scTREAT: Latent Diffusion for Single-Cell Drug-Response Prediction

Anonymous authors
Paper under double-blind review

Abstract

Generating drug-induced transcriptional responses requires a model to account for cellular background together with molecular identity and dose. We present scTREAT, the first diffusion framework to factorize state and drug guidance for single-cell drug-response prediction. A shared latent expression space connects baseline and response, while Guided Decomposable Attention (GD-Attn) injects a baseline representation and a dose-conditioned molecular embedding. Independent channel dropout exposes four conditioning states during training; inference combines three conditional predictions with separately specified state and drug coefficients, and dose modulates drug-guidance strength. On condition-level pseudobulk profiles derived from the Tahoe-100M single-cell atlas, scTREAT achieves the best value on 10 of 12 metrics for unseen cell line–drug combinations (UC) and 11 of 12 for unseen drugs (UD). Ablations examine condition injection, molecular representation, and guidance design. On UD, the relative gains over the second-best model reach 36.11% and 34.21% for DEG logFC-Spearman and logFC-Pearson. These results demonstrate the effectiveness of combining shared latent expression modeling with explicit state and drug conditioning for response prediction.

1 Introduction

Predicting transcriptional responses to unmeasured drug treatments requires modeling how cellular background, molecular identity, and dose jointly shape expression. Single-cell perturbation screens (Klein et al., 2015; Datlinger et al., 2017; Srivatsan et al., 2020) connect compounds to transcriptional readouts and support mechanism-of-action (MoA) analysis and compound evaluation (Lamb et al., 2006; Subramanian et al., 2017; Trapotsi et al., 2022; Ye et al., 2018; Corsello et al., 2020). Yet the cost of screening limits coverage of the cell line $\times$ drug $\times$ dose space (Mayr & Bojanic, 2009; Macarron et al., 2011; Iversen et al., 2012). Conditional response generation offers a way to model the effects of candidate molecules across these unmeasured settings.

The modeling question is how to combine cellular background with treatment information and retain access to their contributions during inference. Latent perturbation models, molecular conditioning, and Transformers offer complementary approaches (Lotfollahi et al., 2019; Hetzel et al., 2022; Qi et al., 2024; Adduri et al., 2025). Diffusion-based methods have also established conditional generation and latent manipulation for single-cell data (Tang et al., 2023; Luo et al., 2024; Sadria & Layton, 2025; He et al., 2026; Shang et al., 2025). Within this line of work, we study a shared conditional generator whose state and drug guidance can be weighted separately at inference.

Approach and contributions. We present scTREAT (Figure 1) and make two contributions:

- **The first diffusion framework to factorize state and drug guidance for single-cell drug-response prediction.** A shared latent space and GD-Attn combine baseline expression with a dose-conditioned molecular representation. Four-state training supplies conditional predictions for factorized inference, with dose also setting drug-guidance strength.
- **Evidence from ablation and generalization.** We examine condition injection, molecular representation, and guidance through ablations, alongside UC/UD prediction on unobserved combinations and held-out drugs.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

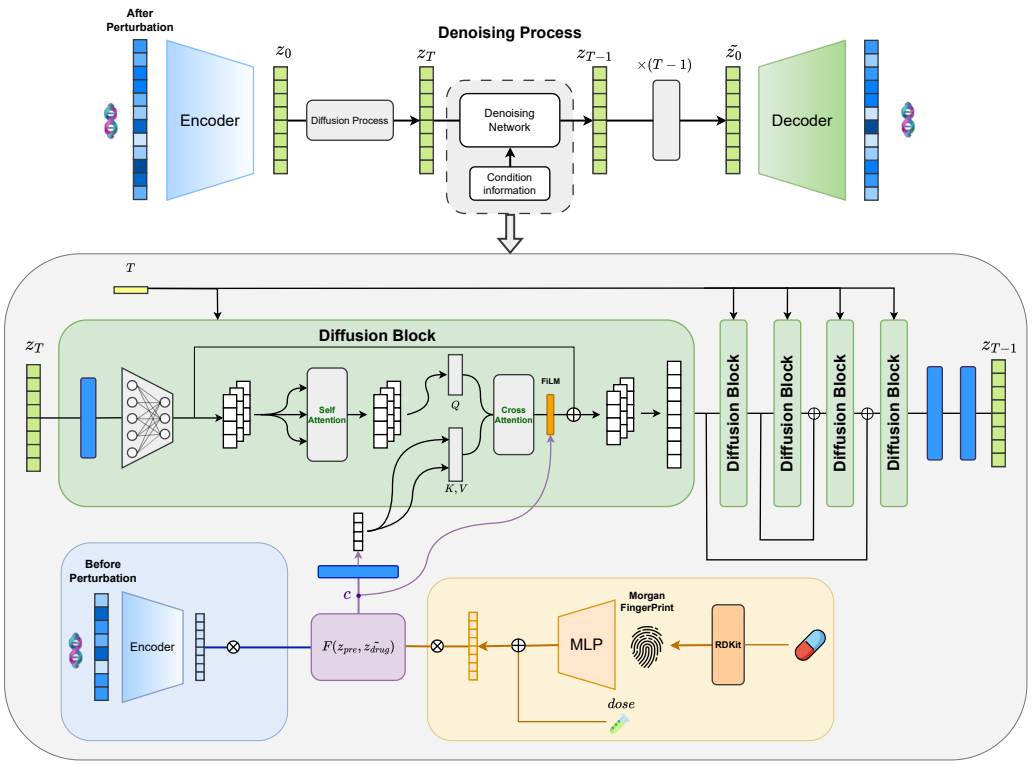

Figure 1: Overview of scTREAT. Top: Expression is encoded in a shared autoencoder latent space; the denoising path summarizes neural noise prediction and reverse-sampling updates to $\hat{z}_0$, followed by decoding to $\hat{x}$. Bottom: The matched-control representation $z_{\text{pre}}{=}E(x_{\text{pre}})$ and dose-fused drug vector $\tilde{z}_{\text{drug}}$ from Morgan fingerprints + MLP (Rogers & Hahn, 2010; Landrum et al.) form a condition token c injected through GD-Attn (Section 3.2).

2 Related Work

Deep learning for single-cell perturbation prediction. Latent-space generative approaches perform counterfactual prediction by mapping expression into a shared latent space: scGen uses latent vector arithmetic across cell types (Lotfollahi et al., 2019); CPA represents cellular background and dose as additive components (Lotfollahi et al., 2021); chemCPA further incorporates molecular structure encoders and transfer from bulk HTS to handle previously unseen compounds at single-cell resolution (Hetzel et al., 2022; Subramanian et al., 2017). Structure-conditional generators combine chemical structure (e.g., SMILES/fingerprints Weininger (1988); Schwartz et al. (2013)), dose, and control expression to predict responses for untested compounds (Qi et al., 2024). Large-scale Transformer pretraining learns transferable gene-expression representations that can serve as initialization or encoders for downstream prediction and generation (Cui et al., 2024). Another line of latent representation learning is exemplified by BioLORD (Piran et al., 2024), which trains cross-modal encoders linking molecular structure and transcriptomic responses, enabling generalization to previously unseen compounds; in contrast, SAMS-VAE (Bereket & Karaletsos, 2024) models perturbation effects only for the perturbations observed during training and does not support unseen-drug prediction. As an alternative to latent generative modeling, optimal-transport (OT) methods formulate control to perturbation alignment from unpaired observations to capture population-level shifts (Bunne et al., 2023; Dong et al., 2023). A hybrid line combines large Transformers with OT, exemplified by STATE (Adduri et al., 2025), though its generative formulation does not support prediction for previously unseen compounds.

Diffusion and transferable techniques from computer vision. In computer vision, diffusion provides an effective, scalable recipe for conditional generation: latent-space denoising with cross-attention (Rombach et al., 2022); controllable sampling via classifier or classifier-free guidance and

compositional guidance (Ho & Salimans, 2022; Liu et al., 2022); Transformer backbones with feature-wise modulation (AdaLN/FiLM) (Vaswani et al., 2017; Peebles & Xie, 2023; Perez et al., 2018); and control adapters for selective layerwise conditioning (Zhang et al., 2023; Mou et al., 2023). More broadly, diffusion injects noise via a forward SDE and denoises via the reverse SDE or the probability-flow ODE (Song et al., 2020b), with accelerated non-Markovian sampling such as DDIM (Song et al., 2020a). For single-cell expression, scDiffusion (Luo et al., 2024) combines latent-space denoising with a pretrained encoder and uses classifier guidance and gradient-based conditional interpolation at inference. Related diffusion approaches model perturbations through conditional generation or latent-space manipulation: scDiff formulates multiple single-cell tasks under a conditional diffusion model; scVAEDer combines a variational autoencoder with latent diffusion; Squidiff uses semantic latent representations for developmental, genetic, and drug perturbations; and dbDiffusion conditions latent generation on perturbation embeddings (Tang et al., 2023; Sadria & Layton, 2025; He et al., 2026; Shang et al., 2025). Together, these studies establish diffusion as a viable tool for single-cell perturbation modeling. scTREAT builds on this foundation by combining shared latent conditioning with four-state training and an explicit rule for separate state and drug guidance. Its methodological focus is this combination and its evaluation through component ablations.

Data resource. Large-scale resources such as LINCS/L1000 and Tahoe-100M support response modeling across compounds, doses, and cellular contexts. We use Tahoe-100M (Zhang et al., 2025), whose public release contains more than 100 million cells across 50 cancer cell lines and 379 drugs.

3 METHOD

scTREAT models post-treatment expression conditioned on a baseline profile, molecular structure, and dose. A shared expression space connects baseline and response profiles; GD-Attn injects state and drug conditions into the denoiser; and four-state training supplies the predictions used for factorized guidance (Figure 1).

3.1 LATENT-SPACE DIFFUSION BACKBONE

We model the post-perturbation expression distribution in an observational manner as the conditional mapping

$$p(x_{\text{post}} \mid x_{\text{pre}}, m, d), \tag{1}$$

where $x_{\text{pre}}, x_{\text{post}} \in \mathbb{R}^G$ denote the pre- and post-perturbation expression over G genes, m denotes the drug's molecular structure (SMILES), and d denotes dose.

A SCimilarity autoencoder (Heimberg et al., 2023), fine-tuned following scDiffusion (Luo et al., 2024), provides an encoder E and decoder D that define a latent space of dimension D_z in which the response dynamics are learned. Once fine-tuned, E and D are frozen. Post-treatment condition-level profiles are mapped to $z_0 = E(x_{\text{post}})$ for training, and the decoder reconstructs predictions via $\hat{x} = D(\hat{z}_0)$ after denoising.

Within this latent schema, a variance-preserving diffusion process injects Gaussian noise along a schedule $\{\beta_t\}_{t=1}^T$ with $\alpha_t = 1 - \beta_t$ and cumulative product $\bar{\alpha}_t = \prod_{s \le t} \alpha_s$. The forward transition is

$$q(z_t \mid z_0) = \mathcal{N}\big(\sqrt{\bar{\alpha}_t}\, z_0,\ (1 - \bar{\alpha}_t)\, \mathbf{I}\big). \tag{2}$$

The denoiser B_θ receives the noisy latent z_t, a sinusoidal time embedding $\gamma(t) \in \mathbb{R}^{d_t}$, and condition c, and predicts the Gaussian noise $\hat{\epsilon}_t = B_\theta\big(z_t, \gamma(t), c\big)$. GD-Attn injects c within the denoising blocks; the effect of this conditioning design is examined by the reported ablations.

3.2 CONDITION REPRESENTATION AND GD-ATTN INJECTION

We encode the pre-perturbation state with the shared expression encoder and build a dose-aware, structure-informed drug vector before fusing both through GD-Attn.

State encoding. We reuse the same fine-tuned expression encoder E of the backbone on the baseline state conditioning branch.

$$z_{\text{pre}} = E^{\text{cond}}(x_{\text{pre}}) \in \mathbb{R}^{D_{\text{pre}}}, \quad D_{\text{pre}} = D_z \tag{3}$$

z_{pre} is the conditioning-side latent for the baseline state; D_{pre} matches the backbone latent dimension D_z; $E^{\text{cond}} \equiv E$ indicates full parameter sharing.

The shared encoder provides a deterministic embedding in the same coordinates as the training target:

$$z_{\text{pre}}^{\text{cond}} = E(x_{\text{pre}}) = z_{\text{pre}}. \tag{4}$$

Parameter sharing places baseline conditions and post-treatment profiles in a common expression space.

Drug encoding. Molecular structure informs the response representation, while dose modulates the drug condition. SMILES are converted to 1024-bit Morgan fingerprints $\phi \in \{0,1\}^{1024}$ and embedded via an MLP to obtain the structure-aware vector

$$z_{\text{drug}} = \text{MLP}(\phi) \in \mathbb{R}^{D_{\text{drug}}} \tag{5}$$

where D_{drug} is the dimensionality of the learned drug embedding. Dose is fused through a FiLM module with a continuous encoding $e_{\text{dose}} \in \mathbb{R}^{D_{\text{dose}}}$, where D_{dose} denotes the dose-embedding dimension.

$$\tilde{z}_{\text{drug}} = \text{FiLM}\left(z_{\text{drug}},\, e_{\text{dose}}\right) \tag{6}$$

Here FiLM applies dose-dependent feature-wise scaling and shifting to z_{drug}. Molecular information and dose jointly determine $\tilde{z}_{\text{drug}}$; this design allows dose to alter the representation without assuming an exact separation of response direction and magnitude.

Condition injection. Guided Decomposable Attention (GD-Attn) performs the non-concatenative injection. The composite condition is

$$c = F\big(z_{\text{pre}},\, \tilde{z}_{\text{drug}}\big) \in \mathbb{R}^{D_c} \tag{7}$$

F is the fusion function and D_c is the dimensionality of the condition.

GD-Attn refines latent tokens through self-attention, injects the fused condition through cross-attention to a single condition token, and applies FiLM modulation and a residual feed-forward update with LayerNorm. This pathway supplies cellular background and treatment information throughout denoising. Channel masking provides the condition subsets used for separate state/drug guidance.

3.3 Four-State Training

Four-state training supplies the conditional predictions needed for separate state and drug guidance at inference. We apply independent dropout to $\{z_{\text{pre}},\, \tilde{z}_{\text{drug}}\}$ in the training process to obtain the four possible combinations.

$$c' \in \{\text{both},\ \text{no-pre},\ \text{no-drug},\ \varnothing\} \tag{8}$$

Unified training objective is shown below.

Injection-side alignment and contrast. For each sample i, define the normalized empirical response direction

$$\hat{v}^{(i)} = \frac{x_{\text{post}}^{(i)} - x_{\text{pre}}^{(i)}}{\max\big(\|x_{\text{post}}^{(i)} - x_{\text{pre}}^{(i)}\|_2, \varepsilon_v\big)}. \tag{9}$$

We project the dose-fused embedding $\tilde{z}_{\text{drug}}$ with a learnable matrix $W \in \mathbb{R}^{G \times D_{\text{drug}}}$ and normalize the projected direction as

$$\hat{u}^{(i)} = \frac{W\, \tilde{z}_{\text{drug}}^{(i)}}{\max\big(\|W\, \tilde{z}_{\text{drug}}^{(i)}\|_2, \varepsilon_u\big)}.$$

Small positive constants ε_v and ε_u safeguard normalization. We align $\hat{u}^{(i)}$ to $\hat{v}^{(i)}$ while maintaining contrast with other response embeddings. The alignment loss is

$$\mathcal{L}_{\text{align}} = \big\|\hat{v}^{(i)} - \hat{u}^{(i)}\big\|_2^2. \tag{10}$$

To encourage discriminative drug–response representations, we use an InfoNCE term (van den Oord et al., 2018; Chen et al., 2020). With $s(u, v) = \langle u, v \rangle / \tau$ and negative response embeddings $\mathcal{N}_i$, the objective is

$$\mathcal{L}_{\text{info}} = \mathbb{E}_{i \in \mathcal{B}} \left[-\log \frac{\exp\big(s(\hat{u}^{(i)}, \hat{v}^{(i)})\big)}{\exp\big(s(\hat{u}^{(i)}, \hat{v}^{(i)})\big) + \sum_{v^- \in \mathcal{N}_i} \exp\big(s(\hat{u}^{(i)}, v^-)\big)} \right]. \tag{11}$$

Here τ is the temperature and $\mathcal{B}$ is the minibatch. In the in-batch form, $\mathcal{N}_i$ contains the other samples' response embeddings. The implementation also provides a module that draws negatives from a queue of detached response embeddings from previous batches, falling back to in-batch negatives while the queue is empty. This objective contrasts matched drug–response pairs with negative response samples.

The injection regularizer combines them as

$$\mathcal{L}_{\text{inject}} = \lambda_{\text{align}} \mathcal{L}_{\text{align}} + \lambda_{\text{info}} \mathcal{L}_{\text{info}} \tag{12}$$

Denoising loss under four-state training. With independent channel dropout c', the diffusion denoiser is trained by the MSE on noise:

$$\mathcal{L}_{\text{denoise}} = \mathbb{E}_{z_0,\, t,\, \epsilon,\, c'} \left[\big\| \epsilon - \epsilon_\theta(z_t, t, c') \big\|_2^2 \right] \tag{13}$$

where $z_t \sim q(z_t \mid z_0)$ follows the VP forward process and ϵ_θ predicts the noise.

Total objective. Our overall objective weights the denoising criterion with the injection regularizer:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{denoise}} + \mathcal{L}_{\text{inject}} \tag{14}$$

with $\lambda_{\text{align}}, \lambda_{\text{info}} \geq 0$ as hyperparameters.

3.4 Factorized Guidance

At inference, we combine three conditional predictions. The coefficients s_p and s_d weight the state and drug guidance terms, respectively:

$$\begin{aligned} \hat{\epsilon} = \; & (1 + s_p + s_d)\, \epsilon(\text{both}) - s_p\, \epsilon(\text{no-pre}) - s_d\, \epsilon(\text{no-drug}) \\ & s_p, s_d \geq 0 \end{aligned} \tag{15}$$

$\epsilon(\cdot)$ denotes the network output under each condition. The unconditional state participates in training but is not evaluated in this inference rule. Equation (15) starts from $\epsilon(\text{both})$ and adds the state contrast $\epsilon(\text{both}) - \epsilon(\text{no-pre})$ weighted by s_p and the drug contrast $\epsilon(\text{both}) - \epsilon(\text{no-drug})$ weighted by s_d. Thus, three predictions from the same denoiser suffice to construct the two guidance terms.

We set drug-guidance strength continuously from dose:

$$s_d(\text{dose}) = s_0 \cdot \sigma\big(\alpha \log(1 + d) + \gamma\big), \qquad \sigma(z) = \tfrac{1}{1+e^{-z}} \tag{16}$$

Here d is dosage and s_0, α, γ are hyperparameters. The log-dose sigmoid (Sebaugh, 2011) is continuous at zero and prevents unbounded amplification at high doses; the inference flow is shown in Figure 2, with additional details in Section B.

4 Experiments

We evaluate response prediction on unseen cell line–drug combinations (UC) and unseen drugs (UD), then examine condition injection, molecular representation, and guidance through component ablations.

4.1 Benchmark and Training Protocol

Data and splits. Tahoe-100M (Zhang et al., 2025) contains 50 cancer cell lines and 379 drugs in its public release. We average single-cell expression within each (cell line, drug, dose, plate) group and match every non-DMSO condition to its same-plate DMSO control. UC holds out complete cell line–drug pairs and UD holds out complete drugs, grouping all doses and plates together (Section C.2).

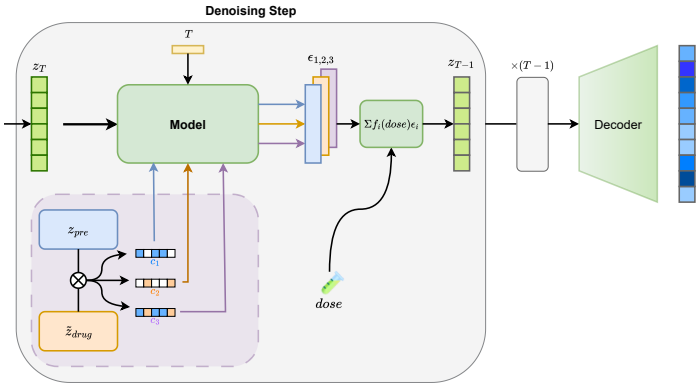

Figure 2: Conditional denoising from z_T. Three noise predictions from a four-state-trained denoiser are combined with separate state/drug guidance for each reverse-sampler update to z_{t-1}. Decoding the final $\hat{z}_0$ gives the predicted expression.

Metrics. Metrics cover logFC and Δ correlations, median explained variance, and response-gene overlap. DEG denotes the Top-200 genes by observed absolute logFC unless specified otherwise. In a global paired pseudobulk analysis, **84% (UC)** and **79% (UD)** of the Top-200 genes satisfy FDR ≤ 0.05 after one-sample t-tests on per-pair logFC and Benjamini–Hochberg correction across all genes (Section D).

Baselines. Our model and other models are trained on the same set and evaluated on UC and UD test sets. All models use the same split, and each is trained to convergence. Baselines include Linear Regression and Multi-layer Perceptron (MLP) with drug encodings implemented analogously to PRnet (Qi et al., 2024) adapter. The Perturb (Adduri et al., 2025) baseline averages observed perturbation offsets for the same perturbation in training data and adds mean offset to control expression to form prediction. The Context (Adduri et al., 2025) baseline averages perturbed expression across samples within the same cell line/context and uses this mean as prediction directly. We further compare six deep-learning baselines: chemCPA (Hetzel et al., 2022), an additive latent-space model with molecular structure encoding and dose scaling, supporting tests on both UC and UD; PRnet (Qi et al., 2024), an encoder–decoder model conditioned on drug perturbations via adapter that generates fingerprint-based embeddings, supporting both tasks above; BioLORD (Piran et al., 2024) trains cross-modal encoders linking molecular structure and transcriptomic profiles and therefore supports both UC and UD evaluation; SAMS-VAE (Bereket & Karaletsos, 2024) models perturbation labels as discrete conditions without molecular-structure encoding and thus supports UC evaluation but cannot generalize to unseen drugs and is not evaluated on UD; STATE (Adduri et al., 2025), a model that takes unperturbed transcriptome together with the perturbation label, evaluated only on UC and not supporting UD prediction; and scGPT (Cui et al., 2024), a Transformer-based single-cell foundation model used as perturbation-prediction baseline, also evaluated only on UC without supporting UD prediction. All models, including baseline methods, are trained and evaluated on the same pseudobulked condition-level pairs for fairness.

Training. Only training-set profiles are used to fine-tune the SCimilarity autoencoder and train the diffusion model. Diffusion is optimized with AdamW (Loshchilov & Hutter, 2017); the learning rate warms up for 3,000 steps and then decays linearly. All models use the same split and condition-level pairs and are trained to convergence (Sections C.1 and 4.1).

4.2 Prediction on Unseen Combinations and Drugs

The full model attains the best value on 10 of 12 UC metrics and 11 of 12 UD metrics (Tables 1 and 2). In UC, representative gains over the second-best model are **+13.46%** on DEG Δ-Spearman and **+12.00%** on DEG logFC-Pearson. In UD, DEG logFC-Spearman/Pearson reaches

Table 1: UC results. Higher is better; bold and underline mark the best and second-best distinct values, with ties sharing formatting. Correlations and median explained variance (EV) are reported for all genes and the Top-200 genes by observed absolute logFC (DEG). P and S denote Pearson and Spearman correlations. Overlap@200 and Overlap@1000 denote the corresponding DEG-accuracy metrics.

All genes

Model	logFC-P	logFC-S	Δ-P	Δ-S	EV_{median}
scTREAT	**0.81**	**0.69**	<u>0.55</u>	<u>0.51</u>	**0.73**
chemCPA	0.70	0.46	**0.59**	0.44	<u>0.70</u>
PRnet	0.61	<u>0.58</u>	0.35	0.27	0.51
STATE	<u>0.77</u>	**0.69**	0.50	0.50	0.69
scGPT	0.46	0.39	0.39	0.35	0.41
Linear	0.50	0.51	0.48	0.41	0.20
MLP	0.55	0.57	0.44	0.36	0.40
<u>Perturb</u>	0.46	0.38	0.21	0.25	0.48
<u>Context</u>	0.48	0.32	0.25	0.24	0.39
BioLORD	0.50	0.34	0.46	0.37	0.63
SAMS-VAE	0.54	0.57	0.53	**0.55**	0.53

DEG and response-gene overlap

Model	logFC-P	logFC-S	Δ-P	Δ-S	EV_{median}	Overlap @200	Overlap @1000
scTREAT	**0.56**	**0.49**	**0.57**	**0.59**	**0.64**	**0.24**	**0.39**
chemCPA	0.40	0.38	0.52	0.46	<u>0.59</u>	<u>0.22</u>	0.28
PRnet	0.31	0.28	0.36	0.28	0.37	0.11	0.21
STATE	<u>0.50</u>	<u>0.48</u>	<u>0.54</u>	<u>0.52</u>	0.53	<u>0.22</u>	<u>0.32</u>
scGPT	0.32	0.23	0.38	0.33	0.34	0.11	0.19
Linear	0.38	0.34	0.45	0.41	0.18	0.14	0.19
MLP	0.37	0.23	0.43	0.36	0.30	0.17	0.18
<u>Perturb</u>	0.35	0.33	0.22	0.24	0.38	0.16	0.17
<u>Context</u>	0.35	0.32	0.24	0.25	0.33	0.13	0.15
BioLORD	0.43	0.29	0.38	0.31	0.51	0.21	0.30
SAMS-VAE	0.44	0.45	0.46	0.49	0.44	0.20	0.29

0.49/0.51 versus 0.36/0.38 for chemCPA (**+36.11%/+34.21%**). Top-1000 DEG-Accuracy improves by **+21.88%/+24.14%** in UC/UD. Together with the ablations, these comparisons connect the conditioning and guidance design to its performance on the response-generation task.

The gains are not uniform: chemCPA leads all-gene Δ-Pearson in UC and UD, while SAMS-VAE leads all-gene Δ-Spearman in UC. The results therefore support broad improvements among the evaluated criteria while retaining distinctions between recovery of fold changes, perturbation shifts, and ranked response genes.

4.3 Ablation Study

We examine how the conditioning and guidance design affects prediction (Figure 3). Removing conditioning collapses performance; replacing GD-Attn weakens fit and Δ alignment; replacing molecular structure with Drug-ID most strongly harms UD; single-channel guidance lowers several metrics; and constant s_d reduces predictive accuracy on selected criteria. Full configurations are in Section F.

4.4 Qualitative Prediction

The qualitative comparison complements the aggregate metrics by showing that predicted profiles retain cell-line structure and reproduce representative gene-wise logFC distributions.

Table 2: UD results. Methods without unseen-drug support are omitted. Higher is better; bold and underline mark the best and second-best distinct values, with ties sharing formatting. Metric abbreviations and the DEG subset follow Table 1.

All genes

Model	logFC-P	logFC-S	Δ-P	Δ-S	EV_{median}
scTREAT	**0.67**	**0.58**	0.50	**0.49**	**0.66**
chemCPA	0.42	0.44	**0.53**	0.43	0.59
PRnet	0.59	0.48	0.33	0.25	0.42
Linear	0.43	0.49	0.46	0.40	0.18
MLP	0.40	0.36	0.43	0.34	0.32
Context	0.31	0.32	0.42	0.39	0.40
BioLORD	0.42	0.30	0.45	0.37	0.60

DEG and response-gene overlap

Model	logFC-P	logFC-S	Δ-P	Δ-S	EV_{median}	Overlap @200	Overlap @1000
scTREAT	**0.51**	**0.49**	**0.53**	**0.52**	**0.56**	**0.22**	**0.36**
chemCPA	0.38	0.36	0.48	0.41	0.47	0.20	0.28
PRnet	0.30	0.28	0.34	0.28	0.31	0.07	0.19
Linear	0.32	0.26	0.40	0.37	0.11	0.12	0.17
MLP	0.29	0.26	0.41	0.33	0.25	0.14	0.16
Context	0.31	0.32	0.43	0.39	0.30	0.15	0.15
BioLORD	0.30	0.26	0.29	0.27	0.51	0.19	0.29

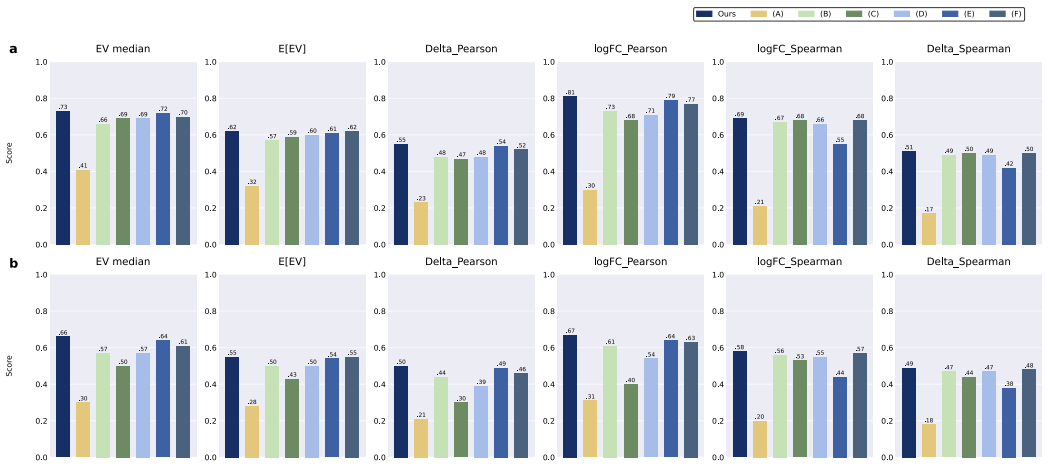

Figure 3: Ablations on UC (a) and UD (b). Conditioning, GD-Attn, and separate guidance support predictive performance; molecular structure supports UD generalization; dose-dependent guidance improves selected metrics. Variants: (A) no conditioning; (B) GD-Attn replaced by Concat+MLP; (C) molecular structure replaced by Drug-ID; (D) single-channel CFG; (E) constant drug-guidance strength; (F) concatenation of condition and latent tokens before attention. Full configurations are in Section F.

4.5 Biological Context

Pathway analysis. The KRAS example in Figure 5 and the PI3K–AKT–mTOR example in Figure 6 show similar broad running-enrichment curve shapes: both predicted and observed profiles rise near the beginning of the ranked list and decline through its middle. These shared features provide a qualitative view of pathway-associated responses. The amplitudes and dominant signed deviations differ, however: the observed curves have larger positive peaks, whereas the predicted curves have larger negative troughs. Under the standard signed-ES interpretation, this limits agreement in

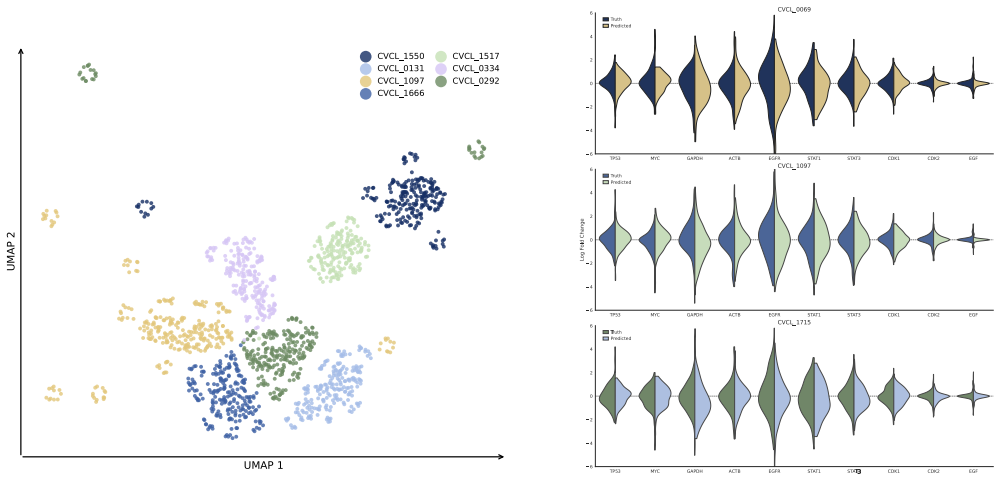

Figure 4: Qualitative response prediction. Predicted profiles separate by cell line in UMAP space (left), while predicted and observed gene-wise logFC distributions agree for representative cell lines (right).

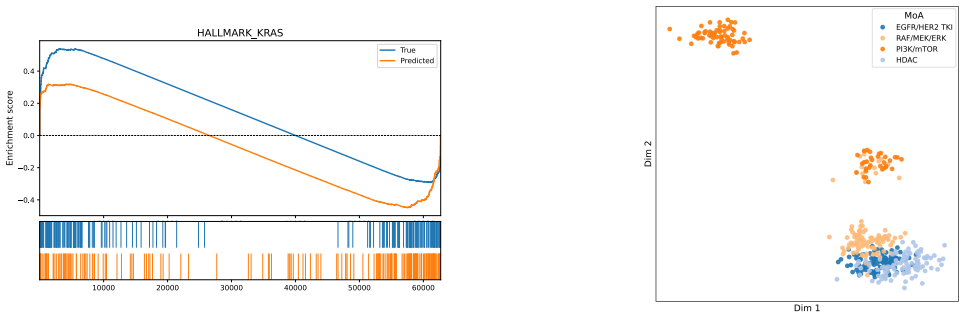

Figure 5: Biological context. Left: KRAS GSEA curves share a broad shape but differ in their dominant signed deviations. Right: drug representations show MoA-associated grouping with overlap between some classes.

enrichment direction, so the examples support partial correspondence rather than complete pathway-response recovery.

MoA-associated representation structure. The MoA-colored UMAP in Figure 5 shows pharmacologically associated organization, including a distinct region enriched for PI3K/mTOR compounds and localized groupings of other annotated classes. EGFR/HER2 and HDAC compounds also overlap. These patterns provide positive qualitative evidence of MoA-associated structure in the learned drug representation, without requiring complete separation of the annotated classes. Because molecular fingerprints are model inputs, this view alone does not establish mechanistic information beyond the supplied structural features.

5 Discussion and Conclusion

scTREAT is the first diffusion framework to factorize state and drug guidance for single-cell drug-response prediction. It combines shared latent expression modeling, non-concatenative condition injection, and four-state training for this inference design. Ablations and UC/UD comparisons connect the design to predictive performance, with competing methods retaining advantages on selected Δ metrics. The results support latent diffusion for predicting transcriptional responses to unmeasured combinations and drugs, while the qualitative analyses provide complementary views of cellular and pharmacological structure.

AI USE STATEMENT

AI tools were used to polish writing, support literature retrieval and discovery, and assist with part of the research execution.

REPRODUCIBILITY STATEMENT

The model formulation and training objective are specified in Section 3, with baseline details in Section 4.1. The appendix describes data construction and split grouping (Section C.2), training (Section C.1), evaluation metrics (Section D), and ablation configurations (Section F).

An anonymized ZIP archive containing the core code has been submitted as supplementary material.

REFERENCES

Abhinav K. Adduri, Dhruv Gautam, Beatrice Bevilacqua, Alishba Imran, Rohan Shah, Mohsen Naghipourfar, Noam Teyssier, Rajesh Ilango, Sanjay Nagaraj, Mingze Dong, Chiara Ricci-Tam, Christopher Carpenter, Vishvak Subramanyam, Aidan Winters, Sravya Tirukkovular, Jeremy Sullivan, Brian S. Plosky, Basak Eraslan, Nicholas D. Youngblut, Jure Leskovec, Luke A. Gilbert, Silvana Konermann, Patrick D. Hsu, Alexander Dobin, Dave P. Burke, Hani Goodarzi, and Yusuf H. Roohani. Predicting cellular responses to perturbation across diverse contexts with state. *bioRxiv*, 2025. doi: 10.1101/2025.06.26.661135. URL https://www.biorxiv.org/content/early/2025/07/10/2025.06.26.661135.

Michael Bereket and Theofanis Karaletsos. Modelling cellular perturbations with the sparse additive mechanism shift variational autoencoder. *arXiv preprint arXiv:2311.02794*, 2024. doi: 10.48550/arXiv.2311.02794. URL https://arxiv.org/abs/2311.02794.

Charlotte Bunne, Stefan G. Stark, Gabriele Gut, Jacobo Sarabia Del Castillo, Mitch Levesque, Kjong-Van Lehmann, Lucas Pelkmans, Andreas Krause, and Gunnar Rätsch. Learning single-cell perturbation responses using neural optimal transport. *Nature Methods*, 20(11):1759–1768, 2023. doi: 10.1038/s41592-023-01969-x. URL https://www.nature.com/articles/s41592-023-01969-x.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 1597–1607. PMLR, 2020. URL http://proceedings.mlr.press/v119/chen20j.html.

S. M. Corsello et al. Discovering the anti-cancer potential of non-oncology drugs by systematic viability profiling. *Nature Cancer*, 1(2):235–248, 2020. doi: 10.1038/s43018-019-0018-6. URL https://www.nature.com/articles/s43018-019-0018-6.

Haotian Cui, Chloe Wang, Hassaan Maan, et al. scGPT: toward building a foundation model for single-cell multi-omics using generative AI. *Nature Methods*, 21:1470–1480, 2024. doi: 10.1038/s41592-024-02201-0. URL https://www.nature.com/articles/s41592-024-02201-0.

Paul Datlinger, André F. Rendeiro, Christian Schmidl, Thomas Krausgruber, Peter Traxler, Johanna Klughammer, Linda C. Schuster, Amelie Kuchler, Donat Alpar, and Christoph Bock. Pooled crispr screening with single-cell transcriptome readout. *Nature Methods*, 14(3):297–301, 2017. doi: 10.1038/nmeth.4177. URL https://www.nature.com/articles/nmeth.4177.

Mingze Dong, Bao Wang, Jessica Wei, Antonio H. de O. Fonseca, Curtis J. Perry, Alexander Frey, Feriel Ouerghi, Ellen F. Foxman, Jeffrey J. Ishizuka, Rahul M. Dhodapkar, and David van Dijk. Causal identification of single-cell experimental perturbation effects with CINEMA-OT. *Nature Methods*, 20(11):1769–1779, 2023. doi: 10.1038/s41592-023-02040-5. URL https://www.nature.com/articles/s41592-023-02040-5.

Siyu He, Yuefei Zhu, Daniel Naveed Tavakol, Haotian Ye, Yeh-Hsing Lao, Zixian Zhu, Cong Xu, Shradha Chauhan, Guy Garty, Raju Tomer, Gordana Vunjak-Novakovic, James Zou, Elham Azizi,

and Kam W. Leong. Squidiff: Predicting cellular development and responses to perturbations using a diffusion model. *Nature Methods*, 23(1):65–77, 2026. doi: 10.1038/s41592-025-02877-y. URL https://doi.org/10.1038/s41592-025-02877-y.

Graham Heimberg, Tony Kuo, Daryle DePianto, Tobias Heigl, Nathaniel Diamant, Omar Salem, Gabriele Scalia, Tommaso Biancalani, Jason Rock, Shannon Turley, Héctor Corrada Bravo, Josh Kaminker, Jason A. Vander Heiden, and Aviv Regev. Scalable querying of human cell atlases via a foundational model reveals commonalities across fibrosis-associated macrophages. *bioRxiv*, 2023. doi: 10.1101/2023.07.18.549537. URL https://www.biorxiv.org/content/early/2023/07/19/2023.07.18.549537.

Leon Hetzel, Simon Böhm, Niki Kilbertus, Stephan Günnemann, Mohammad Lotfollahi, and Fabian J Theis. Predicting cellular responses to novel drug perturbations at a single-cell resolution. In *NeurIPS 2022*, 2022.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv:2207.12598*, 2022. URL https://arxiv.org/abs/2207.12598.

Philip W. Iversen, Benoit Beck, Yun-Fei Chen, Walthere Dere, Viswanath Devanarayan, Brian J. Eastwood, Mark W. Farmen, Stephen J. Iturria, Chahrzad Montrose, Roger A. Moore, Jeffrey R. Weidner, and G. Sitta Sittampalam. HTS assay validation. In *Assay Guidance Manual*. Eli Lilly & Company and the National Center for Advancing Translational Sciences, Bethesda (MD), 2012. URL https://www.ncbi.nlm.nih.gov/books/NBK83783/.

Allon M. Klein, Linas Mazutis, Ilke Akartuna, Naren Tallapragada, Adrian Veres, Victor Li, Leonid Peshkin, David A. Weitz, and Marc W. Kirschner. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell*, 161(5):1187–1201, 2015. doi: 10.1016/j.cell.2015.04.044. URL https://doi.org/10.1016/j.cell.2015.04.044.

J. Lamb et al. The connectivity map: using gene-expression signatures to connect small molecules, genes, and disease. *Science*, 313(5795):1929–1935, 2006. doi: 10.1126/science.1132939.

Greg Landrum et al. Rdkit: Open-source cheminformatics. https://www.rdkit.org. Accessed: 2025-09-22.

Nan Liu, Shuang Li, Yilun Du, Antonio Torralba, and Joshua B. Tenenbaum. Compositional visual generation with composable diffusion models. In *European Conference on Computer Vision (ECCV)*, 2022. URL https://arxiv.org/abs/2206.01714.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2017. URL https://arxiv.org/abs/1711.05101.

Mohammad Lotfollahi, F. Alexander Wolf, and Fabian J. Theis. scGen predicts single-cell perturbation responses. *Nature Methods*, 16(8):715–721, 2019. doi: 10.1038/s41592-019-0494-8. URL https://www.nature.com/articles/s41592-019-0494-8.

Mohammad Lotfollahi, Anna Klimovskaia Susmelj, Carlo De Donno, Yuge Ji, Ignacio L. Ibarra, F. Alexander Wolf, Nafissa Yakubova, Fabian J. Theis, and David Lopez-Paz. Learning interpretable cellular responses to complex perturbations in high-throughput screens. *bioRxiv*, 2021. doi: 10.1101/2021.04.14.439903. URL https://www.biorxiv.org/content/early/2021/05/18/2021.04.14.439903.

Erpai Luo, Minsheng Hao, Lei Wei, and Xuegong Zhang. scDiffusion: conditional generation of high-quality single-cell data using diffusion model. *Bioinformatics*, 40(9):btae518, 2024. doi: 10.1093/bioinformatics/btae518. URL https://academic.oup.com/bioinformatics/article/40/9/btae518/7738782.

Ricardo Macarron, Martyn N. Banks, Dejan Bojanic, David J. Burns, Dragan A. Cirovic, Tina Garyantes, Darren V. S. Green, Robert P. Hertzberg, William P. Janzen, Jeff W. Paslay, Ulrich Schopfer, and G. Sitta Sittampalam. Impact of high-throughput screening in biomedical research. *Nature Reviews Drug Discovery*, 10(3):188–195, 2011. doi: 10.1038/nrd3368. URL https://www.nature.com/articles/nrd3368.

Lorenz M. Mayr and Dejan Bojanic. Novel trends in high-throughput screening. *Current Opinion in Pharmacology*, 9(5):580–588, 2009. doi: 10.1016/j.coph.2009.08.004. URL https://pubmed.ncbi.nlm.nih.gov/19775937/.

Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, Ying Shan, and Xiaohu Qie. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. *arXiv:2302.08453*, 2023. URL https://arxiv.org/abs/2302.08453.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. URL https://openaccess.thecvf.com/content/ICCV2023/papers/Peebles_Scalable_Diffusion_Models_with_Transformers_ICCV_2023_paper.pdf.

Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *AAAI Conference on Artificial Intelligence*, 2018. URL https://arxiv.org/abs/1709.07871.

Zoe Piran, Niv Cohen, Yedid Hoshen, and Mor Nitzan. Disentanglement of single-cell data with biolord. *Nature Biotechnology*, 42(11):1678–1683, 2024. doi: 10.1038/s41587-023-02079-x. URL https://doi.org/10.1038/s41587-023-02079-x.

Xiaoning Qi, Lianhe Zhao, Chenyu Tian, Yueyue Li, Zhen-Lin Chen, Peipei Huo, Runsheng Chen, Xiaodong Liu, Baoping Wan, Shengyong Yang, and Yi Zhao. Predicting transcriptional responses to novel chemical perturbations using deep generative model for drug discovery. *Nature Communications*, 15:9256, 2024. doi: 10.1038/s41467-024-53457-1. URL https://www.nature.com/articles/s41467-024-53457-1.

David Rogers and Mathew Hahn. Extended-connectivity fingerprints. *Journal of Chemical Information and Modeling*, 50(5):742–754, 2010. doi: 10.1021/ci100050t. URL https://doi.org/10.1021/ci100050t.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Rombach_High-Resolution_Image_Synthesis_With_Latent_Diffusion_Models_CVPR_2022_paper.html.

Mehrshad Sadria and Anita Layton. scVAEDer: Integrating deep diffusion models and variational autoencoders for single-cell transcriptomics analysis. *Genome Biology*, 26(1):64, 2025. doi: 10.1186/s13059-025-03519-4. URL https://doi.org/10.1186/s13059-025-03519-4.

J. Schwartz, M. Awale, and J.-L. Reymond. Smifp (smiles fingerprint) chemical space for virtual screening and visualization of large databases of organic molecules. *Journal of Chemical Information and Modeling*, 53(8):1979–1989, August 2013. doi: 10.1021/ci400206h. URL https://pubmed.ncbi.nlm.nih.gov/23845040/.

J. L. Sebaugh. Guidelines for accurate EC50/IC50 estimation. *Pharmaceutical Statistics*, 10(2):128–134, 2011. doi: 10.1002/pst.426. URL https://doi.org/10.1002/pst.426.

Ergan Shang, Yuting Wei, and Kathryn Roeder. Predicting the unseen: A diffusion-based debiasing framework for transcriptional response prediction at single-cell resolution. *Proceedings of the National Academy of Sciences*, 122(52):e2525268122, 2025. doi: 10.1073/pnas.2525268122. URL https://doi.org/10.1073/pnas.2525268122.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020a. doi: 10.48550/arXiv.2010.02502. URL https://arxiv.org/abs/2010.02502. ICLR 2021.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020b. doi: 10.48550/arXiv.2011.13456. URL https://arxiv.org/abs/2011.13456. ICLR 2021 (Oral).

Sanjay R. Srivatsan, José L. McFaline-Figueroa, Vijay Ramani, Lauren Saunders, Junyue Cao, Jonathan Packer, Hannah A. Pliner, Dana L. Jackson, Riza M. Daza, Lena Christiansen, Fan Zhang, Frank Steemers, Jay Shendure, and Cole Trapnell. Massively multiplex chemical transcriptomics at single-cell resolution. *Science*, 367(6473):45–51, 2020. doi: 10.1126/science.aax6234. URL https://www.science.org/doi/10.1126/science.aax6234.

Aravind Subramanian, Rajiv Narayan, Steven M. Corsello, David D. Peck, Ted E. Natoli, Xiaodong Lu, Joshua Gould, John F. Davis, Andrew A. Tubelli, Jacob K. Asiedu, David L. Lahr, Jodi E. Hirschman, Zihan Liu, Melanie Donahue, Bina Julian, Mariya Khan, David Wadden, Ian C. Smith, Daniel Lam, Arthur Liberzon, Courtney Toder, Mukta Bagul, Marek Orzechowski, Oana M. Enache, Federica Piccioni, Sarah A. Johnson, Nicholas J. Lyons, Alice H. Berger, Alykhan F. Shamji, Angela N. Brooks, Anita Vrcic, Corey Flynn, Jacqueline Rosains, David Y. Takeda, Roger Hu, Desiree Davison, Justin Lamb, Kristin Ardlie, Larson Hogstrom, Peyton Greenside, Nathanael S. Gray, Paul A. Clemons, Serena Silver, Xiaoyun Wu, Wen-Ning Zhao, Willis Read-Button, Xiaohua Wu, Stephen J. Haggarty, Lucienne V. Ronco, Jesse S. Boehm, Stuart L. Schreiber, John G. Doench, Joshua A. Bittker, David E. Root, Bang Wong, and Todd R. Golub. A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. *Cell*, 171(6):1437–1452.e17, 2017. doi: 10.1016/j.cell.2017.10.049. URL https://www.cell.com/cell/fulltext/S0092-8674(17)31309-0.

Wenzhuo Tang, Renming Liu, Hongzhi Wen, Xinnan Dai, Jiayuan Ding, Hang Li, Wenqi Fan, Yuying Xie, and Jiliang Tang. A general single-cell analysis framework via conditional diffusion generative models. *bioRxiv*, 2023. doi: 10.1101/2023.10.13.562243. URL https://doi.org/10.1101/2023.10.13.562243.

M.-A. Trapotsi, L. Hosseini-Gerami, and A. Bender. Computational analyses of mechanism of action (moa): data, methods and integration. *RSC Chemical Biology*, 3(2):170–200, 2022. doi: 10.1039/D1CB00069A.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. URL https://arxiv.org/abs/1807.03748.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pp. 5998–6008. Curran Associates, Inc., 2017. URL https://arxiv.org/abs/1706.03762.

David Weininger. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of Chemical Information and Computer Sciences*, 28(1):31–36, 1988. doi: 10.1021/ci00057a005. URL https://pubs.acs.org/doi/10.1021/ci00057a005.

C. Ye et al. Drug-seq for miniaturized high-throughput transcriptome profiling in drug discovery. *Nature Communications*, 9(1):4307, 2018. doi: 10.1038/s41467-018-06500-x. URL https://www.nature.com/articles/s41467-018-06500-x.

Jesse Zhang, Airol A. Ubas, Richard de Borja, Valentine Svensson, Nicole Thomas, Neha Thakar, Ian Lai, Aidan Winters, Umair Khan, Matthew G. Jones, et al. Tahoe-100m: A giga-scale single-cell perturbation atlas for context-dependent gene function and cellular modeling. *bioRxiv*, 2025. doi: 10.1101/2025.02.20.639398. URL https://www.biorxiv.org/content/10.1101/2025.02.20.639398v1. Preprint.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. URL https://openaccess.thecvf.com/content/ICCV2023/papers/Zhang_Adding_Conditional_Control_to_Text-to-Image_Diffusion_Models_ICCV_2023_paper.pdf.

A Additional Context

Experimental motivation. While enabling parallel observation of drug effects at some scale, traditional high-throughput screening (HTS) workflows (Mayr & Bojanic, 2009; Macarron et al., 2011; Iversen et al., 2012) are extremely costly in terms of reagents, personnel time, and turnaround. Their throughput is also limited, making it infeasible to cover all potential cell line $\times$ drug $\times$ dose combinations; even within large public perturbation atlases, measured conditions constitute only a tiny fraction of the combinatorial space. Perturbation maps can additionally support the exploration of precision therapeutic strategies.

B Additional Inference Details

The coefficients s_p and s_d separately weight the state and drug contrast terms in Equation (15). Both terms share the fully conditioned prediction, so this interface does not require biological independence of the two conditions. Dose enters both the FiLM representation and the bounded log-dose sigmoid in Equation (16); the reported ablation measures predictive performance, without establishing monotonic gene-level responses.

The sampler starts from a latent state z_T. At each step, three neural noise predictions are combined by Equation (15), and the reverse sampler updates the latent. The expression encoder is used to obtain the baseline condition, and the decoder maps the final $\hat{z}_0$ back to expression.

C Training and Data Construction

C.1 Training Setup

Following scDiffusion (Luo et al., 2024), we adapt the SCimilarity autoencoder (Heimberg et al., 2023) to the expression dimension and fine-tune it on training-set profiles with a reconstruction objective. We then freeze the encoder and decoder for latent diffusion training. Optimization uses AdamW (Loshchilov & Hutter, 2017); the diffusion learning rate warms up for 3,000 steps and then decays linearly. The held-out test sets are excluded from autoencoder fine-tuning, model training, and model selection.

C.2 Dataset Construction and Splits

We use the Tahoe-100M (Zhang et al., 2025) public single-cell drug-perturbation atlas as our data source. This dataset is at hundred-million scale, spans 50 cancer cell lines, includes 379 drugs in its public release, and is compatible with the conditional mapping $(x_{\text{pre}}, \text{drug}, \text{dose}) \to x_{\text{post}}$ that we aim to learn. We take $(\text{cell_line}, \text{drug}, \text{dose}, \text{plate})$ as conditioning keys, averaging single-cell expression within each key group to obtain the corresponding representative state. For each non-DMSO condition, we identify the matched DMSO control on the same plate to establish a one-to-one pair, yielding the pre-perturbation transcriptome x_{pre} and post-perturbation transcriptome x_{post}.

UC holds out complete cell line–drug combinations: all doses and plates for a held-out combination remain outside training, while its cell line and drug each appear individually in training. UD holds out complete drugs across all cell lines, doses, and plates. These grouping rules prevent the held-out combination or drug from re-entering training through another dose or plate.

D Evaluation Metrics

We rank all genes by absolute $\log_2$ fold change ($|\log_2 \text{FC}|$) under the perturbation and take the top-k genes as the high-impact set. For all DEG-based evaluations except (4), we fix $k = 200$. We retain DEG as shorthand for this effect-size-selected response-gene subset; selection does not require passing an FDR threshold. We evaluate the following metrics, each computed on both the full gene set and the DEG subset: (1) Pearson and Spearman correlations between the ground-truth and predicted log fold-change vectors ($\log_2$ FC); (2) Pearson and Spearman correlations between the ground-truth and predicted perturbation-shift vectors ($\Delta := x_{\text{post}} - x_{\text{pre}}$); (3) median explained variance of the

predictions with respect to the ground truth; and (4) DE-Overlap-Accuracy, for which we evaluate several values of k.

As an additional statistical check, we conducted a global paired pseudobulk analysis separately for UC and UD. For each gene, we performed a one-sample t-test on its per-pair logFC values against a zero mean, followed by Benjamini–Hochberg correction across all genes. Among the Top-200 genes ranked by absolute logFC in this analysis, **84% (UC)** and **79% (UD)** satisfy $\mathrm{FDR} \leq 0.05$. Thus, a large majority of the selected genes also pass the significance threshold. This global check complements the effect-size-based evaluation; it does not redefine the evaluated subset or test performance differences between models.

E Additional Pathway Example

Figure 6 compares the running enrichment scores for the PI3K–AKT–mTOR pathway computed from observed and predicted responses. Both curves rise near the beginning of the ranked list before declining. The predicted curve has a smaller positive peak and a deeper negative trough than the observed curve. Tick marks below the curves show where pathway genes occur in each ranked list. This example complements the KRAS analysis in Figure 5, illustrating similarities in curve shape as well as differences in enrichment magnitude and pathway-gene ranking.

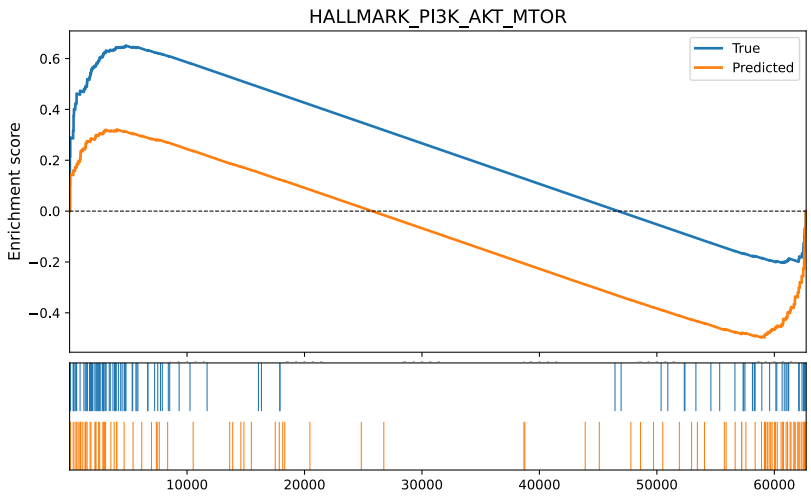

Figure 6: Exploratory GSEA comparison for HALLMARK_PI3K_AKT_MTOR_SIGNALING. Predicted and observed curves share broad shape features but differ in amplitude and dominant signed deviation.

F Complete Ablation Configuration

We examine five design factors through six ablation variants: (A) remove all conditioning (No-Cond); (B) replace GD-Attn (Section 3.2) with Concat+MLP; (C) replace the Morgan/SMILES-based drug encoder (Equation (5)) with Drug-ID; (D) replace decomposable dual-knob guidance (s_p, s_d) (Equation (15)) with single-channel CFG; (E) replace the dose-to-strength mapping $s_d(\mathrm{dose})$ (Equation (16)) with a constant; and (F) replace the non-concatenative GD-Attn injection (Section 3.2) with an input-level concatenation before attention, which concatenates the fused condition token with the latent tokens prior to the attention blocks.

Across UC/UD, these ablations examine conditioning, GD-Attn injection, structure-aware drug encoding, separate guidance, and the dose-to-strength mapping. GD-Attn improves several fit and Δ-alignment metrics relative to the concatenation variants, while separate and dose-dependent guidance

improve selected predictive criteria. These comparisons assess prediction quality rather than directly measuring training stability or dose monotonicity; structure-aware priors are especially important for UD generalization.

G Additional Discussion

The methodological focus of scTREAT is the connection between condition representation and the inference interface. State and drug information enter a shared denoiser through GD-Attn. Independent channel dropout trains the model under four conditioning states, and three conditional predictions are then combined using separately specified state and drug coefficients. This formulation makes the guidance rule explicit while retaining a common response-generation backbone.

The ablations examine different parts of this design: availability of conditions, the injection pathway, molecular representation, and the guidance rule. The UC/UD comparisons provide the task-level evaluation of the resulting generator. The advantages of competing models on selected all-gene Δ metrics show that the benefits depend on which aspects of the response are measured.