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 Background

Cells are the fundamental functional units of biological systems, whose states and behaviors are governed by complex regulatory mechanisms. Understanding how cellular transcriptomes respond to perturbations—such as gene knockouts or drug treatments—across different cell types and conditions is a central problem in biology and medicine.

Although advances in single-cell RNA sequencing and CRISPR-based perturbation technologies have enabled large-scale measurement of transcriptional responses, systematic experimental exploration remains prohibitively expensive and limited in throughput. As a result, only a small fraction of all possible perturbations can be tested in practice.

Early attempts to construct comprehensive "whole-cell" computational models date back to the late 1990s, but were constrained by sparse data, simplified biological assumptions, and limited computational resources. Recently, the availability of large-scale single-cell atlases, rich perturbation datasets, and modern machine learning methods has renewed interest in this direction. In particular, the concept of the AI Virtual Cell (AIVC) has been proposed as a data-driven framework capable of modeling and simulating cellular responses under diverse perturbations and conditions.

Within this context, foundation models for single-cell data, such as scGPT, offer a promising avenue for learning transferable cellular representations and enabling predictive modeling of perturbation responses.

1.2 Problem Setting: Predicting Transcriptomic Responses to Perturbations

In the standard Virtual Cell Challenge (VCC) setting, the task is formulated as predicting transcriptomic responses to perturbations. Given an initial cellular state and a specified perturbation, such as a gene knockout or chemical treatment, the model is required to predict the resulting changes in gene expression induced by the perturbation.

Formally, the input consists of a representation of the pre-perturbation cell state together with a perturbation descriptor, while the output is a high-dimensional vector capturing perturbation-induced gene expression changes across the transcriptome. This task is typically cast as a transcriptome-wide regression problem, where models are trained to minimize errors between predicted and observed expression changes.

Evaluation in this setting commonly relies on correlation-based or mean squared error (MSE)–based metrics, which aggregate prediction errors across all genes. As a result, transcriptome-wide prediction accuracy has become a benchmark for assessing virtual cell models.

1.3 Motivation for Downstream Evaluation

While transcriptome-wide perturbation prediction provides a convenient and standardized proxy task, it does not fully reflect how virtual cell models are used in practice. In many biological and biomedical applications, the primary interest lies in identifying the key genes or pathways driving a perturbation-induced cellular response, rather than accurately predicting expression changes for all genes.

Moreover, transcriptome-wide metrics may obscure biologically meaningful failures. A model can achieve high overall correlation by accurately predicting large numbers of weakly responsive or non-informative genes, while failing to correctly identify the small subset of causal genes that drive the cellular response. This mismatch between evaluation metrics and downstream scientific objectives raises concerns about whether improvements in transcriptome-wide accuracy necessarily translate into greater biological utility.

1.4 Reformulating the Task: Target Gene Identification as a Downstream Evaluation

To address this limitation, we reformulate the evaluation of virtual cell models as a downstream target gene identification task. Given a population of cells subjected to a known perturbation, the model is required to assign scores to candidate genes and rank them according to their likelihood of being true perturbation targets.

This formulation directly evaluates whether a model can correctly prioritize biologically relevant genes whose perturbation induces meaningful cellular changes. Compared to transcriptome-wide regression, target gene identification provides a more structured and lower-noise supervision signal. Ground-truth targets are typically available as binary labels or partial rankings, reducing the influence of non-informative genes whose expression may fluctuate due to technical noise or unrelated biological variation.

As a result, downstream target gene identification offers a complementary and biologically grounded perspective for assessing the practical utility of virtual cell models.

1.5 Our Contributions

In this project, we make the following contributions:

- We reimplemented a more biologically informed set of evaluation metrics for the Virtual Cell Challenge (VCC) and applied them to a subset of 150 known gene perturbations, with a revised train/validation/test split. Using these metrics, we benchmarked the performance of two representative virtual cell models: STATE and scGPT.

- We 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 model predictions.
- We extended our evaluation beyond gene perturbations to drug perturbations and provide an exploratory study on the **Tahoe-100M** dataset, highlighting the challenges.

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 Related Work

2.1 Gene Perturbation Benchmarks

Standardized benchmarks are essential for evaluating virtual cell models. The Virtual Cell Challenge (VCC) provides dedicated single-cell genetic perturbation datasets generated by silencing 300 selected genes in H1 human embryonic stem cells using CRISPR interference. The data include approximately 300,000 single-cell RNA-seq profiles measured with 10x Genomics GEM-X Flex and Illumina sequencing. For the challenge, the dataset is split into training (150 perturbations, 183,000 cells), validation (50 perturbations), and final test (100 held-out perturbations). Perturbations are categorized based on their effect sizes: strong (more than 100 differentially expressed genes), subtle (10-100 DEGs), and negligible (<10 DEGs). This benchmark enables systematic assessment of transcriptome-wide prediction.

The Norman dataset is another commonly used benchmark in single-cell perturbation modeling. In scGPT, it was used for a reverse perturbation prediction task, where the model predicts the source of genetic perturbations given the resulting cell state. For example, a subset of 20 genes is used to construct single- or double-gene perturbation combinations, yielding a closed candidate set of 210 perturbation conditions. The model is fine-tuned on a subset of known perturbations and evaluated by retrieving the top-K candidate perturbations that best match the observed cell states.

2.2 State-of-the-art Models

In this work, we focus on two representative approaches for virtual cell modeling.

2.2.1 STATE

STATE (Single-cell Transformer for Adaptive Transcriptomic Effects) [AGB$^+$] is a transformer-based model designed to predict transcriptomic responses to genetic, signaling, and chemical perturbations while accounting for cellular heterogeneity within and across experiments. It leverages large-scale single-cell data, trained on over 100 million perturbed cells, to capture complex relationships between genes and accurately model perturbation effects.

The model consists of two main modules:

- **State Embedding (SE)**: This module learns high-dimensional representations of single cells using self-supervised training. Each gene is first embedded using a protein language model (ESM2) and then processed via a Transformer to model gene-gene interactions. For each cell, the top high-expression genes are selected to form a sequence, augmented with

expression-weighted embeddings. The output of a [CLS] token is combined with a dataset token ([DS]) to produce the final cell embedding vector.

- **State Transition (ST)**: This module learns how perturbations alter cell states through supervised training. The input includes embeddings of control cells, perturbation conditions, and batch information to account for technical variability. These are combined and passed through a Transformer backbone to predict perturbed gene expression profiles. The model is trained using a Maximum Mean Discrepancy (MMD) loss to align the predicted distribution with the observed distribution of perturbed cells, ensuring that the predicted cellular responses capture realistic biological variation.

STATE has been shown to improve the discrimination of perturbation effects in large datasets and effectively identify differentially expressed genes. Moreover, it can generalize to novel cellular contexts where perturbations were not observed during training, highlighting the advantages of large-scale, data-driven models for virtual cell modeling.

2.2.2 scGPT

scGPT [CWM$^+$24] is a foundation model for single-cell biology...

2.3 Drug Perturbation Datasets

Beyond gene-level perturbations, large-scale datasets such as **Tahoe-100M** provide systematic profiling of drug responses across diverse cell states. Tahoe-100M is currently the largest single-cell perturbation dataset, containing over 100 million cells sampled from 1,344 experimental conditions, including 380 chemical compounds applied at multiple concentrations across approximately 50 cancer cell lines and 14 experimental plates.

The dataset includes cell-level metadata such as experimental plate ID, cell cycle phase, and cell cycle scores (S_score, G2M_score), enabling more detailed analyses of cellular heterogeneity under chemical perturbations. These data allow evaluation of virtual cell models in pharmacological contexts and testing their generalization beyond single-gene interventions.

2.4 Connection to Our Work

Our work builds upon and integrates the methodologies of both STATE and scGPT for practical evaluation and target gene identification. Specifically:

- We utilize the **State Transition (ST)** module from STATE to predict transcriptomic responses, applied to a subset of 150 known gene perturbations from the VCC H1 perturbation dataset. We also implement a revised train/validation/test split to systematically benchmark model performance on these perturbations.
- Using **scGPT**, we finetune the model to leverage the improved VCC evaluation metrics, applied to the same 150 known perturbations with the revised data split. This enables a direct comparison of scGPT and STATE on the VCC benchmark under more biologically informed metrics.
- We further extend scGPT for **target gene identification**, creating scGPT-based models capable of ranking candidate genes by their likelihood of being true perturbation targets. For these downstream tasks, we incorporate the Norman dataset as a source of perturbation data to fine-tune and validate the model.
- Finally, we conduct preliminary exploration on the **Tahoe-100M** drug perturbation dataset to assess the feasibility of applying scGPT to larger-scale pharmacological perturbations, although full target gene prediction on this dataset remains future work.

This progression illustrates the workflow of our study: starting from transcriptome-wide perturbation prediction using STATE's ST module, benchmarking and finetuning scGPT with improved VCC metrics, and moving towards biologically meaningful downstream evaluation in target gene identification.

3 Methods

3.1 Transcriptome-wide Perturbation Prediction

3.1.1 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, 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.

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 without directly accessing expression-level supervision.

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 to generate predicted perturbed transcriptomes.

3.1.2 scGPT

3.2 From Transcriptome-wide Predictions to Target Gene Identification

4 Experiments

4.1 Transcriptome-wide Perturbation Prediction

4.1.1 Dataset and Data Split

Instead of using the original VCC split, we re-split 150 gene perturbations into training, validation, and test sets at the perturbation level to evaluate generalization to unseen perturbations.

Specifically, 64% of perturbation genes are used for training, 16% for validation, and the remaining 20% (30 genes) form a held-out test set. 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 entirely unseen perturbation targets.

4.1.2 Evaluation metrics: MAE, PDS, DES

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 $y_{\mathrm{ntc}} \in \mathbb{R}^G$. Perturbation-induced expression changes are then defined as $\delta_k = y_k - y_{\mathrm{ntc}}$ and $\hat{\delta}_k = \hat{y}_k - y_{\mathrm{ntc}}$. For each predicted perturbation $\hat{\delta}_k$, we compute cosine similarities to all true perturbation deltas $\{\delta_j\}_{j=1}^N$ as

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

These similarities are ranked in descending order, and the rank of the correct perturbation k is denoted by R_k (with 1 being the most similar). The final score is reported as the mean rank $\mathrm{PDS}_{\mathrm{rank}} = \frac{1}{N}\sum_{k=1}^{N} R_k$, or its normalized version $\mathrm{nPDS}_{\mathrm{rank}} = \frac{1}{N^2}\sum_{k=1}^{N} R_k$, where lower values indicate better perturbation discrimination.

MAE on Top 2000 Genes by Ground-Truth Fold Change. To focus evaluation on biologically meaningful signals, we additionally compute the mean absolute error (MAE) restricted to the genes with the largest true perturbation effects. For each perturbation k, we start from the raw (unnormalized) average counts $c_k \in \mathbb{R}^G$ and control counts $c_{\mathrm{ntc}} \in \mathbb{R}^G$, and compute the absolute $\log_2$ fold change for each gene g using a pseudocount of 1:

$$\mathrm{LFC}_{k,g} = \left|\log_2(c_{k,g}+1) - \log_2(c_{\mathrm{ntc},g}+1)\right|.$$

Genes are ranked by $\mathrm{LFC}_{k,g}$ in descending order, and the top 2000 genes form the index set Ω_k. Using log1p-normalized pseudobulk expressions y_k and $\hat{y}_k$ (the same representation as in PDS), the MAE is computed only over these selected genes and then averaged across perturbations:

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

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

4.1.3 Experiment Setup

4.1.4 Main Results

4.2 Target Gene Identification

5 Extending to Drug Perturbations: Tahoe-100M

5.1 Motivation

While our primary experiments focus on gene perturbations, an important long-term goal of virtual cell models is to support drug target discovery in realistic pharmacological settings. Compared to genetic perturbations, drug perturbations introduce additional complexity, including indirect regulatory effects, multi-target mechanisms, and strong cell-line–specific responses.

To explore whether perturbation modeling frameworks can be extended beyond gene-level interventions, we conduct an exploratory study on the Tahoe-100M drug perturbation dataset.

5.2 Data Curation and Target Annotation

The original Tahoe-100M dataset does not provide explicit annotations of drug target genes, which makes supervised target identification infeasible. To enable evaluation, we integrate external drug–target annotations obtained from a curated dataset released on Hugging Face.

Based on the availability and reliability of target annotations, we select a subset of 52 drugs for our experiments. Details of Tahoe-100M data extraction and preprocessing are provided in Appendix A. After preprocessing, the extracted gene expression matrix contains 62,710 genes in total, with approximately 31,597 genes observed in the selected subset.

5.3 Problem Formulation and Data Splitting

We formulate drug target identification as a ranking problem: given drug-treated cells and control cells, the model is expected to prioritize true target genes among all candidate genes.

To reduce information leakage across drugs, we perform a drug-level split. Single- target drugs are retained in the training and validation sets. For multi-target drugs, a drug is included in the test set only if all its target genes have appeared in the training set. Drugs with a large number of targets are kept in training to stabilize learning. In practice, overlap of target genes across different drugs is limited, which further constrains the size of the test set.

5.4 Baseline Modeling and Evaluation

We evaluate drug perturbation effects using two simple baselines. First, we analyze whether true target genes appear among the top 5, 10, 50, or 100 differentially expressed genes. Second, we train a logistic regression classifier using gene-level features derived from expression changes.

Across both evaluations, the baselines fail to reliably identify true target genes, with performance close to random guessing. These results indicate that naïve expression-based approaches are insufficient for drug target discovery in this setting.

5.5 Challenges and Biological Observations

We observe several factors that make drug target identification substantially more challenging than gene perturbation prediction. First, expression changes of target genes are often weak or even absent, potentially because drug effects primarily act on protein activity rather than directly suppressing mRNA levels. Second, the time elapsed after drug treatment is unknown, making it difficult to determine whether the drug has taken effect beyond cell-cycle–related changes.

Moreover, many drugs act on multiple targets and exhibit highly cell-line–specific responses. In our experiments, data from different cancer cell lines are aggregated, which may further obscure target-specific signals. These characteristics suggest that Tahoe-100M reflects the complexity of real-world drug target discovery and poses challenges that are not present in controlled gene perturbation benchmarks.

6 Discussion

7 Conclusion

References

[1]

References

[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.

[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.

Appendix A Tahoe-100M Data Extraction and Preprocessing Pipeline

Appendix A.1 Input Data and Key Files

We extract drug perturbation data from the Tahoe-100M dataset using locally downloaded parquet files (`train*.parquet`). Each record contains gene token IDs, expression values, drug identifiers, cell line IDs, plate information, and sample metadata.

To enable supervision for drug target identification, we integrate drug–target annotations from an external curated dataset released on Hugging Face. In addition, a JSON-based mapping from Tahoe gene token IDs to scGPT vocabulary IDs is used to align the gene space with scGPT-based models. Genes not present in the scGPT vocabulary are discarded.

Appendix A.2 Target Space Construction

Only drugs with at least one annotated target gene are retained. Each target gene is mapped to a categorical label, forming a classification space defined by the union of all observed targets. For multi-target drugs, each (drug, cell) sample is expanded into multiple rows, one per target gene, with identical expression features but different supervision labels. To prevent severe class imbalance, the number of samples per target gene is capped.

Appendix A.3 Control Cell Matching Strategy

Drug-treated cells are paired with control cells treated with DMSO. A control cell is identified if the drug name contains `DMSO`. For efficiency, only the first encountered DMSO-treated cell for each (cell line, plate) key is stored and reused for all corresponding treated cells. This design choice significantly reduces computational overhead while providing a consistent reference state.

Appendix A.4 Gene Space Alignment and Quality Filtering

Since treated and control cells may contain different observed gene sets, control expression vectors are aligned to the gene order of treated cells. Alignment is performed by a two-pointer scan over sorted gene ID lists. Genes present in treated cells but missing in controls are filled with zeros, while control-only genes are discarded.

To ensure sufficient overlap between treated and control gene spaces, we compute an overlap ratio defined as $|G_{\text{treat}} \cap G_{\text{ctrl}}|/|G_{\text{treat}}|$. Samples with overlap ratio below a predefined threshold are discarded, preventing excessive zero-filling in control expressions.

Appendix A.5 Output Dataset Format

The final dataset is stored as a sharded parquet collection. Each row corresponds to a paired (treated, control) sample at single-target resolution. Stored fields include treated gene IDs and expressions, aligned control expressions, drug and target identifiers, cell metadata, and the overlap ratio used for filtering.

Appendix A.6 Key Configuration Parameters

- Control matching key: (`cell_line_id`, `plate`)
- Control fill value: 0.0
- Overlap ratio threshold: 0.05
- Maximum samples per target gene: 100,000