From Transcriptome-wide Prediction to Target Gene Discovery: Improving Virtual Cell Models with scGPT

Anrui Wang
2023533015
wangar2023@shanghaitech.edu.cn

Jiawen Dai
2023533132
daijw2023@shanghaitech.edu.cn

Yiting Qi
2023533043
qiyt2023@shanghaitech.edu.cn

Abstract

Predicting cellular responses to genetic or drug perturbations is a fundamental challenge in computational biology, with significant implications for drug discovery and functional genomics. Traditional virtual cell modeling, exemplified by the Virtual Cell Challenge (VCC), formulates this task as transcriptome-wide regression. However, transcriptome-wide evaluation suffers from high dimensionality, noisy measurements, and metrics dominated by non-informative genes, limiting its biological interpretability. In this work, we reformulate the task as target gene identification, ranking candidate genes by their likelihood of being true perturbation targets. We implement a refined evaluation framework for the VCC benchmark, re-splitting 150 known gene perturbations into train, validation, and test sets, and assess the performance of STATE and scGPT models. Furthermore, we develop scGPT-based models tailored for target gene discovery and extend our evaluation to drug perturbations using the Tahoe-100M dataset. Our study highlights both the limitations of transcriptome-wide metrics and the promise of target-focused modeling, providing insights for more biologically meaningful virtual cell predictions.

1 Introduction

1.1 Phenotypic screening creates an inverse problem

Phenotypic drug discovery has historically been a cornerstone of therapeutic development, uniquely capable of identifying compounds that rescue disease states in a physiologically relevant context without requiring prior knowledge of a specific molecular target. However, this "target-agnostic" success comes with a fundamental **inverse problem**: while the therapeutic effect—the phenotype—is observed, the underlying molecular mechanism of action (MoA) or target often remains unknown, necessitating a complex and laborious target deconvolution process. Unlike target-based approaches that proceed from cause to effect, phenotypic screening presents us with the effect, requiring us to solve for the cause.

High-throughput transcriptomic profiling has emerged as a critical bridge in this gap, offering a high-dimensional, information-rich "fingerprint" of cellular states that reflects the downstream consequences of upstream perturbations. By defining the biological phenotype as a computable vector, it becomes possible to systematically map the causal relationship between a perturbation (whether genetic or chemical) and its resulting cellular state. This capability transforms the biological question into a data-driven inference task: if we can quantify the phenotypic shift, can we identify the perturbation that caused it?

Recent advances in artificial intelligence have begun to address this by proposing the construction of "Virtual Cells"—high-fidelity simulators learned directly from large-scale biological data. As highlighted by Roohani et al. [RHT+25], these systems are expected to learn the fundamental relationship between cell state and function, with the primary intent of predicting the consequences of perturbations—such as gene knockdowns or drug treatments—across diverse cell contexts. While much of the current effort focuses on this **forward prediction**, the ultimate utility of AI Virtual Cells (AIVCs) extends beyond mere simulation to "closing the loop" for discovery. Bunne et al. [BRR+24] articulate a vision where AIVCs serve as engines for **in silico experimentation**, capable of simulating the effects of varying interventions to propose potential causal factors behind observed phenotypes. Although computation alone may not fully resolve all causal links, AIVCs offer the critical potential to drastically **"reduce the space of possible hypotheses"**, thereby accelerating the identification of underlying mechanisms and new drug targets. This positions the Virtual Cell not just as a simulator, but as a computational partner for target deconvolution in phenotypic drug discovery.

1.2 The community's asymmetry: optimizing forward metrics vs answering actionable questions

The field currently faces a critical asymmetry: while immense effort is poured into optimizing forward prediction metrics, empirical gains remain marginal. A systematic benchmark across diverse datasets reveals that complex deep learning models do not consistently outperform simpler alternatives, with performance often heavily confounded by perturbation effect sizes rather than architectural superiority [LYF+24]. More strikingly, recent rigorous evaluations serve as a wake-up call, demonstrating that state-of-the-art models for perturbation prediction do not yet consistently outperform simple linear baselines or even trivial mean predictions [AE]. This suggests that the singular pursuit of minimizing reconstruction error (MSE) may not inherently lead to the actionable causal discovery required for drug development.

1.3 Reframing: invert the causal chain to make models actionable

To bridge the gap between model performance and biological discovery, we propose a fundamental reframing of the Virtual Cell's utility: shifting the focus from **forward simulation** (predicting the expression profile of a known perturbation) to **reverse retrieval** (inferring the causal perturbation of an observed profile). This is not an arbitrary task invention but a rigorous operationalization of the "in silico reverse perturbation prediction" paradigm already emerging in foundational models. For instance, **scGPT** [CWM+24] explicitly formalizes this as a top-K retrieval task, demonstrating that foundation models can learn to rank potential genetic drivers based on their alignment with a target cell state. By adopting this retrieval-based formulation, we align the model's output directly with the decision-making logic of drug discovery: generating a ranked list of probable targets for experimental validation.

This inversion transforms the Virtual Cell from a descriptive simulator into a prescriptive engine for **therapeutic target identification**. This utility has been notably demonstrated by **Geneformer** [TXC+23], which leveraged transfer learning on large-scale data to move beyond mere expression forecasting, successfully prioritizing candidate therapeutic targets and key network regulators for cardiomyopathy. However, reliable reverse inference requires models to possess true mechanistic reasoning rather than simple pattern matching, particularly when dealing with the combinatorial complexity of biological systems. As evidenced by **GEARS** [RHL24], incorporating biological inductive biases—such as gene regulatory graphs—is essential for generalizing to unseen perturbations and combinatorial sets that lack direct experimental training data. Similarly, achieving robust target discovery requires models (like the transformer-based STATE [AGB+]) to model distributions across **sets of cells**, enabling mechanistic generalization to novel cellular contexts where training data may be sparse or absent. By closing this causal loop, we move from optimizing abstract error metrics to answering the actionable question: *what perturbation caused this phenotype?*

1.4 Contributions

In this work, we close the loop between forward simulation and reverse inference, presenting a unified framework that repurposes the Virtual Cell from a predictive simulator into a prescriptive engine for target discovery. Our primary contributions are as follows:

1. **Rigorous Evaluation of Forward Generalization:** We establish a strict evaluation protocol for forward perturbation prediction using the **Virtual Cell Challenge (VCC)** dataset [RHT+25]. Unlike standard benchmarks that often rely on random cell-splits, we enforce **perturbation-level splits** to test true mechanistic generalization. By evaluating state-of-the-art models (including scGPT and STATE) against biologically grounded metrics (DES, PDS) rather than simple reconstruction error, we validate their capacity to capture causal perturbation-response maps beyond mere data fitting.

2. **A Unified Paradigm for Reverse Target Retrieval:** We formally define the task of **in silico reverse perturbation prediction** as a retrieval and ranking problem. We demonstrate that this paradigm—which takes a transcriptomic phenotype as query to retrieve the most probable causal agents—can be effectively generalized across different perturbation modalities. This validates the feasibility of using deep perturbation models to bridge the gap between phenotypic screening readouts and target identification, aligning drug-induced signatures with genetic ground truths.

3. **Decoupling Geometric and Semantic Inference:** We uncover a critical **"Geometry vs. Semantics" trade-off** in phenotypic retrieval. Our comparative analysis reveals that while simple heuristic baselines (e.g., PCA-KNN) excel at "head" precision for strong phenotypes via geometric similarity, pre-trained foundation models (e.g., scGPT) demonstrate superior robustness and "tail" recall. This finding suggests that foundation models uniquely capture the semantic context required for deciphering complex or subtle biological signals that lie beyond simple geometric proximity.

Author Contributions

- **Anrui Wang**: Implemented scGPT experiments to fulfill the Virtual Cell Challenge (VCC) requirements, evaluated model performance, adapted scGPT to leverage the improved VCC evaluation metrics, and created scGPT-based models specifically for target gene identification, enabling a more biologically meaningful assessment of predictions.
- **Jiawen Dai**: Reimplemented a more biologically informed set of evaluation metrics for the Virtual Cell Challenge (VCC), conducted baseline experiments for VCC models, and extracted relevant data from the Tahoe-100M dataset.
- **Yiting Qi**: Implemented STATE experiments to fulfill the VCC requirements and evaluate performance, conducted studies on the Tahoe-100M dataset, and developed baseline models for target gene prediction on the Tahoe-100M dataset.

2 Forward validation as a credibility scaffold

2.1 Forward prediction on VCC under perturbation-level splits

To establish a rigorous baseline for virtual cell capabilities, we first benchmarked state-of-the-art models on the **Virtual Cell Challenge (VCC)** dataset. Unlike standard evaluations that often rely on random cell splits—which risk simplifying the task to cellular interpolation—we enforced a strict **perturbation-level split**. In this setting, the models were tasked with predicting the transcriptional effects of genetic perturbations that were entirely unseen during training, thereby testing true mechanistic generalization rather than pattern memorization.

We evaluated two representative model architectures: **STATE**, which leverages cell-set modeling to capture distributional shifts across cellular contexts, and **scGPT**, a foundational model that utilizes large-scale pre-training to learn generalizable gene-gene interactions. By systematically comparing these models alongside linear baselines, we defined the current performance boundary for forward prediction (Figure 2). This "forward validation" step serves as a necessary credibility scaffold; it confirms that the underlying embeddings capture sufficient biological signal to support the more complex inference tasks required for target discovery, without assuming that forward simulation is the endpoint of the modeling effort. All implementation details, including data splits and hyperparameter configurations, are detailed in Supplementary Note 1.

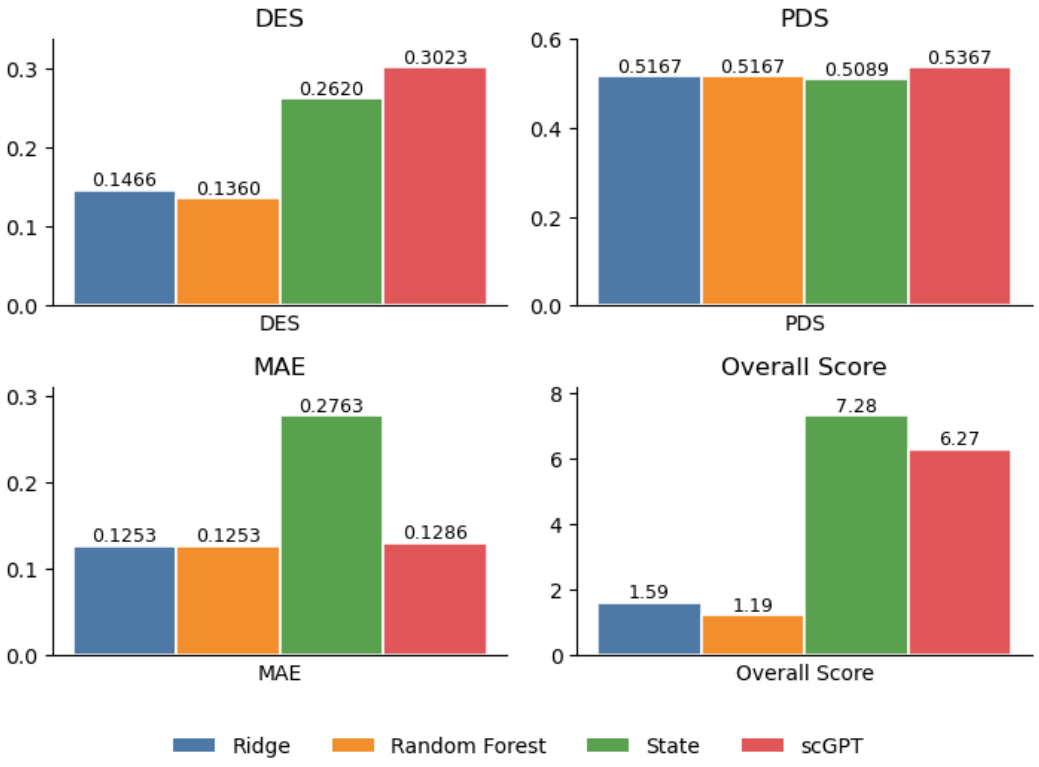

Figure 1: **Evaluation Metrics and Scoring Framework.** The Virtual Cell Challenge utilizes a composite scoring system derived from three complementary metrics to assess forward prediction performance. (A) Differential Expression Score (DES): Measures the model's fidelity in recovering biologically interpretable signals by calculating the overlap (precision and recall) between predicted and ground-truth differentially expressed genes (DEGs). (B) Perturbation Discrimination Score (PDS): A ranking-based metric that evaluates whether a predicted pseudo-bulk profile is more similar to its corresponding ground truth than to other perturbations. To mitigate biases toward vectors with small norms, PDS utilizes cosine similarity to emphasize directional alignment over magnitude. (C) Mean Absolute Error (Top-2k MAE): Captures the global reconstruction accuracy by computing the element-wise error across the top 2,000 highly variable genes. (D) Overall Score: A weighted aggregation of the individual metrics used to rank model performance, enforcing a balance between global fit (MAE), directional correctness (PDS), and biological specificity (DES).

2.2 Why we evaluate biology, not just numbers

In evaluating these models, we depart from a reliance on raw reconstruction errors, which can be misleading in transcriptomic data. Conventional metrics like Mean Absolute Error (MAE) are frequently dominated by the high baseline expression of invariant genes, potentially awarding high scores to trivial "mean-prediction" baselines while failing to distinguish true biological replicates from unrelated perturbations. Instead, as illustrated in Figure 2, we prioritize metrics that assess the **perturbation delta** (δ)—the specific causal shift in gene expression induced by the intervention relative to controls. We focus on two biologically interpretable metrics to serve as our primary credibility scaffold:

- **Differential Expression Score (DES):** This metric evaluates the model's ability to recover the specific set of differentially expressed genes (DEGs). By measuring the overlap (precision and recall) between predicted and ground-truth DEGs, DES verifies whether the model faithfully recapitulates the biologically interpretable effects of a perturbation, rather than merely fitting global statistical distributions.

- **Perturbation Discrimination Score (PDS):** Adapted here to utilize cosine similarity, PDS assesses the directional fidelity of the predicted shift. This ensures that the model correctly ranks the true perturbation as the most similar candidate among all possibilities. Unlike L1-based rankings, this approach emphasizes the alignment of up- or down-regulation vectors rather than signal magnitude, mitigating the penalty on "attenuated" predictions often seen in single-cell modeling.

This evaluation philosophy aligns with the "delta-based" validation advocated in recent foundational studies, ensuring that high performance reflects the capture of causal regulatory signals rather than the preservation of static cell-type identity.

Metric Analysis and Failure Modes The selection of these metrics addresses specific failure modes in predictive modeling (Figure 2D). While MAE provides a necessary baseline for global statistical fit, it remains largely insensitive to subtle, biologically meaningful shifts. Consequently, DES serves as the strictest test of biological utility, ensuring the model identifies the precise gene sets driving the phenotype. Furthermore, the adaptation of PDS addresses the "attenuation" phenomenon, where models correctly predict the direction of regulation but with dampened magnitude compared to ground truth. By prioritizing vector direction (cosine similarity) over Euclidean distance, we reward models that capture the correct regulatory mechanism even if the signal strength is conservative.

2.3 Takeaway: forward is necessary but not sufficient

While forward benchmarks verify that a model can simulate *effects*, therapeutic discovery necessitates inferring *causes* from observed effects. This fundamental asymmetry motivates an explicit inversion of the perturbation–phenotype mapping, moving from verifying simulation accuracy to validating retrieval capability.

3 Methods

4 Discussion

5 Conclusion

References

[AE] Constantin Ahlmann-Eltze. Deep-learning-based gene perturbation effect prediction does not yet outperform simple linear baselines.

[AGB+] 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.

[BRR+24] 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, October 2024. arXiv:2409.11654 [q-bio].

[CWM+24] 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, August 2024.

[LYF+24] Lanxiang Li, Yue You, Yunlin Fu, Wenyu Liao, Xueying Fan, Shihong Lu, Ye Cao, Bo Li, Wenle Ren, Jiaming Kong, Shuangjia Zheng, Jizheng Chen, Xiaodong Liu, and Luyi Tian. A Systematic Comparison of Single-Cell Perturbation Response Prediction Models, December 2024.

[RHL24] Yusuf Roohani, Kexin Huang, and Jure Leskovec. Predicting transcriptional outcomes of novel multigene perturbations with GEARS. *Nature Biotechnology*, 42(6):927–935, June 2024.

[RHT+25] 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*, 188(13):3370–3374, June 2025.

[TXC+23] 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, June 2023.

Appendix A Supplementary Note 1: Methods and Experimental Details

Appendix A.1 Methods

Appendix A.1.1 Transcriptome-wide Perturbation Prediction

STATE Implementation Pipeline We implement the STATE model following the official open-source implementation released by the authors. In this work, we focus on the State Transition (ST) module of STATE, which is responsible for predicting perturbation-induced transcriptomic responses given control cells and perturbation conditions.

Data preprocessing. We use the Arc Institute VCC competition support dataset (`competition_support_set`), which provides preprocessed single-cell RNA-seq data for H1 human embryonic stem cells. Gene expression values are transformed using a log1p normalization ($\log(1+x)$), which stabilizes variance and reduces the dominance of highly expressed genes. Control cells corresponding to non-targeting perturbations are retained and used as a reference state (x_{ntc}).

Perturbation features. Perturbation identities are represented using pretrained protein embeddings. Specifically, gene perturbations are encoded using ESM2-based perturbation feature vectors provided by the official STATE repository. These embeddings allow the model to incorporate prior biological information about gene identity (e.g., sequence homology) without directly accessing expression-level supervision, facilitating generalization to unseen genes.

Model architecture. The ST module takes as input control cell expression profiles, perturbation embeddings, and batch embeddings to account for technical variability. All inputs are projected into a shared hidden space and processed by a Transformer backbone, which models how perturbations induce shifts in cellular state distributions. The output representations are mapped back to the gene expression space via a decoding head to generate predicted perturbed transcriptomes ($\hat{x}_{\mathrm{pert}}$).

Appendix A.1.2 scGPT Implementation

We adapt scGPT, a foundation model for single-cell biology, for the perturbation prediction task. The model is initialized with weights pre-trained on a massive scale single-cell corpus.

Forward Prediction (Fine-tuning). For the forward task, we fine-tune scGPT using a modified masked language modeling (MLM) objective. The input sequence consists of gene expression tokens conditioned on a special perturbation token. The model is trained to reconstruct the expression of the perturbed state given the gene tokens and the perturbation condition. To prevent the model from learning technical artifacts, we apply random masking to batch indices during training.

Appendix A.2 From Transcriptome-wide Predictions to Target Gene Identification

We extend the modeling framework to solve the inverse problem: identifying the causal perturbation from a transcriptomic phenotype.

Task Formalization. We define the task as a retrieval ranking problem. Given a query phenotypic signature $x_{\mathrm{phenotype}}$ (e.g., a differentially expressed profile), the model must output a ranked list of candidate perturbations $\mathrm{Rank}(\mathcal{P})$ from the target space $\mathcal{P}$.

Reverse Mode Architecture. We utilize the scGPT backbone in a classification setup. The observed cell state (phenotype) is tokenized and processed by the transformer. A classification head is attached to the `[CLS]` token embedding to predict a probability distribution over the candidate targets.

Hybrid Ranking Loss. To handle the long-tail distribution of potential targets and improve retrieval performance, we implement a hybrid loss function:

$$\mathcal{L} = \mathcal{L}_{\text{rank}} + \lambda \mathcal{L}_{\text{bce}}$$

where $\mathcal{L}_{\text{rank}}$ is a ListMLE ranking loss that optimizes the relative order of candidates, and $\mathcal{L}_{\text{bce}}$ is a binary cross-entropy loss that learns independent target probabilities. To address class imbalance, we employ a dynamic negative sampling strategy, sampling 10 non-causal perturbations as negatives for every positive pair during each training epoch.

Appendix A.3 Experiments

Appendix A.3.1 Transcriptome-wide Perturbation Prediction

Dataset and Data Split. To rigorously evaluate mechanistic generalization, we do not use a random cell split. Instead, we re-split the 150 gene perturbations from the VCC dataset into training, validation, and test sets at the perturbation level.

- **Training:** 64% of perturbation genes (approx. 96 genes).
- **Validation:** 16% of perturbation genes.
- **Test:** 20% (30 genes) held-out perturbation genes.

Control (non-targeting) cells are included in all splits to provide a consistent reference state. Importantly, there is no overlap of perturbation genes between the training, validation, and test sets, ensuring that models are evaluated on their ability to predict the effects of entirely unseen perturbation targets based on their embeddings.

Appendix A.3.2 Evaluation metrics: PDS, MAE, DES

To align evaluation with biological utility, we employ three complementary metrics.

Perturbation Discrimination Score (PDS). The Perturbation Discrimination Score (PDS) measures whether the predicted gene expression changes induced by different perturbations preserve the correct global directionality across all genes. For each perturbation $k \in \{1, \ldots, N\}$, we first construct pseudobulk expressions by averaging all cells under the same perturbation, yielding the ground-truth pseudobulk $y_k \in \mathbb{R}^G$, the predicted pseudobulk $\hat{y}_k \in \mathbb{R}^G$, and the control (NTC) pseudobulk $y_{\text{ntc}} \in \mathbb{R}^G$. Perturbation-induced expression changes are then defined as $\delta_k = y_k - y_{\text{ntc}}$ and $\hat{\delta}_k = \hat{y}_k - y_{\text{ntc}}$.

For each predicted perturbation $\hat{\delta}_k$, we compute cosine similarities to all true perturbation deltas $\{\delta_j\}_{j=1}^N$:

$$S_{k,j} = \frac{\hat{\delta}_k \cdot \delta_j}{\|\hat{\delta}_k\|_2 \, \|\delta_j\|_2}.$$

These similarities are ranked, and the rank of the correct perturbation k is denoted by R_k. The final score is reported as the mean rank normalized to the number of perturbations: $\text{PDS}_{\text{rank}} = \frac{1}{N} \sum_{k=1}^{N} R_k$.

Motivation: In our task, we modify the original VCC PDS metric by replacing L1 distance with cosine similarity. An L1-based ranking can be biased toward vectors with smaller norms (conservative predictions), even when their directions are poorly aligned. The revised PDS places greater emphasis on directional correctness—whether genes are consistently predicted to be up- or down-regulated—addressing the "attenuation" often seen in single-cell models.

MAE on Top 2000 Genes (Top-2k MAE). To focus evaluation on biologically meaningful signals rather than invariant background genes, we compute the Mean Absolute Error (MAE) restricted to genes with the largest true perturbation effects. For each perturbation k, genes are ranked by their absolute $\log_2$ fold change ($\text{LFC}_{k,g}$) relative to control. The top 2000 genes form the index set Ω_k.

$$\text{MAE}_{\text{top2k}} = \frac{1}{N} \sum_{k=1}^{N} \left(\frac{1}{2000} \sum_{g \in \Omega_k} |\hat{y}_{k,g} - y_{k,g}| \right)$$

This metric emphasizes prediction accuracy on genes that exhibit the strongest true responses.

Differential Expression Score (DES). The Differential Expression Score (DES) measures how accurately a model recovers perturbation-induced differential gene expression. For each perturbation k, we identify significantly differentially expressed (DE) genes in both ground truth ($G_{k,\text{true}}$) and prediction ($G_{k,\text{pred}}$) using a Wilcoxon rank-sum test (FDR $\alpha = 0.05$, Benjamini–Hochberg correction).

The score is defined as the recall of the true DE set:

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

Table 1: Performance comparison of perturbation prediction models.

Model	PDS	MAE	DES	Overall
Cell-mean baseline	0.5167	0.1258	0.1075	0.00
Ridge Regression	0.5167	**0.1253**	0.1466	1.59
Random Forest Regression	0.5167	**0.1253**	0.1360	1.19
State Model	**0.5367**	0.1286	**0.3023**	**7.28**
scGPT finetune	0.5089	0.2763	0.2620	6.27

If the model predicts more DE genes than exist in the truth ($n_{k,\text{pred}} > n_{k,\text{true}}$), we truncate the predicted set to the top $n_{k,\text{true}}$ genes by LFC magnitude to prevent gaming the metric by over-prediction.

Overall Score. To rank models, we calculate a scaled composite score based on the improvement over a cell-mean baseline.

Appendix A.3.3 Main Results

We evaluated STATE and scGPT against representative baseline models (Ridge Regression, Random Forest Regression, and a trivial Cell-Mean baseline). The results are summarized in Table S1.

Analysis of Results. Among all methods, the STATE Model achieves the best overall performance, with substantial improvements in biologically relevant metrics (DES: 0.3023 vs 0.1466 for Ridge) and directional accuracy (PDS). The scGPT fine-tuned model also demonstrates competitive performance in DES (0.2620), confirming that deep models capture non-linear regulatory patterns better than linear baselines.

However, complex models do not consistently outperform simpler baselines on MAE. The simple linear regressions achieve the lowest MAE (0.1253), only marginally better than the cell-mean baseline (0.1258). This saturation suggests that MAE is dominated by invariant genes and is a poor differentiator of model quality in this context.

The discrepancy between high DES/PDS and average MAE in deep models highlights the specific nature of their learning:

- **Attenuation:** Deep models learn meaningful single-cell patterns (high DES), but the predicted magnitudes are often dampened to minimize uncertainty, leading to smaller δ vectors. This lowers PDS if magnitude is considered, justifying our use of cosine similarity.
- **Noise vs. Signal:** Single-cell data is inherently noisy. Deep models trained on single-cell losses may not optimize perfectly for pseudo-bulk metrics, whereas regression baselines trained to minimize variance often converge to the mean.

These findings reinforce our argument that evaluation must prioritize biological fidelity (DES/PDS) over global reconstruction error (MAE).