Benchmarking diffusion models for predicting perturbed cellular responses

Zijun Song[1,3] **Changwen Zheng**[3] **Jiangmeng Li**[3] **Linhai Xie**[4,*] **Yujia Xiang**[2,4,*]

[1] School of Advanced Interdisciplinary Sciences, University of Chinese Academy of Sciences, Beijing, China
[2] School of Life Sciences, Tsinghua University, Beijing, China
[3] Institute of Software, Chinese Academy of Sciences, Beijing, China
[4] State Key Laboratory of Medical Proteomics, National Center for Protein Sciences (Beijing),
Research Unit of Proteomics Driven Cancer Precision Medicine (Chinese Academy of Medical Sciences),
Beijing 102206, China
*Corresponding authors: xielinhai@ncpsb.org.cn, yujia.xiang@outlook.com

Abstract

Predicting cellular responses to perturbations is important for understanding biological processes. Although several benchmarks exist, the performance of diffusion models in perturbation prediction has not been systematically studied. Here, we compare the VAE-based model scGen against five diffusion models to assess their suitability for perturbation prediction. Our benchmark covers the prediction of known perturbations, unseen conditions, and stress-test evaluations. The results show that scGen outperforms diffusion models in most settings, while diffusion models demonstrate robustness to noisy data. We also find that encoder design strongly influences model stability and that evaluation metrics can lead to different conclusions. Diffusion models capture the responses of differentially expressed genes, but perform less well on non-DE genes. Overall, our study provides insights into the strengths and limitations of diffusion models, informing their future application in perturbation prediction. Code is available at https://github.com/ZijunSong/PertDiffBench.

1 Introduction

Characterizing how cells respond to perturbations, such as drug treatments, antibody stimulation, or genetic modifications, is crucial for elucidating gene regulatory networks and identifying potential therapeutic targets[3, 20]. The development of high-throughput perturbation sequencing technologies (e.g. CROP-seq and Perturb-seq) has enabled researchers to capture the effects of perturbations on cellular states at single-cell resolution on a large scale[7, 6]. However, such experiments still have limitations in high costs, long timelines, and insufficient coverage of the perturbation space. It is challenging to meet the demands of diverse biological hypothesis validation and large-scale drug screening.

Due to these constraints, computational approaches that simulate cellular responses to perturbations offer a scalable and cost-effective alternative to experimental profiling, enabling broader exploration of hypothetical conditions and paving the way toward constructing a functional virtual cell. These methods are implemented with diverse modeling strategies, for example, scGen, which is based on variational autoencoders, GEARS based on graph neural networks, CellOT based on optimal transport, CellOracle based on linear regression, and STATE based on transformer[1, 11, 4, 21, 15].

Recently, diffusion models have achieved significant breakthroughs in multiple biomedicine-related fields such as image generation, molecular sequence design, and omics data imputation. They have

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Biosecurity Safeguards for Generative AI.

also been applied to cellular perturbation modeling tasks, such as scDiffusion, scVAEder, scDiff, Squidiff, and MorphDiff[25, 22, 17, 24, 10].

Building upon these modeling efforts, several benchmarking studies have emerged. scPerturb and PerturBase provide curated perturbation datasets[26, 18]. Other works investigate specific aspects of perturbation modeling: for example, Ahlmann-Eltze et al. reported that in certain gene perturbation tasks, simple linear models outperform existing deep learning based approaches[2]. PerturBench systematically compared single- and combinatorial-perturbation models and studied the effects of data scaling and imbalance[28]. Pert-Eval explored the use of large language model (LLM)-derived embeddings as inputs to enhance perturbation prediction[27]. Some researchers also conducted a systematic comparison of multiple perturbation modeling methods[14, 13].

Despite these advances, none of the existing benchmarking efforts have systematically evaluated the performance of diffusion models in perturbation modeling tasks, leaving a gap in understanding their robustness, generalization, and potential biosafety risks. To bridge this gap, we present the first benchmark that systematically evaluates diffusion models in cell perturbation modeling tasks, providing a unified framework to test their robustness. Our main contributions include:

(1) Systematically evaluating the predictive stability of diffusion models under varying feature dimensions, and data noise conditions;

(2) Comprehensively testing their generalization capabilities in scenarios involving unseen drug perturbations, unseen cell types, and cross-species contexts;

(3) Analyzing the failure modes of different models in specific scenarios, revealing the limitations of current methods under extreme conditions, and exploring the intrinsic robustness mechanisms required to address biosafety challenges.

2 Related Works

Perturbation modeling. Perturbation modeling in cellular systems aims to predict how cells response to interventions, including genetic perturbations and non-genetic stimuli. Such responses can be characterized using various quantitative readouts, such as changes in cell morphology, alterations in multi-omics profiles, and measurements related to cell viability. Existing approaches to perturbation modeling can be broadly categorized into several families:

(1) **Shallow machine learning models**, such as CellOracle, which use linear regression to infer relationships between transcription factors and target genes, and then predict gene regulatory network changes after perturbing specific transcription factors.

(2) **Deep learning models**

- **Generative modeling approaches**, including VAE-based methods such as scGen, trVAE, and CellCap [16, 29]. Diffusion-based approaches, such as scVAEder, scDiffusion, scDiff, SquiDiff, and MorphDiff, which leverage diffusion models for generating post-perturbation states[24, 10]. Large language model (LLM)-based approaches, including scGPT, scFoundation, and STATE[5, 9].

- **Discriminative modeling approaches**, such as GEARS and GenKI, leveraging graph neural networks for perturbation modeling[30].

(3) **Mathematical and statistical modeling approaches**, such as CellOT, which employs optimal transport to model perturbation effects.

3 Benchmarking Settings

3.1 Overview of benchmarking tasks

We evaluate perturbation modeling methods through three task families:

- **Task 1: Response prediction of known conditions.** This task involves predicting perturbed gene expression from control profiles in both scRNA-seq and bulk RNA-seq settings. We split control–perturbation pairs evenly into training and test sets, and evaluate models by their ability to generate perturbed gene expression profiles for unseen controls in the test set.

• **Task 2: Response prediction of unknown conditions.** This task assesses out-of-distribution generalization across (1) unseen drug perturbations, (2) unseen perturbed cell types, and (3) cross-species settings. Models are trained on control and perturbed gene expression for condition A, and then predict perturbed gene expression for condition B given only its control data.

• **Task 3: Stress-test evaluation.** This task evaluates model robustness under (1) different feature dimensions and (2) different Gaussian noise levels in the input dataset. These scenarios test models' stability and help identify potential failure modes.

3.2 Datasets and preprocessing

All publicly available datasets analyzed in this study are summarized in the provided table. For preprocessing, all datasets were centered and log-normalized.

For the response prediction of known conditions task, we selected Kang18 and Ramaiahgari19 as benchmarks[12, 19]. We evaluate the performance of the model in generating perturbed single-cell gene expression and perturbed bulk gene expression.

For the response prediction of unknown conditions task. In the unseen drug perturbation prediction task, we used the Srivatsan20 dataset, randomly splitting the perturbation data into 70% for training and 30% for testing, and repeated the random split three times[23]. For unseen cell type prediction, we trained models on the Kang18 dataset using both control and perturbed gene expression profiles of CD4 T cells, then predicted post-perturbation responses in natural killer (NK) and B cells. For cross-species prediction, models were trained on mouse perturbation-control data from the Hagai18 dataset and subsequently applied to predict perturbation-induced expression changes in rat, rabbit and pig[8].

For stress-test evaluation, (1) different numbers of highly variable genes (from 1000 to 6000 genes) were selected for CD4 T cells. (2) the gene expression of CD4 T cells was perturbed with Gaussian noise of varying levels, with standard deviations ranging from 0.1 to 1.5.

Table 1: Summary of datasets used for perturbation modeling tasks.

Datasets	Cell lines	Sample size	Perturbation methods	Tasks
Kang18	PBMC	18,868 cells	IFN-β	Task1, 2, 3
Ramaiahgari19	HepaRG	30 samples	Aflatoxin	Task1
Srivatsan20	MCF7	84,674 cells	188 drug perturbations	Task2
Hagai18	Mononuclear phagocytes	77,642 cells	LPS	Task2

3.3 Models selection

We benchmarked six methods, including two self-implemented baselines and four publicly available models. The baselines (DDPM and DDPM+MLP) were implemented to provide reference performance for diffusion-based approaches and to examine how encoder design influences model robustness and generalization. Among the public methods, scGen represents a VAE-based approach widely used in perturbation prediction, while Squidiff, scDiff, and scDiffusion are representative diffusion-based models covering different design paradigms. We excluded scVAEder from comparison due to the lack of sufficient documentation for reproducible execution. For each task, all experiments were repeated three times for each model to ensure robustness of the results.

3.4 Baseline models

Denoising diffusion probabilistic models (DDPM) is a conditional diffusion model designed for perturbation prediction. A key modification from standard DDPMs is the replacement of the U-Net architecture with a Multi-Layer Perceptron (MLP) to better suit the gene expression data. The MLP takes the noisy expression vector and the control cell profile as conditional input to predict the added noise. The model was trained for 1000 epochs using the AdamW optimizer with a learning rate of 1×10^{-5} and a batch size of 512. The diffusion process consists of 1000 timesteps.

DDPM with MLP (DDPM+MLP) operates within a latent space to predict perturbed cellular responses. The model first employs an MLP-based autoencoder to encode the high-dimensional gene expression data into a low-dimensional latent representation (256 dimensions). A conditional DDPM then generates the perturbed latent vector based on the control latent vector. Finally, an MLP decoder maps the generated latent vector back to the gene expression space. The model was trained for 1000 epochs with the AdamW optimizer at a learning rate of 5×10^{-6} and a batch size of 256. The diffusion process uses 1000 timesteps and a linear noise schedule from $\beta_1 = 1 \times 10^{-4}$ to $\beta_T = 5 \times 10^{-3}$.

3.5 Metrics

Three types of metrics were used to evaluate model performance: (1) Error metric: Mean Absolute Error (MAE); (2) Correlation metrics: Pearson Correlation Coefficient (PCC), Delta Pearson Correlation (DPC), and Delta Pearson Correlation DEG (This is a variation of Delta Pearson Correlation, it measures the delta pearson correlation of top differentiation expressed genes); (3) Distributional similarity metric: Maximum Mean Discrepancy (MMD). The detailed information for these metrics is in the supplementary materials.

4 Results

4.1 Response prediction of known conditions

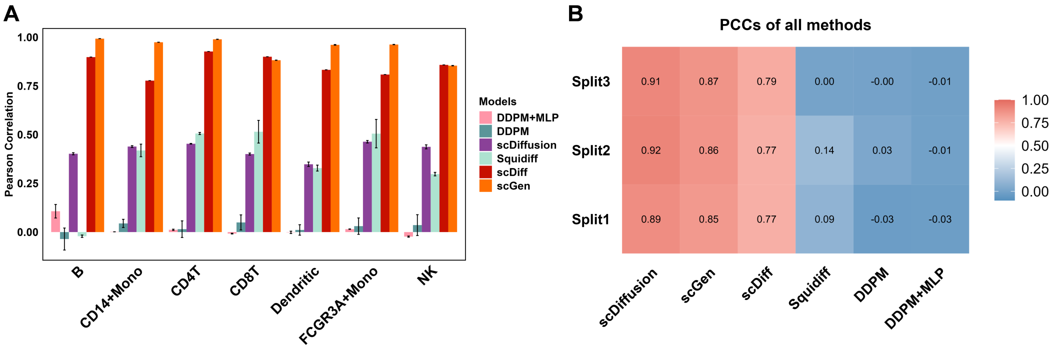

Figure 1: The performance of models on scRNA-seq and bulk RNA-seq. (A) Comparison of mean gene expression between observed and predicted cells using Pearson correlation. (B) Evaluation of models on bulk RNA-seq. Datasets were randomly split into training and testing sets three times.

We assessed the model's performance on scRNA-seq and bulk RNA-seq. The task is primarily to evaluate the model's basic ability to generate perturbed gene expression. In Figure 1A, the results show that scGen and scDiff performed highest on single-cell transcriptomics data, while the DDPM and DDPM+MLP performed the worst. DDPM without MLP encoder exhibited high variance, indicating that an encoder for gene expression representation is important for the predictivity stability of diffusion models.

In Figure 1B, scDiffusion, scGen and scDiff performed well on bulk RNA-seq data. But the performance of squidiff declines rapidly, indicating that its design has limitations in generalizing to bulk RNA-seq.

4.2 Response prediction of unknown conditions

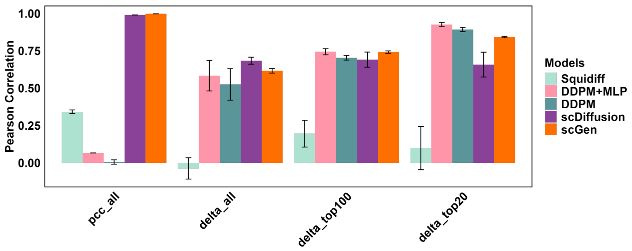

Figure 2: Performance of models on the unseen drug perturbation task, evaluated by PCC (all genes), Delta PCC (all genes), and Delta PCC DEG (top 100 and top 20).

(1) Unseen drug perturbations prediction

We evaluated models for predicting gene expression under unseen drug perturbations. Due to implementation issues, scDiff results were unavailable and therefore excluded from the evaluation. In Figure 2, for the all-gene set Pearson correlation metrics, scGen achieved the highest performance (PCC = 0.997), scDiffusion also performed well (PCC = 0.990), while most other models were close to zero. Interestingly, when considering the Delta Pearson DEG metrics, the performance of DDPM, and DDPM+MLP improved.

These results show that diffusion models capture strong perturbation signals in core DEGs. However, most of them fail to generalize across non-DEG background genes. In contrast, scGen maintains robust performance across the full transcriptome. This robustness is likely due to its latent space regularization and smooth linear shifts.

(2) Unseen cell type prediction

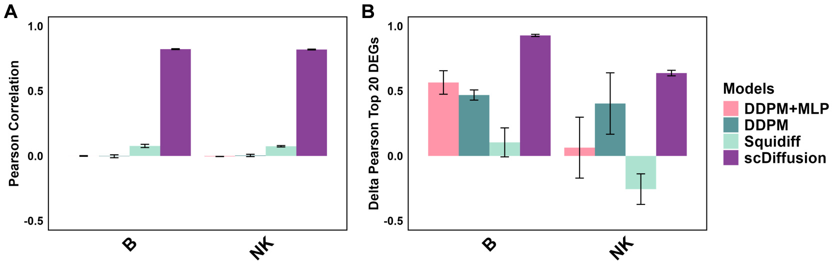

Figure 3: Performance of models on the unseen cell types prediction task.

This task is a zero-shot task. We trained models using both control and perturbed gene expression profiles of CD4 T cells, then predicted post-perturbation responses in natural killer (NK) and B cells. However, scDiff and scGen require the control gene expression of predicted cell types as input for their training stages. Therefore, we excluded scDiff and scGen from our evaluation.

Among the remaining models, scDiffusion achieved the best performance. In contrast, most other models showed near-zero Pearson correlation for all genes, and their Delta Pearson DEG scores were also poor.

(3) Cross-species prediction

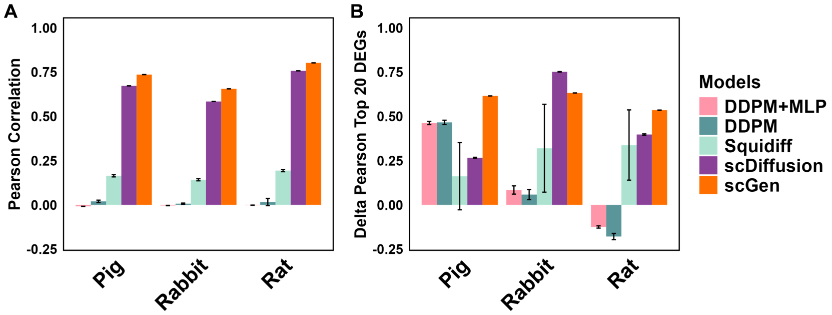

Figure 4: Evaluation of model performance in a cross-species setting.

We then evaluate models on the cross-species scenario. Since scDiff results are not available for this task due to implementation issues, we excluded it from our evaluation. All models were trained on mouse control and perturbed gene expression, then predicted the perturbed gene expression of other species (rat, rabbit and pig).

The results in Figure 4 show that scGen achieved the best performance on both PCC (all genes) and Delta PCC DEG metrics, highlighting its robustness across species boundaries. Among the diffusion models, scDiffusion performed better than the others.

4.3 Stress test evaluation

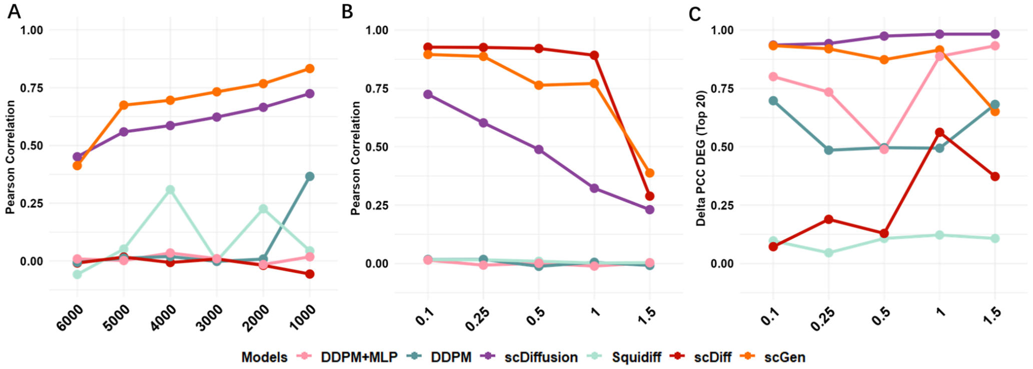

Figure 5: Evaluation of model performance in different feature dimensions and noise levels. (A) Model performance based on the number of selected highly variable genes. (B) and (C) Model performance under different levels of Gaussian noise perturbation. The x-axis shows the standard deviation (sd) of the added Gaussian noise, and the y-axis quantifies the PCC and Delta PCC DEG (top 20) metrics.

(1) Effect of feature selection. As the number of highly variable genes decreases, the performance of all models improves, reflecting a reduction in prediction complexity. The performance of scGen and scDiffusion is stable, while other models are unstable.

(2) Robustness to noise. The performance of all models decreases as the level of Gaussian noise increases. In terms of the Pearson correlation (all genes) metric, scGen and scDiff performed better than the other models. Notably, for the Delta Pearson DEG indicator, scDiffusion's performance gradually improves as the standard deviation of Gaussian noise increases, while scGen's performance gradually declines. At a standard deviation of 1.5, scDiffusion, DDPM+MLP, and DDPM all outperformed scGen. This suggests that the noise addition and denoise process intrinsic to the diffusion model provides them with inherent robustness against noisy data.

5 Discussion

In this study, we benchmarked diffusion models for perturbation modeling task, the results showed that VAE based model scGen outperforms diffusion models in various testing scenarios. However, in noise robustness tests, diffusion models perform better, particularly on the Delta Pearson DEG metric. A likely explanation is that the diffusion process inherently involves diffusion and denoise processes, which help these models recover key signals under noisy conditions. This suggests that diffusion models have potential application value in modeling noisy data where technical noise is unavoidable.

We observed that most diffusion models perform relatively well on the Delta Pearson DEG metric but poorly in predicting non-differentially expressed genes. This result highlights the importance of metric selection in evaluation. In practice, perturbations typically impact a small group of genes. Relying only on the Pearson metric can mislead us about the model's true performance. Therefore, it is important to establish more comprehensive metrics for biological perturbation prediction tasks.

Our experiments further highlight the crucial role of encoder design. The DDPM baseline without an encoder showed high variance and unstable results across tasks, indicating that encoders are critical for effective feature representation and stable training. Recent diffusion-based models have explored different encoder strategies: for instance, scDiffusion leverages pre-trained SCimilarity model to encode single-cell omics data, while squidiff integrates a semantic encoder for omics and temporal information. In future work, integrating prior knowledge with encoders that jointly represent data and perturbation conditions could improve the stability and generalization of diffusion models.

In summary, our study revealed the limits and advantages of diffusion models in perturbation prediction tasks. Choosing appropriate evaluation metrics and improving encoder design will be crucial for better performance, and may guide the development of next-generation models for complex biological perturbations.

Limitations. This study focused on conventional perturbation prediction tasks, future work should evaluate diffusion models under more diverse and biologically realistic perturbation settings. Constructing benchmark datasets that incorporate not only Gaussian noise but also intrinsic biological variability and sequencing artifacts will be critical to rigorously assess model robustness and stability.

6 Acknowledgments

This work was supported by the China Postdoctoral Science Foundation (Grant No. 2025T180417).

References

[1] Abhinav K Adduri et al. "Predicting cellular responses to perturbation across diverse contexts with State". In: *bioRxiv* (2025), pp. 2025–06.

[2] Constantin Ahlmann-Eltze, Wolfgang Huber, and Simon Anders. "Deep learning-based predictions of gene perturbation effects do not yet outperform simple linear baselines". In: *bioRxiv* (2024), pp. 2024–09.

[3] Charlotte Bunne et al. "How to build the virtual cell with artificial intelligence: Priorities and opportunities". In: *Cell* 187.25 (2024), pp. 7045–7063.

[4] Charlotte Bunne et al. "Learning single-cell perturbation responses using neural optimal transport". In: *Nature methods* 20.11 (2023), pp. 1759–1768.

[5] Haotian Cui et al. "scGPT: toward building a foundation model for single-cell multi-omics using generative AI". In: *Nature methods* 21.8 (2024), pp. 1470–1480.

[6] Paul Datlinger et al. "Pooled CRISPR screening with single-cell transcriptome readout". In: *Nature methods* 14.3 (2017), pp. 297–301.

[7] Atray Dixit et al. "Perturb-Seq: dissecting molecular circuits with scalable single-cell RNA profiling of pooled genetic screens". In: *cell* 167.7 (2016), pp. 1853–1866.

[8] Tzachi Hagai et al. "Gene expression variability across cells and species shapes innate immunity". In: *Nature* 563.7730 (2018), pp. 197–202.

[9] Minsheng Hao et al. "Large-scale foundation model on single-cell transcriptomics". In: *Nature methods* 21.8 (2024), pp. 1481–1491.

[10] Siyu He et al. "Squidiff: Predicting cellular development and responses to perturbations using a diffusion model". In: *bioRxiv* (2024), pp. 2024–11.

[11] Kenji Kamimoto et al. "Dissecting cell identity via network inference and in silico gene perturbation". In: *Nature* 614.7949 (2023), pp. 742–751.

[12] Hyun Min Kang et al. "Multiplexed droplet single-cell RNA-sequencing using natural genetic variation". In: *Nature biotechnology* 36.1 (2018), pp. 89–94.

[13] Chen Li et al. "Benchmarking AI models for in silico gene perturbation of cells". In: *bioRxiv* (2024), pp. 2024–12.

[14] Lanxiang Li et al. "A systematic comparison of single-cell perturbation response prediction models". In: *bioRxiv* (2024), pp. 2024–12.

[15] Mohammad Lotfollahi, F Alexander Wolf, and Fabian J Theis. "scGen predicts single-cell perturbation responses". In: *Nature methods* 16.8 (2019), pp. 715–721.

[16] Mohammad Lotfollahi et al. "Conditional out-of-distribution generation for unpaired data using transfer VAE". In: *Bioinformatics* 36.Supplement_2 (2020), pp. i610–i617.

[17] Erpai Luo et al. "scDiffusion: conditional generation of high-quality single-cell data using diffusion model". In: *Bioinformatics* 40.9 (2024), btae518.

[18] Stefan Peidli et al. "scPerturb: harmonized single-cell perturbation data". In: *Nature Methods* 21.3 (2024), pp. 531–540.

[19] Sreenivasa C Ramaiahgari et al. "The power of resolution: contextualized understanding of biological responses to liver injury chemicals using high-throughput transcriptomics and benchmark concentration modeling". In: *Toxicological Sciences* 169.2 (2019), pp. 553–566.

[20] Jennifer E Rood, Anna Hupalowska, and Aviv Regev. "Toward a foundation model of causal cell and tissue biology with a Perturbation Cell and Tissue Atlas". In: *Cell* 187.17 (2024), pp. 4520–4545.

[21] Yusuf Roohani, Kexin Huang, and Jure Leskovec. "Predicting transcriptional outcomes of novel multigene perturbations with GEARS". In: *Nature Biotechnology* 42.6 (2024), pp. 927–935.

[22] Mehrshad Sadria and Anita Layton. "scVAEDer: integrating deep diffusion models and variational autoencoders for single-cell transcriptomics analysis". In: *Genome Biology* 26.1 (2025), pp. 1–17.

[23] Sanjay R Srivatsan et al. "Massively multiplex chemical transcriptomics at single-cell resolution". In: *Science* 367.6473 (2020), pp. 45–51.

[24] Wenzhuo Tang et al. "A general single-cell analysis framework via conditional diffusion generative models". In: *bioRxiv* (2023), pp. 2023–10.

[25] Xuesong Wang et al. "Prediction of cellular morphology change under perturbations with transcriptome-guided diffusion model". In: *bioRxiv* (2025), pp. 2025–07.

[26] Zhiting Wei et al. "PerturBase: a comprehensive database for single-cell perturbation data analysis and visualization". In: *Nucleic Acids Research* 53.D1 (2025), pp. D1099–D1111.

[27] Aaron Wenteler et al. "Perteval-scfm: benchmarking single-cell foundation models for perturbation effect prediction". In: *bioRxiv* (2024), pp. 2024–10.

[28] Yan Wu et al. "PerturBench: Benchmarking Machine Learning Models for Cellular Perturbation Analysis". In: *NeurIPS 2024 Workshop on AI for New Drug Modalities*. 2024. URL: `https://openreview.net/forum?id=1EJXtK3AWr`.

[29] Yang Xu et al. "Explainable modeling of single-cell perturbation data using attention and sparse dictionary learning". In: *Cell Systems* 16.4 (2025).

[30] Yongjian Yang et al. "Gene knockout inference with variational graph autoencoder learning single-cell gene regulatory networks". In: *Nucleic Acids Research* 51.13 (2023), pp. 6578–6592.

Supplementary Materials

A Models

scDiffusion builds on a conditional diffusion model framework. It includes three parts: a pre-trained foundation model SCimilarity as the autoencoder to embed the gene expression, a denoising network with full connected layers and a skip-connected structure as the backbone, and a classifier for conditional generation. Default parameters were used for scDiffusion training.

scDiff is a conditional diffusion model which consists of four components: an embedder to embed input gene expression, multiple conditioners to extract representation of different conditions, a crossattention encoder and a linear decoder. We used the default parameters to train scDiff on different tasks.

Squidiff is a conditional denoising diffusion implicit model with a semantic encoder. The semantic encoder is a multilayer perceptron (MLP), which utilized a residual connection to embed both time and semantic features.

scGen is a variational autoencoder based model. It learns cell state shifts in latent space and uses a decoder to reconstruct gene expression, enabling accurate prediction of unseen perturbations.

B Metrics

(1) Error Metric: These metrics directly measure the discrepancy between the predicted values and the ground truth values.

Mean Absolute Error (MAE): MAE calculates the average of the absolute prediction errors. It provides a robust measure of the mean deviation between predicted and true values and is less sensitive to outliers.

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

where y_i and $\hat{y}_i$ denote the observed and predicted values for the i-th sample, respectively, and n is the number of samples.

(2) Correlation Metrics: These metrics assess the degree to which the patterns and trends in the predicted gene expression profiles align with the ground truth.

Pearson Correlation Coefficient (PCC): This metric measures the linear correlation between the overall predicted gene expression profile and the true expression profile. A PCC value closer to 1 indicates a higher consistency between the model's predictions and the actual data.

$$\text{PCC}(Y, \hat{Y}) = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2 \sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}})^2}} \tag{2}$$

where Y and $\hat{Y}$ are the vectors of true and predicted expression values, and $\bar{y}$ and $\bar{\hat{y}}$ are their respective means.

Delta Pearson Correlation (DPC): To evaluate the model's ability to capture the trend of gene expression changes induced by a perturbation, we introduced DPC. This metric specifically calculates the Pearson correlation between the predicted change in gene expression ($\Delta_{\text{pred}} = \hat{Y}_{\text{pert}} - Y_{\text{ctrl}}$) and the true change ($\Delta_{\text{pred}} = Y_{\text{pert}} - Y_{\text{ctrl}}$).

Delta Pearson Correlation on DEGs (DPC DEG): This is a crucial variant of DPC designed to focus on key biological signals. We first identify the top differentially expressed genes (DEGs) based on the ground truth data, and then compute the DPC exclusively on this subset of genes. This metric more accurately measures the model's predictive performance on the core perturbation effects.

(3) Distributional Similarity Metric:

Maximum Mean Discrepancy (MMD): MMD measures the distance between two probability distributions. In this project, we use it to assess the similarity between the distribution of model-generated gene expressions and the distribution of true gene expressions. A smaller MMD value indicates that the predicted data distribution is closer to the true data distribution, demonstrating the model's ability to better learn the intrinsic data structure. The MMD is estimated as:

$$\text{MMD}(Y, \hat{Y}) = \sqrt{\frac{1}{m^2}\sum_{i=1}^{m}\sum_{j=1}^{m} k(y_i, y_j) + \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} k(\hat{y}_i, \hat{y}_j) - \frac{2}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n} k(y_i, \hat{y}_j)} \tag{3}$$

where k is a kernel function (e.g., the Radial Basis Function kernel), y_i and $\hat{y}_i$ denote the i-th true and predicted samples, respectively, and Y and $\hat{Y}$ represent the corresponding sets of samples.

C Additional benchmark results

C.1 Response prediction of known conditions

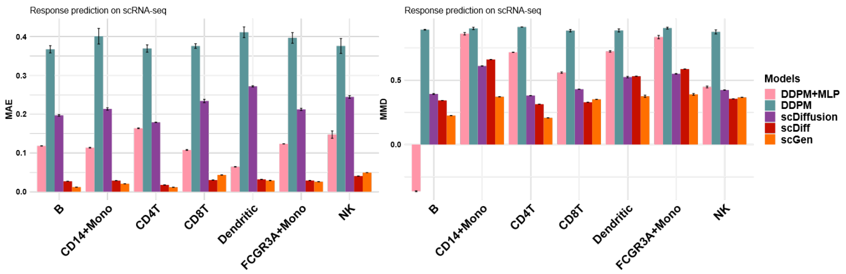

Figure 6: Evaluation of model performance in known conditions prediction, evaluated by MAE and MMD.

Since the MAE value of Squidiff was extremely high (MAE is from 2.260 to 65.750), it was not included in the figure for clarity.

C.2 Performance of models on the unseen drug perturbation task

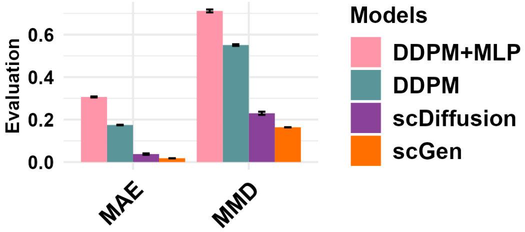

Figure 7: Performance of models on the unseen drug perturbation task, evaluated by MAE and MMD.

Since the MAE value of Squidiff was extremely high (MAE = 40.547), it was not included in the figure for clarity.

C.3 Performance of models on the unseen cell type prediction task

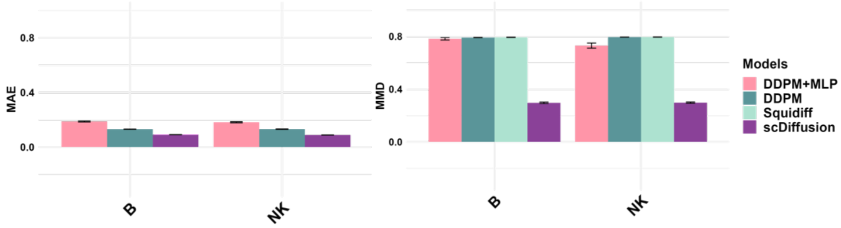

Figure 8: Performance of models on the unseen cell type prediction task, evaluated by MAE and MMD.

Since the MAE value of Squidiff was extremely high (MAE nearly 56), it was excluded from the figure for clarity.

C.4 Performance of models on the cross-species prediction task

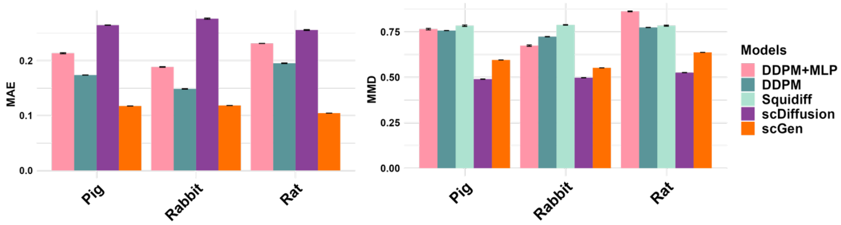

Figure 9: Performance of models on the cross-species prediction task, evaluated by MAE and MMD.

Since the MAE value of Squidiff was extremely high (MAE nearly 55), it was excluded from the figure for clarity.