—Modeling cellular responses to genetic and chemical perturbations remains a central challenge in single-cell biology. Existing data-driven framework have advanced perturbation prediction through variational autoencoders, chemically conditioned autoencoders, and large-scale transformer pretraining. However, these models are prone to local optima in the nonconvex Waddington landscape of cell fate decisions, where poor initialization can trap trajectories in spurious lineages or implausible differentiation outcomes. While executable gene regulatory networks complement these approaches, automated design frameworks incorporate biological priors through multi-agent optimization. Yet, an approach that is completely data-driven with well-designed initialization to escape local optima and converge to a proper lineage remains elusive. In this work, we introduce a multistage reinforcement learning algorithm tailored for single-cell perturbation modeling. We first compute an explicit natural gradient update using Fisher-vector products and a conjugate gradient solver, scaled by a KL trust-region constraint to provide a safe, curvature-aware the first step for the policy. Starting with these preconditioned parametrs, we then apply a second phase of proximal policy optimization (PPO) with clipped surrogates, exploiting minibatch efficiency to refine the policy. We demonstrate that this initialization substantially improves generalization on Single-cell RNA sequencing (scRNA-seq) and Single-cell ATAC sequencing (scATAC-seq) data. https://github.com/boabangf/GNN_RL_gene_trajectory_perturbation/tree/main/Multi-Step%20Differentiation%20Experiment**

***Index Terms*—Machine Learning, Waddington Landscape, Single Cell, Convex Optimization, Gene Regulation, Reinforcement Learning, Protein Interaction**

I. Introduction

Understanding how stem cells make fate decisions is a central question in developmental biology and regenerative medicine. Stem cell differentiation is orchestrated by dense networks of interacting transcription factors (TFs) forming gene regulatory networks (GRNs), which govern the timing and nature of cell state transitions [1], [2]. Insights derived from GRNs during differentiation have enabled more rational design of cell culture systems and have direct implications for cell therapy and regenerative applications [3], [4]. A key example is the use of TFs to reprogram embryonic or adult somatic cells into a pluripotent state, where specific driver TFs can induce a network configuration consistent with pluripotency [5]. These processes can be modelled as executable Boolean GRNs that capture both the network topology and the regulatory rules governing gene interactions, enabling simulation of time-evolving cell states [6]–[8].

Despite their utility, deriving informative, predictive, and executable GRNs remains challenging. Traditionally, constructing GRNs requires integrating evidence from gene perturbation experiments, a process that is labour-intensive, time-consuming, and costly [7]. Notable progress has been made through automated formal reasoning, which has successfully identified minimal GRNs underlying naive pluripotency in mice, accurately predicting outcomes for a substantial fraction of new experiments [9], [10]. However, such approaches have been limited by the scale of data and the reliance on low-throughput measurements.

The advent of single-cell profiling technologies has dramatically expanded the capacity to study cellular differentiation at unprecedented resolution. scRNA-seq, in particular, provides high-coverage, flexible, and accurate measurements of gene expression, enabling robust clustering and pseudotime inference [11]–[13]. These datasets facilitate the derivation of more precise and high-throughput GRNs. While formal reasoning approaches have successfully inferred executable GRNs from single-cell quantitative PCR (sc qPCR) data [14], leveraging scRNA-seq remains underdeveloped. scRNA-seq presents challenges, including dropout effects and reduced sensitivity for lowly expressed TFs, but offers significant advantages in data richness and flexibility. Despite these opportunities, the field lacks an integrated platform for inferring, simulating, and analyzing executable GRNs directly from single-cell transcriptomic data [1].

scATAC-seq and Cellular Indexing of Transcriptomes and Epitopes by sequencing (CITE-seq) are other powerful techniques that provide complementary insights into cellular states by profiling chromatin accessibility and surface protein expression, respectively. These methodologies enable high-resolution, multi-dimensional analyses of single cells, facilitating a deeper understanding of cellular heterogeneity and regulatory mechanisms.

scATAC-seq is an advanced technique that assesses chromatin accessibility by measuring the integration of Tn5 transposase into open chromatin regions within individual nuclei. This approach allows researchers to identify regulatory elements such as promoters and enhancers, and to infer transcription factor activity and cell-type-specific regulatory networks. The method has been instrumental in elucidating the epigenetic underpinnings of various biological processes, including cell

Francis Boabang is with the Concordia Institute for Information and Systems Engineering (CIISE), Concordia University, Montréal, QC, Canada.

Samuel Asante Gyamerah is with the Department of Mathematics, Toronto Metropolitan University, Toronto, Ontario, Canada. (Corresponding author: asante.gyamerah@torontomu.ca)

differentiation, disease progression, and response to environmental stimuli [15]–[17].

CITE-seq combines single-cell RNA sequencing with the measurement of surface protein expression by incorporating antibody-derived tags (ADTs) into the sequencing process. This integration allows for the simultaneous profiling of gene expression and protein levels in individual cells, providing a comprehensive view of cellular identity and function. The technique has been widely adopted in immunology, oncology, and developmental biology to dissect complex cellular populations and identify novel biomarkers [18]–[20].

The integration of scRNA-seq, scATAC-seq and scCITE-seq data offers a holistic view of cellular states by linking chromatin accessibility with gene expression and protein profiles. This multi-omics approach enables the identification of regulatory elements that control gene expression and elucidates how these elements influence cellular phenotypes. Computational tools like Seurat and Harmony can facilitate integration across modalities, enabling identification of cell types and states [16], [18].

RNA-seq, scATAC-seq and CITE-seq are transformative technologies that, when used together, offer a powerful platform for dissecting the complexities of cellular regulation. Their integration facilitates a multi-dimensional analysis of single cells, providing deeper insights into the molecular basis of health and disease. As these technologies continue to evolve, they hold the promise of uncovering novel therapeutic targets and advancing personalized medicine strategies.

These challenges are further compounded when incorporating perturbation data. Genetic knockouts, drug treatments, or cytokine stimulations induce highly context-dependent transcriptional responses that can be subtle, nonlinear, and influenced by cell-type–specific regulatory architectures. Capturing such complexity demands models that can move beyond simple statistical associations to uncover causal regulatory mechanisms [21].

While machine learning models, particularly deep learning, offer powerful tools for high-dimensional representation learning, their application to regulatory network modeling remains nontrivial. Nonconvex optimization landscapes, limited interpretability, and difficulties in generalizing to unseen perturbations or novel cell types often hinder their ability to outperform simpler, more interpretable baselines. Thus, the task of extracting meaningful insights from single-cell data requires approaches that marry the expressive capacity of machine learning with the mechanistic grounding of regulatory network theory, while carefully accounting for data heterogeneity across modalities.

The growing interest in reinforcement learning (RL) within computational biology stems from its ability to model sequential decision-making processes under uncertainty, a paradigm highly relevant for single-cell perturbation analysis [21]. Unlike supervised learning approaches that passively map inputs to outputs, RL agents learn by actively interacting with an environment, making decisions that maximize long-term rewards. This framework naturally aligns with the challenges of predicting cellular responses to perturbations, where interventions such as gene knockouts, drug treatments, or cytokine stimulations can be viewed as actions applied to a dynamic and high-dimensional cellular state space.

In the context of single-cell biology, RL offers several unique advantages. First, it provides a principled mechanism for exploring perturbation spaces that are combinatorial and intractable with brute-force experimentation. Instead of exhaustively testing all possible single and combinatorial perturbations, RL agents can learn efficient exploration strategies to identify interventions most likely to induce desired cell fate transitions. Second, RL is inherently suited for modeling trajectories, making it a natural tool for approximating the Waddington landscape [22] and capturing lineage-specific differentiation dynamics as shown in 1. Third, by framing perturbation analysis as a feedback-driven process, RL facilitates adaptive learning, where agents refine predictions as new single-cell data and perturbation experiments become available [21].

The scAgents framework leverages these advantages by embedding RL into an autonomous multi-agent system for end-to-end single-cell perturbation analysis [21]. Here, multiple agents can specialize in tasks such as modality integration (e.g., scRNA-seq, scATAC-seq, CITE-seq), regulatory network inference, or intervention policy optimization, while coordinating through shared objectives. This multi-agent perspective enables distributed reasoning across heterogeneous data types and perturbation mechanisms, addressing the fragmentation that limits current approaches. As a result, RL within scAgents not only provides a computational strategy for predicting perturbation effects but also opens the door to closed-loop experimental design, where models actively guide which perturbations should be tested next, accelerating discovery in cell biology. To further enhance the capabilities of ScAgents, there is the need to improve the existing policy optimization techniques in reinforcement learning to help model escape local optima and converge to a good lineage or cell fate. Cell reprogramming can be viewed as an optimization trajectory escaping a local optimum, driven by external perturbations (i.e., chemical, genetic, or environmental) that reshape the underlying potential landscape of gene regulation. Within the framework of the Waddington epigenetic landscape, each cell fate represents a local attractor basin stabilized by transcriptional and epigenetic feedback loops. Transitions between these basins require mechanisms that promote exploration and controlled deviation from stability, analogous to optimization strategies in machine learning that prevent premature convergence.

Policy optimization in deep reinforcement learning ranges from first-order, minibatch algorithms such as Proximal Policy Optimization to curvature-aware trust-region methods such as Trust Region Policy Optimization (TRPO). PPO stabilizes policy updates using a clipped (or penalized) surrogate and multiple minibatch passes, giving strong empirical performance with simple implementation, but the proximal behaviour is heuristic. TRPO instead enforces an average-KL trust region and computes a natural-gradient step via Fisher-vector products and a conjugate gradient solver, producing principled curvature correction at increased per-update cost. The trust-region updates ensure stable yet effective movement across curvature-sensitive regions of the landscape, entropy regularization pro-

motes diversity and exploration of alternative cell fates, and exploration-driven updates simulate stochastic fluctuations or biological noise that can trigger fate transitions. Together, these mechanisms form a reinforcement learning–inspired model of cell reprogramming, where policy optimization corresponds to navigating and reshaping the epigenetic landscape to reach new, biologically meaningful attractor states.

Prior work has sought to combine the advantages of both approaches. Wang *et al.* propose Trust Region–Guided PPO (TRGPPO), which adapts PPO's clipping range using KL-based trust-region analysis to tighten guarantees while preserving the PPO workflow [23]. Lascu *et al.* provide a Fisher-Rao geometric reinterpretation of PPO (FR-PPO), deriving geometry-aware surrogates that align PPO updates with Fisher/KL structure; FR-PPO is primarily a theoretical reformulation rather than an explicit natural-gradient solver [24]. Other lines of work explore first-order preconditioning to approximate curvature cheaply [25], or alter the PPO surrogate to induce greater conservatism (e.g., COPG) [26].

Our TRPO-preconditioned PPO multistage is explicitly sequential: we compute a natural-gradient direction using exact Fisher structure (FVP/CG), scale it to satisfy a KL budget (TRPO style), accept a safe major step, and then initialize a short PPO clipped fine-tune from this anchor. This differs from FR-PPO (no CG/FVP major step), from first-order preconditioning (we use the exact Fisher curvature rather than an approximate preconditioner), and from objective-shaping approaches (we retain a TRPO trust-region guarantee for the major update while leveraging PPO's minibatch fine-tuning for additional local improvements). Initializing PPO from a TRPO-preconditioned step provides several practical and theoretical benefits. The Fisher-vector product and conjugate gradient step in TRPO computes an approximate natural gradient, making the initial update curvature-aware and aligned with the local geometry of the policy manifold [27]. Furthermore, the KL-divergence constraint enforced during the TRPO step ensures a trust-region guarantee, preventing overly aggressive updates that could destabilize the policy [28]. As a result, subsequent PPO fine-tuning becomes more sample-efficient, requiring fewer minibatch updates to achieve performance gains, while reducing oscillations and redundant gradient steps for faster convergence. Finally, this initialization balances exploration and stability: the TRPO step preserves policy safety, while PPO's clipped surrogate refines the policy locally to exploit the available trajectories effectively [24], [28]. Overall, the multistage approach leverages the strengths of both TRPO and PPO, combining principled, curvature-aware updates with efficient, practical fine-tuning.

In a nutshell, the contributions of the paper are as follows:

1) we compute an explicit natural gradient step using Fisher-vector products and a conjugate gradient solver, scaled to satisfy a KL trust-region constraint, providing a safe, curvature-aware major-step for the policy.
2) Starting from the preconditioned parameters, we perform a short PPO clipped surrogate fine-tuning phase, exploiting minibatch efficiency to refine the policy locally without violating the trust region. Unlike previous approaches that either approximate curvature or modify the surrogate geometry (e.g., FR-PPO [24], first-order preconditioning [25], COPG [26]), our method explicitly combines exact TRPO preconditioning with PPO's sample-efficient optimization, achieving a best-of-both-worlds algorithm.
3) Finally, we demonstrate that this initialization improves sample efficiency, stability, and convergence speed, while balancing exploration and exploitation in Single Cell perturbation analysis.

II. Related work

Machine learning and regulatory network has improved prediction accuracy and scalability across modalities (scRNA-seq, scATAC-seq, CITE-seq) but can act as black boxes that do not directly expose causal regulatory logic.

The first paper focuses on the inference and simulation of *executable* gene regulatory network (GRN). Platforms such as IQCELL infer logic-based GRNs from pseudo-time ordered scRNA-seq, construct compact Boolean or logic models constrained by mutual information and temporal ordering, and then simulate developmental trajectories to predict qualitative effects of gene knockouts and perturbations [1]. IQCELL emphasizes interpretability and causal hypotheses: the inferred GRNs can recapitulate many experimentally validated causal interactions and enable in-silico perturbation experiments that help understand developmental fate decisions. This logic/GRN approach excels at mechanistic insight but can be limited in modelling dose-dependence, continuous dynamics, and very high-dimensional multi-omic inputs. The drawback of this paper is that since Boolean models define discrete attractor states, cells may converge to artificial "stable states" that correspond to bad differentiation fates.

One strand emphasizes *data-driven* machine learning models that treat perturbation prediction as a conditional mapping from pre-perturbation cell state and perturbation descriptor to post-perturbation expression. Early simple approaches used linear models and tree-based regressors, while more recent methods leverage deep generative models, conditional GANs, and transformer-style architectures to model high-dimensional, heterogeneous single-cell modalities and generalize to unseen perturbations [30]–[32]. A landmark in data-driven approaches is [30], who introduced *scGen*, a variational autoencoder framework that predicts how single cells respond to perturbations such as drug treatments or gene knockouts. scGen learns a latent representation of cells and models perturbations as vector arithmetic in latent space, allowing generalization to unseen perturbations and cell types. This work demonstrated that generative neural networks could accurately extrapolate perturbation responses in scRNA-seq, setting the stage for subsequent machine learning approaches.

Then, [31] proposed *chemCPA*, a conditional perturbation autoencoder that integrates molecular structure information with single-cell gene expression data to predict transcriptional responses to small-molecule perturbations. By incorporating chemical embeddings and adopting an adversarial training strategy, chemCPA not only generalized to unseen compounds but also provided insights into drug mechanism of action. This highlighted the potential of integrating domain knowledge such as molecular chemistry into single-cell perturbation models.

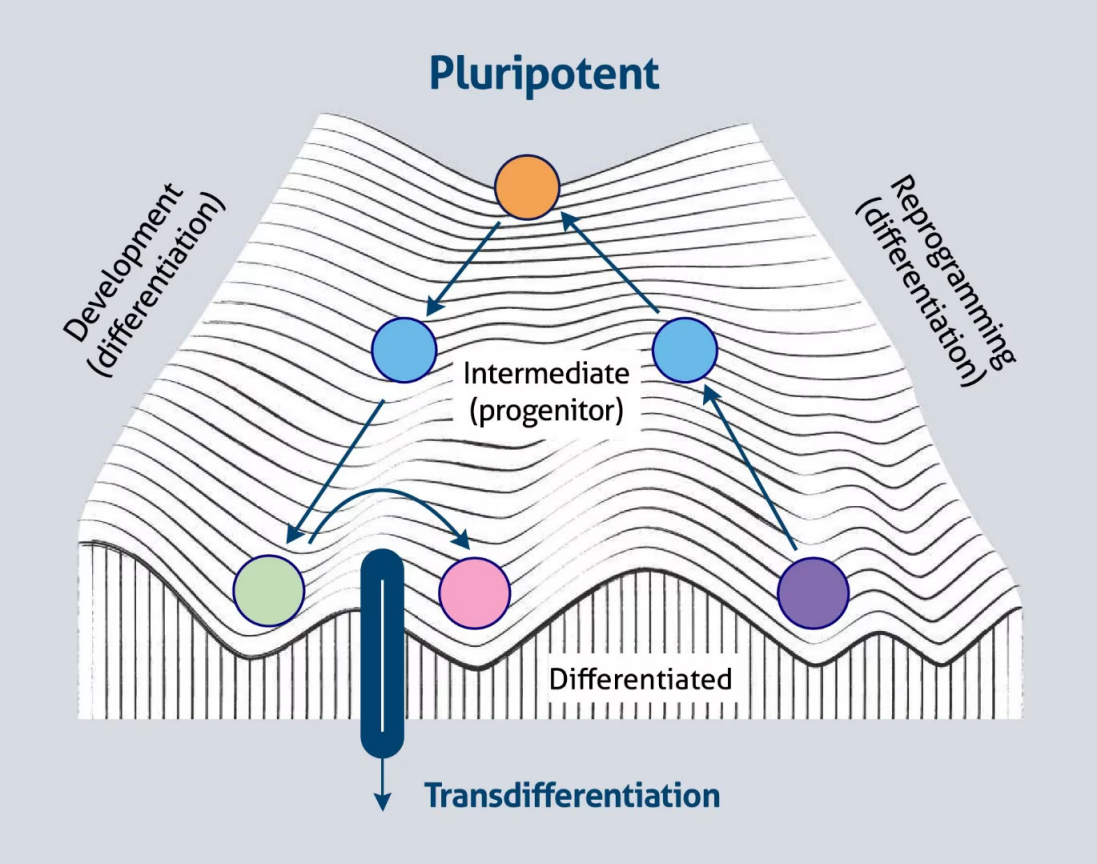

Figure 1 Cell-fate plasticity refers to the ability of a cell to adopt multiple potential developmental trajectories in response to intrinsic and extrinsic signals. This concept is often visualized through Waddington's epigenetic landscape, where cell states are represented as valleys and hills, and differentiation is depicted as a ball rolling down branching paths. In this metaphor, the shape of the landscape encodes both the stability of cell fates and the potential for plastic transitions between states. High plasticity corresponds to shallow valleys or flexible branching points, allowing cells to respond dynamically to perturbations and environmental cues. Understanding how molecular mechanisms reshape this landscape is crucial for interpreting developmental processes, cellular reprogramming, and disease progression. [22], [29].

More recently, [32] presented *scGPT*, a foundation model for single-cell multi-omics inspired by advances in large language models. scGPT leverages transformer architectures and large-scale pretraining across millions of cells to enable cross-modal prediction, perturbation response modeling, and transfer learning. Unlike earlier models trained on specific datasets, scGPT offers a general-purpose backbone adaptable to a wide range of single-cell analysis tasks, representing a shift toward foundation models in computational biology.

Also, the authors in [30] proposed *scGen*, a variational autoencoder that learns a latent representation of cells and models perturbations as vector operations in this space. scGen demonstrated that generative neural networks could extrapolate responses of unseen perturbations or cell types, highlighting the power of latent generative modeling for perturbation prediction. However, such VAE-based frameworks implicitly rely on convex optimization heuristics in a highly nonconvex landscape. When viewed through the lens of Waddington's epigenetic landscape, scGen's latent arithmetic can become trapped in local optima that correspond to spurious differentiation paths, leading to biologically implausible cell states if initialization is not carefully managed.

To improve generalization, [31] introduced *chemCPA*, a conditional perturbation autoencoder that integrates molecular structure information with single-cell gene expression data. By conditioning on chemical embeddings and leveraging adversarial training, chemCPA showed improved predictive accuracy for novel compounds and provided mechanistic interpretability for drug responses. Yet, the framework still inherits the same limitations of deep autoencoders: the learned perturbation manifold may become locally biased, forcing cells into "false valleys" of the Waddington landscape. Without mechanisms to escape these local traps, the model risks predicting lineage decisions that correspond to bad minima, cellular fates that do not reflect biological reality.

The latest development in this line is [32], who proposed *scGPT*, a foundation model for single-cell multiomics trained at scale with transformer architectures. scGPT offers impressive transferability, enabling cross-modal prediction and perturbation response modeling by pre-training on millions of cells. This represents a conceptual shift toward general-purpose backbones for single-cell biology. However, transformer-based foundation models exacerbate nonconvexity: their immense parameterization can memorize local valleys of the differentiation landscape, and inappropriate initialization or inadequate inductive bias may cause the model to reinforce suboptimal lineage bifurcations. Although scGPT captures global cell state representations, its optimization trajectory does not guarantee

exploration of the full Waddington landscape, limiting its ability to faithfully model rare or hard-to-reach differentiation outcomes. In addition, Ahlmann-Eltze et al. demonstrated that deep learning–based methods, despite their complexity, did not outperform simple linear baseline models in predicting transcriptome responses to single and double perturbations [33]. The authors attributed this limitation to challenges inherent in nonconvex optimization, where the models failed to converge effectively or differentiate robustly across lineages. This finding highlights the need for more reliable approaches that balance model expressiveness with stability and generalizability.

In summary, while the above models advance the state of single-cell differentiation modeling through generative latent modeling, chemically informed conditioning, and large-scale pretraining, they all suffer from the fundamental challenge of navigating a rugged, nonconvex Waddington landscape. The risk of becoming trapped in local optima shows as biologically implausible differentiation trajectories, incorrect lineage assignments, or exaggerated drug responses. Overcoming these limitations requires methods that explicitly incorporate nonconvex dynamics and robust initialization strategies to help the perturbation model escape local optima and converge to a solution, lineage, or cell fate.

III. TRPO-Preconditioned PPO multistage

Let $\pi_\theta(a|s)$ denote the policy with parameters θ, and let θ_{old} be the parameters used to collect a batch of T on-policy trajectories. Let $\hat{A}_t$ denote the estimated advantage at time step t. Define the importance sampling ratio:

$$r_\theta(s_t, a_t) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{\text{old}}}(a_t|s_t)}. \tag{1}$$

A. Surrogate Objective and Gradient

The standard policy gradient surrogate is

$$L_{\text{surr}}(\theta) = \mathbb{E}_t\left[r_\theta(s_t, a_t)\,\hat{A}_t\right]. \tag{2}$$

Its gradient at θ_{old} is

$$g = \nabla_\theta L_{\text{surr}}(\theta)\big|_{\theta_{\text{old}}} = \frac{1}{T}\sum_{t=1}^{T} \nabla_\theta \log \pi_\theta(a_t|s_t)\big|_{\theta_{\text{old}}}\,\hat{A}_t. \tag{3}$$

B. Fisher-Vector Product and Conjugate Gradient

Define the mean KL divergence:

$$\bar{D}_{KL}(\theta_{\text{old}}\|\theta) = \frac{1}{T}\sum_{t=1}^{T} D_{KL}\big(\pi_{\theta_{\text{old}}}(\cdot|s_t)\,\|\,\pi_\theta(\cdot|s_t)\big), \tag{4}$$

and its Hessian (Fisher matrix) $H = \nabla^2_\theta \bar{D}_{KL}(\theta_{\text{old}}\|\theta)|_{\theta_{\text{old}}}$. We solve $Hd = g$ approximately using matrix-free Conjugate Gradient (CG). For any vector v, the Fisher-vector product is

$$Hv \approx \nabla_\theta\Big(\nabla_\theta \bar{D}_{KL}(\theta)\cdot v\Big)\Big|_{\theta_{\text{old}}} + \lambda_{\text{damp}}\, v, \tag{5}$$

where λ_{damp} is a small damping constant for numerical stability.

C. Natural Gradient Step with KL Scaling

The natural step is scaled to satisfy a trust-region KL constraint δ:

$$\alpha = \sqrt{\frac{2\delta}{g^\top d + \epsilon}}, \quad \Delta\theta_{\text{nat}} = \alpha d, \quad \theta' = \theta_{\text{old}} + \Delta\theta_{\text{nat}}, \tag{6}$$

where ϵ is a small constant to prevent division by zero. Optionally, a backtracking line search can be applied to ensure $\bar{D}_{KL}(\theta_{\text{old}}\|\theta') \le \delta$.

D. PPO Clipped Fine-Tuning

Starting from θ', we perform a PPO-style fine-tune with the clipped surrogate:

$$L_t^{\text{CLIP}}(\theta) = \min\Big(r_\theta(s_t,a_t)\,\hat{A}_t,\ \text{clip}(r_\theta(s_t,a_t), 1-\varepsilon_{\text{clip}}, 1+\varepsilon_{\text{clip}})\,\hat{A}_t\Big), \tag{7}$$

and the total fine-tune objective is

$$\mathcal{L}_{\text{PPO}}(\theta) = -\mathbb{E}_t[L_t^{\text{CLIP}}(\theta)] + c_v\,\mathbb{E}_t\Big[\tfrac{1}{2}(V_\theta(s_t) - \hat{R}_t)^2\Big] - c_e\,\mathbb{E}_t[\mathcal{H}(\pi_\theta(\cdot|s_t))], \tag{8}$$

where $V_\theta(s_t)$ is the value function estimate, $\hat{R}_t$ is the return, $\mathcal{H}(\cdot)$ denotes policy entropy, and c_v, c_e are weighting coefficients. Fine-tuning is performed over a few minibatch epochs to exploit PPO's sample efficiency without violating the trust region as shown in 1 and figure 2.

IV. Evaluation

Insilico CRISPR-based perturbations provide a powerful framework to investigate gene function by systematically eliminating or overexpressing target genes. In single-cell experiments, such perturbations allow the study of cellular responses at high resolution, revealing gene regulatory relationships and context-specific effects.

The open AI gym environment simulates the effects of CRISPR interference on individual cells, modeling gene expression changes under controlled perturbations. Each episode represents a single cell, and actions correspond to predicted adjustments in gene expression. The environment incorporates stochastic perturbations, including gene knockouts and overexpression, enabling the evaluation of predictive models under realistic, variable gene regulatory dynamics. Rewards are defined to encourage models to accurately approximate observed post-perturbation expression profiles, making this setup suitable for training reinforcement learning agents to predict single-cell responses to genetic interventions.

The dataset is organized into multiple modalities, such as RNA, ATAC expression, and split into training and testing sets. During training, models observe a subset of cells and perturbations, while evaluation is performed on held-out cells and unseen perturbations to assess generalization. Gene expression matrices are fused across modalities according to pre-defined splits, and pseudotime information is incorporated to account for temporal dynamics.

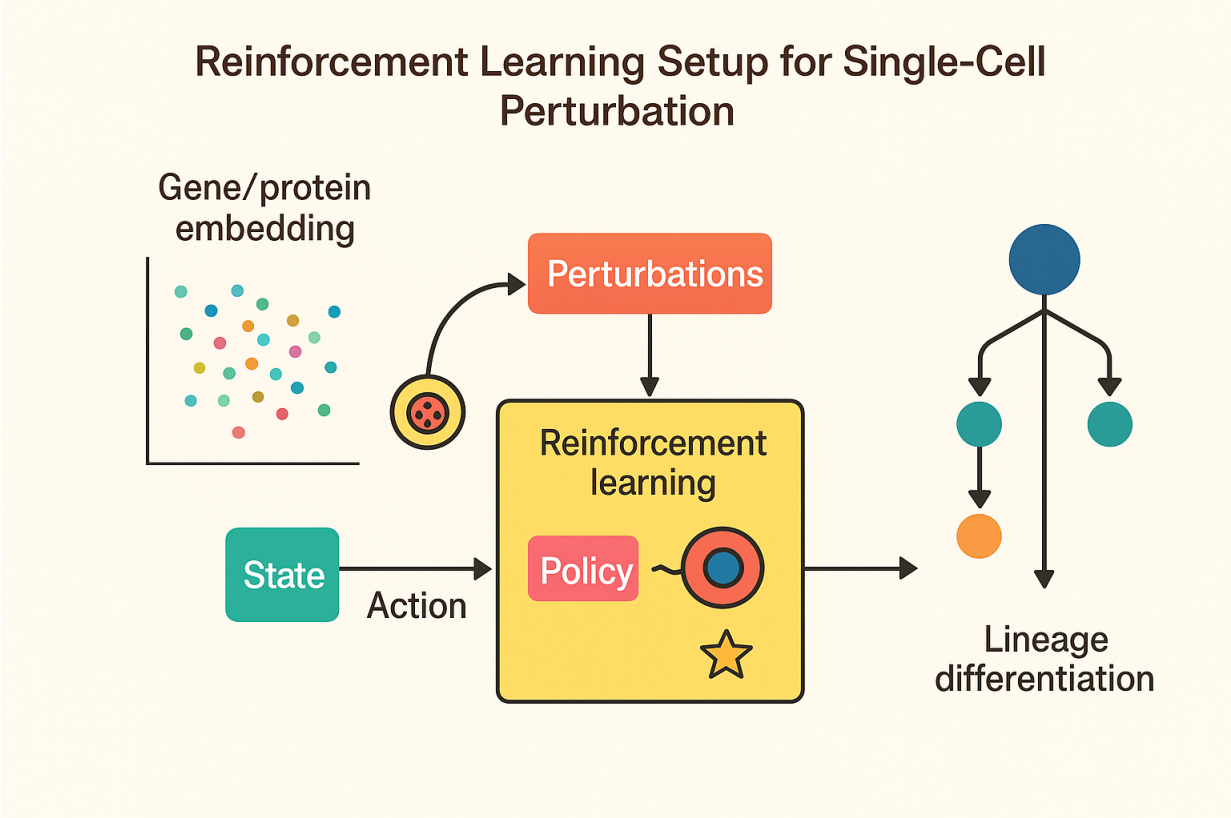

Figure 2 Single Cell Perturbation Analysis with the Proposed TRPO_PPO Method

We train three types of reinforcement learning agents: PPO, TRPO, and a multi-stage TRPO_PPO approach. PPO and TRPO are optimized with modality-specific policy networks (multiple hidden layers of 256 neurons per branch), and the TRPO_PPO pipeline allows for warm-starting PPO with parameters learned by TRPO. Training is performed for up to 1000000 timesteps per agent, with reward signals derived from the reduction in mean squared error between predicted and target expression profiles. For the multistage TRPO_PPO method, we used 10_000 steps for the first phase of TRPO phase training and the remaining steps for the PPO phase.

Evaluation Metrics

We evaluate the performance of our models across different modalities using the following metrics:

- **Accuracy (ACC):** The fraction of correctly predicted perturbation effects for each gene.

$$\text{ACC} = \frac{\text{Number of correct predictions}}{\text{Total predictions}}$$

- **Precision (PREC):** The proportion of true positive predictions among all positive predictions.

$$\text{PREC} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

- **Recall (REC):** The proportion of true positive predictions among all actual positives.

$$\text{REC} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

- **F1 Score (F1):** The harmonic mean of precision and recall.

$$\text{F1} = 2 \cdot \frac{\text{PREC} \cdot \text{REC}}{\text{PREC} + \text{REC}}$$

- **Area Under Precision-Recall Curve (AUPRC):** Measures the area under the precision-recall curve for binary predictions.
- **Mean Squared Error (MSE):** Average squared difference between predicted and observed gene expression.

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2$$

- **Root Mean Squared Error (RMSE):** Square root of MSE.

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2}$$

- **Mean Absolute Error (MAE):** Average absolute difference between predicted and observed gene expression.

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}|\hat{y}_i - y_i|$$

- **Coefficient of Determination (R^2):** Proportion of variance in observed expression explained by predictions.

$$R^2 = 1 - \frac{\sum_i(\hat{y}_i - y_i)^2}{\sum_i(y_i - \bar{y})^2}$$

Algorithm 1 TRPO-Preconditioned PPO multistage

Require: Initial policy parameters θ_0, KL threshold δ, damping λ_{damp}, PPO clip ε, number of PPO fine-tune epochs N_{ppo}

for iteration $k = 0, 1, 2, \ldots$ **do**

Collect a batch of trajectories $\{s_t, a_t, r_t\}_{t=1}^{T}$ using current policy π_{θ_k}

Compute advantages $\hat{A}_t$ and returns $\hat{R}_t$

Compute policy gradient at θ_k:

$$g = \frac{1}{T}\sum_{t=1}^{T} \nabla_\theta \log \pi_\theta(a_t|s_t)\hat{A}_t\Big|_{\theta_k}$$

Compute Fisher-Vector Product (FVP) function:

$$v \mapsto \nabla_\theta(\nabla_\theta \bar{D}_{KL}(\theta_k) \cdot v) + \lambda_{\text{damp}} v$$

Solve $Hd = g$ approximately via Conjugate Gradient to get natural step d

Scale step to satisfy KL constraint:

$$\alpha = \sqrt{\frac{2\delta}{g^\top d + \epsilon}}, \quad \Delta\theta_{\text{nat}} = \alpha d$$

Update parameters: $\theta' = \theta_k + \Delta\theta_{\text{nat}}$

▷ TRPO stage complete; switch to PPO fine-tuning

for epoch = 1 to N_{ppo} **do**

for each minibatch in the batch **do**

Compute PPO clipped surrogate:

$$L_t^{\text{CLIP}}(\theta) = \min\left(r_\theta \hat{A}_t,\ \text{clip}(r_\theta, 1-\varepsilon, 1+\varepsilon)\hat{A}_t\right)$$

Compute total PPO loss:

$$\mathcal{L}_{\text{PPO}} = -\mathbb{E}[L^{\text{CLIP}}] + c_v\,\mathbb{E}[(V_\theta(s_t) - \hat{R}_t)^2] - c_e\,\mathbb{E}[\mathcal{H}(\pi_\theta)]$$

Update θ' with optimizer (i.e., SGD/Adam)on $\mathcal{L}_{\text{PPO}}$

end for

end for

Set $\theta_{k+1} \leftarrow \theta'$

end for

- **Pearson Correlation Coefficient (PCC):** Linear correlation between predicted and observed gene expression.

$$\text{PCC} = \frac{\text{Cov}(\hat{y}, y)}{\sigma_{\hat{y}}\sigma_y}$$

Here, $\hat{y}_i$ and y_i represent the predicted and observed expression for gene i, respectively, and N is the number of test samples. TP, FP, and FN refer to true positives, false positives, and false negatives for binary prediction of gene perturbation effects.

A. RNA single cell dataset evaluation results

Multistage TRPO_PPO not only achieves higher classification-style metrics but also demonstrates superior performance in continuous expression prediction metrics (MSE, R^2, and Pearson correlation). This indicates that the hybrid policy produces policy/value function representations that more effectively capture the mapping from biological state variables (e.g., cell features, perturbation, pseudotime) to expression levels. The consistent improvement across both discrete and continuous evaluation measures relative to PPO alone highlights the importance of the optimization trajectory and the local minima reached. The multistage approach likely leverages TRPO's stable and conservative policy updates for robust exploration, followed by PPO's efficient fine-tuning within a favorable basin of attraction resulting in the notable gains in R^2 and Pearson correlation and the reduction in MSE observed in Table II.

B. ATAC single cell dataset evaluation results

The standard Policy optimization for controlling perturbations and predicting expression is inherently nonconvex: neural-network policies plus environment dynamics produce many local minima and saddle points. We expected rugged loss surfaces, not simple convex bowls.

TRPO enforced a trust-region step that can be seen as a conservative, curvature-aware move in parameter space. Pretraining with TRPO likely placed the optimizer in a good basin-of-attraction of the nonconvex landscape (a broad, low-loss valley), making subsequent PPO fine-tuning (clipped surrogate) more effective. The spectacular gains of Multistage TRPO_PPO are consistent with finding a much better basin rather than merely local overfitting.

The very low test MSE and high R^2 / Pearson for Multistage TRPO_PPO suggested the solution lies in a flat, robust region of the landscape (flat minima usually correlate with better generalization), rather than a sharp overfitted point as shown in Table III. The strong algorithmic gap indicates multimodality (multiple attractors), geometry matters, and algorithmic trajectory (trust-region then clipped updates) changed which attractor was found.

Furthermore, Multistage TRPO_PPO appeared to find perturbations that move cells into attractor basins that are much closer to ground-truth expression states (low MSE, high Pearson). In landscape terms: it found actions that push cells onto the correct valley floor with small residuals, rather than skimming the ridge or landing in the wrong nearby basin.

ATAC signals are sparser/noisier than RNA expression but contain regulatory accessibility information. The fact that Multistage TRPO_PPO still achieves very high R^2 and Pearson for ATAC suggests the model captures the regulatory to expression mapping well using attention-based graph neural network. Also, the learned policy identified perturbations that consistently translated to correct chromatin/activity readouts (i.e., robustly landing in the right attractor).

C. Joint RNA and ATAC single cell dataset results

The large performance gap and dramatic reduction in expression error for Multistage TRPO_PPO indicated that it found a much better region of the nonconvex Waddington landscape (likely a flatter, robust basin), translating into more accurate placement of cells in the Waddington-like attractor landscape. This implied that the Multistage TRPO_PPO training strategy is effective for both classification and continuous expression tasks in your single-cell perturbation setting, for both ATAC and RNA modalities as shown in Table I.

Table I Performance comparison of different algorithms on testing Joint dataset(RNA and ATAC single cell data).

Algorithm	Accuracy	Precision	Recall	F1	AUPRC	MSE	RMSE	MAE	R^2	Pearson Corr
PPO	0.6281	0.6577	0.6281	0.6283	0.6095	0.6044	0.7755	0.7004	0.1741	0.6388
TRPO_PPO	**0.6661**	**0.6940**	**0.6661**	**0.6675**	**0.6428**	**0.5948**	**0.7672**	**0.6935**	**0.1948**	**0.6495**

Table II Comparison of PPO and TRPO_PPO Performance Metrics on RNA single-cell dataset.

Algorithm	**Accuracy**	**Precision**	**Recall**	**F1**	**AUPRC**	**MSE**	**RMSE**	**MAE**	**R^2**	**PearsonCorr**
PPO	0.575	0.5892	0.575	0.5739	0.5863	0.6756	**0.8169**	**0.7563**	-0.1147	0.5275
TRPO_PPO	**0.5938**	**0.6095**	**0.5938**	**0.5878**	**0.6001**	**0.6714**	0.8165	0.7560	**-0.1017**	**0.5506**

Table III Performance comparison of different algorithms on testing ATAC single-cell dataset.

Algorithm	**Accuracy**	**Precision**	**Recall**	**F1**	**AUPRC**	**MSE**	**RMSE**	**MAE**	**R^2**	**PearsonCorr**
PPO	0.6083	0.6436	0.6083	0.6033	0.5930	0.6704	0.8148	0.7500	0.2546	0.6405
TRPO_PPO	**0.6521**	**0.6798**	**0.6521**	**0.6510**	**0.6356**	**0.6638**	**0.8111**	**0.7444**	**0.2418**	**0.6475**

In summary, the proposed framework aligns with Waddington's landscape model by conceptualizing cell differentiation and reprogramming as an optimization process over a complex developmental landscape. In this context, the multistage TRPO_PPO optimization mimics the biological progression of cells transitioning from degenerative to regenerative states. TRPO's stable exploration allows the model to navigate the high-dimensional epigenetic ridges that separate cell fates, while PPO's fine-tuning efficiently refines trajectories toward target regenerative valleys, corresponding to desired differentiated or reprogrammed states. By learning optimal control policies over this landscape, the framework can predict and guide reprogramming strategies that move cells toward pluripotent or lineage-specific outcomes. Consequently, this approach supports personalized therapeutic prescription for patients with degenerative conditions by integrating cell regeneration strategies such as induced pluripotent stem cell reprogramming, enabling the identification of optimal regenerative interventions tailored to each patient's molecular and cellular profile as shown in Figure 1.

V. Conclusion

In this work, we introduced a multistage TRPO_PPO reinforcement learning algorithm designed to overcome the challenges of modeling single-cell perturbation effects in the rugged and nonconvex Waddington landscape. By combining the curvature-aware, trust-region updates of TRPO with the sample-efficient fine-tuning of PPO, our approach successfully balances stability, exploration, and efficiency. Across RNA, ATAC, and joint multimodal datasets, the TRPO_PPO pipeline consistently outperformed PPO baseline, achieving substantial improvements in accuracy, F1, and precision-recall metrics, while dramatically reducing reconstruction error and boosting correlation with ground-truth expression.

These results demonstrated that the optimization trajectory, not just the final model, plays a critical role in navigating the high-dimensional landscape of gene expression. The conservative initialization of TRPO enables the policy to escape poor local optima, while PPO efficiently exploits the favorable basin to refine predictions. This synergy translates into robust generalization across modalities, indicating that our approach not only identifies correct perturbation responses but also places cells into biologically plausible attractor states.

Beyond its immediate performance gains, the multistage TRPO_PPO framework provides a general principle for perturbation modeling in systems biology: coupling curvature-aware initialization with efficient local refinement can enhance convergence in nonconvex settings. Future directions include extending this approach to multi-agent reinforcement learning systems, as well as improving activation functions, optimizers, and loss function formulations to better escape local optima in the Waddington landscape. In addition, integrating machine learning methods for the detection of class II antigens could expand the applicability of our framework to immunogenomics and related biomedical domains.

References

[1] T. Heydari, M. A. Langley, C. L. Fisher, D. Aguilar-Hidalgo, S. Shukla, A. Yachie-Kinoshita, M. Hughes, K. M. McNagny, and P. W. Zandstra, "Iqcell: A platform for predicting the effect of gene perturbations on developmental trajectories using single-cell rna-seq data," *PLOS Computational Biology*, vol. 18, no. 2, p. e1009907, 2022. [Online]. Available: https://doi.org/10.1371/journal.pcbi.1009907

[2] S. Semrau and A. van Oudenaarden, "Studying lineage decision-making in vitro: Emerging concepts and novel tools," *Annu Rev Cell Dev Biol*, vol. 31, pp. 317–345, 2015.

[3] Y. Lipsitz, N. Timmins, and P. Zandstra, "Quality cell therapy manufacturing by design," *Nat Biotechnol*, vol. 34, pp. 393–400, 2016.

[4] L. Prochazka, Y. Benenson, and P. Zandstra, "Synthetic gene circuits and cellular decision-making in human pluripotent stem cells," *Curr Opin Syst Biol*, vol. 5, pp. 93–103, 2017.

[5] K. Takahashi and S. Yamanaka, "Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors," *Cell*, vol. 126, pp. 663–676, 2006.

[6] S.-J. Dunn, H. Kugler, and B. Yordanov, "Formal analysis of network motifs links structure to function in biological programs," *IEEE/ACM Trans Comput Biol Bioinform*, 2019.

[7] I. Peter, E. Faure, and E. Davidson, "Predictive computation of genomic logic processing functions in embryonic development," *Proc Natl Acad Sci*, vol. 109, pp. 16 434–16 442, 2012.

[8] A. Yachie-Kinoshita, K. Onishi, J. Ostblom, M. Langley, E. Posfai, and J. Rossant, "Modeling signaling-dependent pluripotency with boolean logic to predict cell fate transitions," *Mol Syst Biol*, vol. 14, 2018.

[9] S.-J. Dunn, G. Martello, B. Yordanov, S. Emmott, and A. Smith, "Defining an essential transcription factor program for naive pluripotency," *Science*, vol. 344, pp. 1156–1160, 2014.

[10] B. Yordanov, S.-J. Dunn, H. Kugler, A. Smith, G. Martello, and S. Emmott, "A method to identify and analyze biological programs through automated reasoning," *Npj Syst Biol Appl*, vol. 2, p. 16010, 2016.

[11] A. Babtie, T. Chan, and M. Stumpf, "Learning regulatory models for cell development from single cell transcriptomic data," *Curr Opin Syst Biol*, vol. 5, pp. 72–81, 2017.

[12] M. Fiers, L. Minnoye, S. Aibar, C. Bravo Gonzalez-Blas, Z. Kalender Atak, and S. Aerts, "Mapping gene regulatory networks from single-cell omics data," *Brief Funct Genomics*, vol. 17, pp. 246–254, 2018.

[13] A. Pratapa, A. Jalihal, J. Law, A. Bharadwaj, and T. Murali, "Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data," *Nat Methods*, 2020.

[14] F. Hamey, S. Nestorowa, S. Kinston, D. Kent, N. Wilson, and B. Göttgens, "Reconstructing blood stem cell regulatory network models from single-cell molecular profiles," *Proc Natl Acad Sci*, vol. 114, pp. 5822–5829, 2017.

[15] J. D. Buenrostro, P. G. Giresi, L. C. Zaba, H. Y. Chang, and W. J. Greenleaf, "Single-cell chromatin accessibility reveals principles of regulatory variation," *Nature*, vol. 523, no. 7561, pp. 486–490, 2015.

[16] Z. Li, Y. Zhang, J. Li, and et al., "Chromatin-accessibility estimation from single-cell atac-seq data," *Nature Communications*, vol. 12, no. 1, pp. 1–11, 2021.

[17] E. R. Gur and et al., "scatac-seq generates more accurate and complete chromatin accessibility profiles compared to bulk atac-seq," *Scientific Reports*, vol. 15, no. 1, pp. 1–11, 2025.

[18] M. Stoeckius and et al., "Cite-seq: Coupling antibody-based proteomics with single-cell rna-seq," *Nature Methods*, vol. 14, no. 9, pp. 865–868, 2017.

[19] Y. Chen and et al., "A joint analysis of single cell transcriptomics and proteomics using cite-seq," *Nature Computational Biology*, vol. 1, no. 1, pp. 1–13, 2025.

[20] N. Tjoonk, "Cite-seq: Single-cell rna sequencing + surface protein analysis," 2023, single Cell Discoveries. [Online]. Available: https://www.scdiscoveries.com/blog/knowledge/cite-seq/

[21] Anonymous Authors, "scagents: A multi-agent framework for fully autonomous end-to-end single-cell perturbation analysis," in *ICML 2025 Workshop on GenBio (preprint)*, 2025, preprint (ICML submission). Code and models: https://anonymous.4open.science/r/scAgents-2025-242E/.

[22] S. F. Gilbert, "Epigenetic landscaping: Waddington's use of cell fate bifurcation diagrams," *Biology and Philosophy*, vol. 6, no. 2, pp. 135–154, 1991.

[23] Y. Wang, H. He, X. Tan, and Y. Gan, "Trust region–guided proximal policy optimization," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2019, neurIPS 2019 paper / arXiv preprint. [Online]. Available: https://arxiv.org/abs/1901.10314

[24] R.-A. Lascu, D. Šiška, and Ł. Szpruch, "Ppo in the fisher–rao geometry," arXiv preprint arXiv:2506.03757, 2025. [Online]. Available: https://arxiv.org/abs/2506.03757

[25] T. Moskovitz, R. Wang, J. Lan, S. Kapoor, T. Miconi, J. Yosinski, and A. Rawal, "First–order preconditioning via hypergradient descent," in *International Conference on Learning Representations (ICLR) – OpenReview*, 2020, arXiv and OpenReview material (first posted 2019 as arXiv:1910.08461). [Online]. Available: https://openreview.net/forum?id=Skg3104FDS

[26] J. Markowitz and E. W. Staley, "Clipped-objective policy gradients for pessimistic policy optimization (copg)," arXiv preprint arXiv:2311.05846, 2023. [Online]. Available: https://arxiv.org/abs/2311.05846

[27] J. Schulman, S. Levine, P. Abbeel, M. Jordan, and P. Moritz, "Trust region policy optimization," in *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, 2015. [Online]. Available: https://arxiv.org/abs/1502.05477

[28] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," in *arXiv preprint arXiv:1707.06347*, 2017. [Online]. Available: https://arxiv.org/abs/1707.06347

[29] Proteintech Group, "Cell fate commitment and the waddington landscape model," https://www.ptglab.com/news/blog/cell-fate-commitment-and-the-waddington-landscape-model/, 2023, accessed: 2025-10-08.

[30] M. Lotfollahi, F. A. Wolf, and F. J. Theis, "scgen predicts single-cell perturbation responses," *Nature Methods*, vol. 16, no. 8, pp. 715–721, 2019.

[31] L. Hetzel, S. Böhm, N. Kilbertus, S. Günnemann, M. Lotfollahi, and F. Theis, "Predicting cellular responses to novel drug perturbations at a single-cell resolution," *arXiv preprint*, 2022, arXiv:2204.13545.

[32] H. Cui, C. Wang, H. Maan, K. Pang, F. Luo, N. Duan, and B. Wang, "scgpt: Toward building a foundation model for single-cell multi-omics using generative ai," *Nature Methods*, 2024.

[33] C. Ahlmann-Eltze, W. Huber, and S. Anders, "Deep-learning-based gene perturbation effect prediction does not yet outperform simple linear baselines," *Nature Methods*, vol. 22, pp. 1657–1661, 2025.