scCycleMol: Modeling Cell-Cycle-Aware Single-Cell Drug Perturbation Responses

Dingping Zhao*
Division of Pharmacognosy, School of Pharmaceutical Sciences,
State Key Laboratory of Natural and Biomimetic Drugs, Peking University, China

Jie Lin†
Department of Computer Science at School of Informatics,
Xiamen University, Xiamen, China

Abstract

Single-cell drug perturbation models should predict not only transcriptional response magnitude, but also whether a treatment changes the proliferative state of the cell. This is difficult because cell-cycle variation is often treated as a nuisance factor, and benchmark processing rarely makes drug-induced phase changes a first-class prediction target. We introduce scCycleMol, a cell-cycle-aware perturbation prediction framework built on a curated 24-hour SciPlex3 benchmark with standardized molecule identities, dose and cell-line metadata, modeled genes, and expression-derived cell-cycle supervision. Instead of using cell-cycle state as an input covariate, scCycleMol derives supervision from the treated state and applies it to the predicted treated expression. The main model uses a learnable full-expression cell-cycle head with circular G1/S/G2M phase targets, so the auxiliary loss can backpropagate through the decoder, dose-response module, and molecular drug representation. We also evaluate marker-score supervision, molecular representation choices, and pretraining sources to isolate where the gains arise. On a processed 24-hour SciPlex3 benchmark with 635,541 cells, 186 non-control molecule-level perturbation units, 188 valid label-level compound embeddings, three cancer cell lines, four nonzero treatment doses plus DMSO control, and 5,080 modeled genes, scCycleMol improves out-of-distribution expression metrics over included conditional perturbation reference baselines. The best LINCS-pretrained circular variant reaches 0.9093 expected all-gene r^2 and 0.6843 expected DE-gene r^2, compared with reference values of 0.6800 and 0.5400 for LINCS-pretrained ChemCPA. Closed-loop cell-cycle supervision raises phase accuracy by 0.54–0.62 absolute points while keeping expected all-gene r^2 within 0.003 of matched no-cell-cycle models; Tahoe-pretrained circular supervision reaches 0.9609 phase accuracy.

1 Introduction

Drug perturbations do not merely shift average gene expression. They can arrest cells, accelerate transitions, or redistribute a population across cell-cycle phases. Live-cell studies of anti-cancer agents show drug- and dose-specific changes in cell-cycle phasing [5], while long-term single-cell imaging under cisplatin shows that proliferation status can shape arrest-versus-death outcomes [4]. For single-cell perturbation prediction, this matters: a model can reconstruct a plausible mean transcriptome while missing whether a drug pushes cells toward S phase, G2/M, or another proliferative state. Such errors directly affect how we interpret drug mechanism, response heterogeneity, and out-of-distribution response.

*Email: 2411110131@stu.pku.edu.cn
†Corresponding author. Email: jaylin@stu.xmu.edu.cn

Recent single-cell perturbation models, including conditional autoencoder and compositional perturbation frameworks, have made substantial progress in predicting transcriptional responses from cell state, drug identity, dose, and covariates [6, 8, 9]. At the same time, recent benchmarks caution that perturbation-effect prediction remains difficult: gene-perturbation models do not always outperform simple linear baselines, and broad single-cell benchmarks emphasize generalization across unseen contexts and perturbation settings [1, 19]. Two limitations are especially important for drug response modeling. First, cell-cycle variation is often left implicit, treated as an ordinary covariate, or removed as unwanted variation, even when drug-induced phase changes are part of the biological response to be predicted. Second, benchmark preprocessing often emphasizes expression reconstruction while leaving cell-cycle labels, marker scores, molecule identities, and pretraining resources weakly connected. This makes it difficult to ask whether a perturbation model predicts drug-induced proliferation state rather than only a plausible transcriptome.

We propose scCycleMol, a cell-cycle-aware framework for single-cell drug perturbation prediction. The model builds on a ComPert/ChemCPA-style conditional autoencoder, but changes the learning signal around the predicted treated state. Instead of feeding cell-cycle phase to the model as an input covariate, scCycleMol derives supervision from treated single-cell profiles and applies it to the predicted treated expression. This design asks the model to explain drug-induced proliferation changes through its predicted response rather than by conditioning on the observed phase label. To make this supervision dense and differentiable, the main model attaches a lightweight cell-cycle head to the full predicted expression vector and trains it with circular G1/S/G2M phase targets. When this loss is allowed to update the upstream decoder, dose-response module, and molecular drug encoder, it forms a closed-loop training signal: the molecular representation predicts a perturbation response, and the predicted biological state gives feedback on whether that response is cell-cycle consistent.

scCycleMol also studies how drug representation and pretraining interact with this biological supervision. The drug representation can use SMILES language embeddings, molecular graph embeddings, or their projected combination, but we treat these choices as empirical components within the cell-cycle-aware prediction task. Our embedding ablations show that the proposed embedding improves LINCS-pretrained OOD expression metrics over the ChemCPA embedding, while SciPlex-pretrained embeddings are nearly tied and differ mainly in variance preservation. This motivates joint reporting of expression response, DE-gene response, variance preservation, and cell-cycle fidelity.

Our contributions are:

1. We organize a cell-cycle-aware drug perturbation benchmark from 24-hour SciPlex3, with standardized molecule identities, dose and cell-line metadata, modeled genes, marker scores, phase labels, and compatible LINCS and Tahoe pretraining resources.

2. We formulate treated-state cell-cycle supervision for single-cell drug perturbation prediction, using a learnable full-expression head and circular G1/S/G2M phase targets instead of supplying phase as an input nuisance covariate.

3. We show that scCycleMol improves OOD expression metrics over included conditional perturbation reference baselines while substantially increasing cell-cycle phase fidelity, with phase accuracy gains of 0.54–0.62 absolute points under closed-loop supervision.

2 Related Work

2.1 Single-Cell Perturbation Prediction

Single-cell perturbation models seek to predict the transcriptional response of a cell under genetic, chemical, or environmental interventions. Latent translation methods such as scGen model perturbation effects

as transformations in a learned expression space [8]. Compositional perturbation models and CPA-style approaches extend this idea by separating basal cell state, perturbation identity, dose, and covariates, often using adversarial objectives to encourage disentanglement [9]. ChemCPA further incorporates chemical representations to support drug generalization [6]. Related approaches address complementary regimes. GEARS models combinatorial genetic perturbations with graph priors [13], CellOT learns perturbation responses as neural optimal transport between single-cell distributions [2], and PRnet and PerturbNet target chemical or mixed chemical/genetic generalization to unseen perturbations [10, 22]. Recent diffusion-based models further frame single-cell drug-response and perturbation prediction as conditional generative modeling of treated cell states [7, 15]. scCycleMol builds on this family of conditional response predictors, but differs in the target of supervision: instead of optimizing only expression reconstruction and disentanglement, it requires predicted treated expression to preserve a specific biological state axis, namely cell cycle. Chemical representation remains important for drug generalization. SMILES language models, molecular graph neural networks, and multi-view molecular learning capture complementary chemical information [3, 12, 21, 24]. In this work, these representations are evaluated as components of a cell-cycle-aware perturbation model rather than as the central object of study.

2.2 Cell-Cycle State Modeling and Perturbation Resources

Cell cycle is a central source of variation in single-cell transcriptomics. Standard workflows often score cells using S and G2/M marker-gene sets from prior single-cell studies and assign coarse G1/S/G2M labels, with implementations in Seurat and Scanpy [14, 18, 20]. Depending on the scientific question, cell cycle may be corrected as a confounder or retained as a biological signal. In drug perturbation response, the latter is often crucial: many compounds alter proliferation, arrest, and phase distributions. scCycleMol therefore reframes cell cycle as a prediction target. This differs from using phase labels as covariates, because the model cannot explain away drug-induced state changes by conditioning on the answer.

Large-scale resources such as LINCS L1000, SciPlex, and more recent high-throughput single-cell drug screens provide complementary regimes of scale, modality, and biological resolution [16, 17, 23]. Bulk or reduced-gene resources offer broad chemical coverage, while single-cell resources expose heterogeneity and state redistribution. For cell-cycle-aware perturbation modeling, these resources must be connected through consistent molecule identities, gene vocabularies, treatment metadata, and phase annotations. Accordingly, scCycleMol uses processed SciPlex3 as the primary OOD benchmark and uses L1000 and Tahoe-100M as broader resources for pretraining and cross-resource analysis where compatibility supports direct comparison.

3 Method

3.1 Problem Setup

Let $\mathbf{x}_i \in \mathbb{R}^G$ denote the basal or control gene-expression profile of cell i over G genes. Each training example is associated with a drug $d_i \in \mathcal{D}$, dose u_i, and biological covariates such as cell type $\mathbf{c}_i$. The target is treated expression $\mathbf{y}_i \in \mathbb{R}^G$. Unlike standard perturbation prediction, we also observe or derive cell-cycle supervision from the treated cell, denoted s_i. Depending on the variant, s_i may be a pair of S/G2M marker scores, a discrete phase label in $\{\mathrm{G1}, \mathrm{S}, \mathrm{G2M}\}$, or a circular encoding of that phase. The goal is to learn a predictor

$$\hat{\mathbf{y}}_i = f_\theta(\mathbf{x}_i, d_i, u_i, \mathbf{c}_i) \tag{1}$$

that accurately predicts treated expression while preserving the cell-cycle state of the treated population. Cell-cycle information is used only as supervision on the predicted treated state, not as an input covariate.

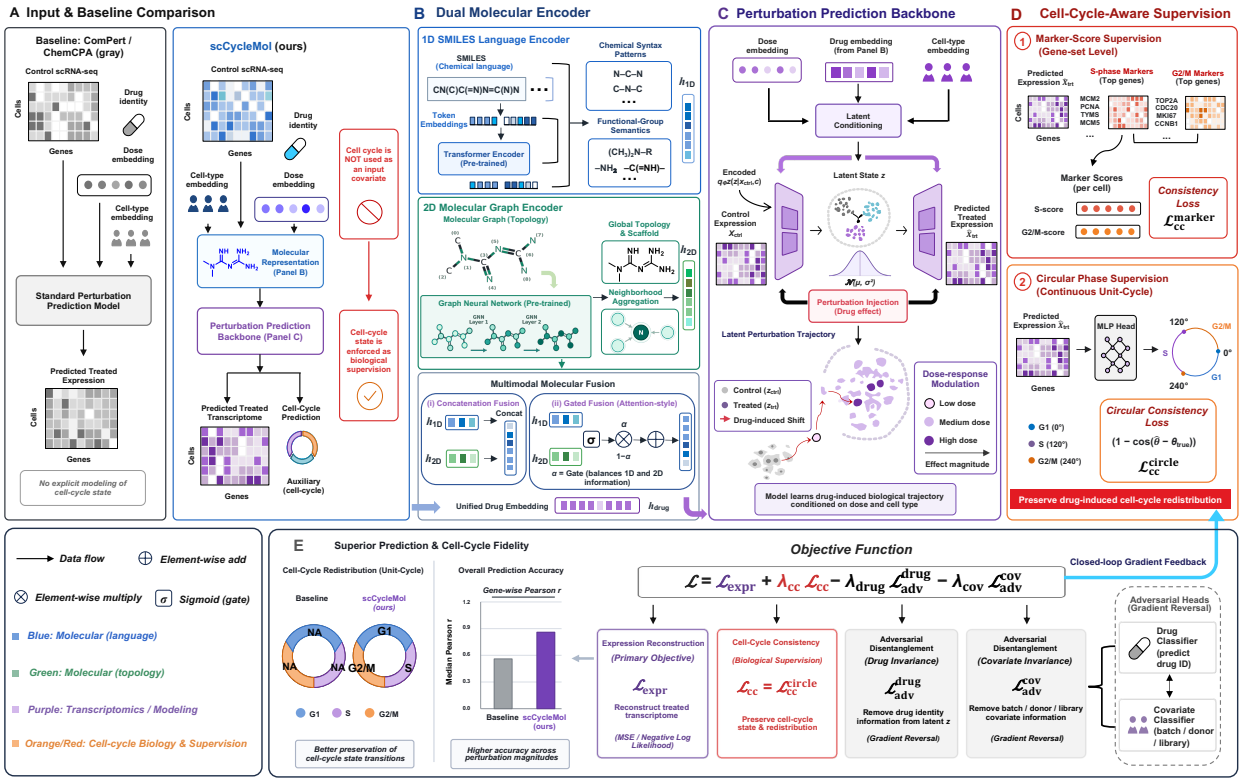


Figure 1: Overview of scCycleMol. The model encodes basal single-cell expression, conditions on drug and dose, and applies closed-loop cell-cycle supervision to the predicted treated state.

This prevents the model from conditioning on the observed phase label and instead requires drug-induced proliferation or phase changes to be represented through the predicted expression profile.

3.2 Base Perturbation Model

scCycleMol follows the conditional autoencoder design used in ComPert/ChemCPA-style models [6, 9]. An encoder maps control expression into a latent basal state,

$$\mathbf{z}_i = E_\theta(\mathbf{x}_i). \tag{2}$$

A drug encoder maps drug identity and molecular representation into a perturbation vector, optionally modulated by a dose-response function. Cell-type or other covariate embeddings are added in the latent space. The decoder then produces predicted treated expression:

$$\hat{\mathbf{y}}_i = D_\theta(\mathbf{z}_i + r_\theta(d_i, u_i) + q_\theta(\mathbf{c}_i)). \tag{3}$$

The expression loss is

$$\mathcal{L}_{\text{expr}} = \frac{1}{N}\sum_{i=1}^{N} \ell_{\text{expr}}(\hat{\mathbf{y}}_i, \mathbf{y}_i), \tag{4}$$

where ℓ_{expr} can be mean squared error or the reconstruction objective used by the base model. When inherited from ComPert/ChemCPA, adversarial objectives are used to discourage unwanted leakage of drug or covariate information into the basal latent representation.

3.3 Cell-Cycle-Aware Treated-State Supervision

A natural auxiliary objective is to compute S and G2/M marker scores from the predicted expression profile and match them to observed scores. This objective is biologically interpretable, but it routes the auxiliary gradient through only the marker-gene subset and relies on a fixed scoring rule. To provide a denser and more flexible supervisory signal, scCycleMol uses a learnable cell-cycle head that operates on the full predicted treated expression profile.

Marker-score auxiliary objective. We also consider a marker-score auxiliary objective as an interpretable ablation. Let $m_S(\cdot)$ and $m_{\mathrm{G2M}}(\cdot)$ denote differentiable or precomputed marker-score functions for S-phase and G2/M genes. Given predicted treated expression $\hat{\mathbf{y}}_i$, the model computes predicted scores

$$\hat{s}_i = [m_S(\hat{\mathbf{y}}_i), m_{\mathrm{G2M}}(\hat{\mathbf{y}}_i)]\,, \tag{5}$$

and minimizes

$$\mathcal{L}_{\mathrm{cc}}^{\mathrm{score}} = \frac{1}{N}\sum_{i=1}^{N} \|\hat{s}_i - s_i^{\mathrm{score}}\|_2^2\,. \tag{6}$$

This objective directly ties the predicted expression profile to expression-derived cell-cycle biology.

Cell-cycle prediction head. The main model replaces fixed score extraction with a lightweight learnable head. Given the predicted treated expression $\hat{\mathbf{y}}_i \in \mathbb{R}^G$, the head

$$\hat{p}_i = H_\theta(\hat{\mathbf{y}}_i) \tag{7}$$

maps the full expression vector to a two-dimensional phase representation. We implement H_θ as a small multilayer perceptron with a 64-dimensional hidden layer. Because the head receives all predicted genes rather than a fixed marker subset, the cell-cycle loss can shape the predicted treated profile through a dense end-to-end signal while adding only a small parameter overhead.

Circular phase encoding. Discrete phase labels impose an artificial ordering if treated as class indices. We therefore map the three coarse phases to equally spaced prototypes on the unit circle:

$$\mathrm{G1} \mapsto (1,0), \quad \mathrm{S} \mapsto \left(\cos\frac{2\pi}{3}, \sin\frac{2\pi}{3}\right), \quad \mathrm{G2M} \mapsto \left(\cos\frac{4\pi}{3}, \sin\frac{4\pi}{3}\right). \tag{8}$$

The circular loss is

$$\mathcal{L}_{\mathrm{cc}}^{\mathrm{circ}} = \frac{1}{N}\sum_{i=1}^{N} \|\hat{p}_i - p_i\|_2^2\,, \tag{9}$$

where p_i is the unit-circle target. At evaluation time, the predicted phase is obtained by nearest prototype on the circle.

Closed-loop supervision. The cell-cycle head is attached to the predicted treated expression without stopping gradients. Therefore, the circular phase loss backpropagates through the decoder, dose-response module, and drug-conditioned perturbation pathway. This creates a closed-loop training signal: the drug-conditioned perturbation pathway determines the predicted response, and the predicted biological state provides feedback on whether that response is cell-cycle consistent:

$$(d_i, u_i) \to h_d \to r_\theta(d_i, u_i) \to \hat{\mathbf{y}}_i \to H_\theta(\hat{\mathbf{y}}_i) \to \hat{p}_i. \tag{10}$$

3.4 Drug Representation and Pretraining Variants

The drug encoder is trained in the context of treated-state prediction rather than standalone molecular property prediction. Its representation must support both expression reconstruction and cell-cycle-consistent state prediction. We therefore evaluate molecular views before passing the drug embedding into the perturbation predictor. The 1D branch encodes the SMILES string with a chemical language model:

$$z_d^{1d} = \mathrm{LM}_\phi(\mathrm{SMILES}_d). \tag{11}$$

This representation captures token context, substructure syntax, and chemical language regularities. The 2D branch constructs a molecular graph $\mathcal{G}_d$ from the same molecule and encodes it with a graph neural network:

$$z_d^{2d} = \mathrm{GNN}_\psi(\mathcal{G}_d), \tag{12}$$

capturing atom-bond neighborhoods and topology. Both representations are projected into the perturbation model's drug space:

$$h_d^{1d} = W_{1d} z_d^{1d}, \qquad h_d^{2d} = W_{2d} z_d^{2d}. \tag{13}$$

We consider two fusion strategies. Static concatenation forms

$$h_d^{\mathrm{cat}} = W_f \,\mathrm{concat}(h_d^{1d}, h_d^{2d}). \tag{14}$$

Gated fusion learns a molecule-specific gate,

$$g_d = \sigma\left(\mathrm{MLP}_g(\mathrm{concat}(h_d^{1d}, h_d^{2d}))\right), \tag{15}$$

and combines views as

$$h_d^{\mathrm{gate}} = g_d \odot h_d^{1d} + (1 - g_d) \odot h_d^{2d}. \tag{16}$$

Let h_d denote the selected drug representation, which may be a single projected view or a combined representation. This drug embedding is passed to the dose-response module and downstream perturbation predictor. Because the downstream losses are applied to both predicted expression and predicted cell-cycle phase, the gate is optimized to select molecular views that are useful for biological perturbation response rather than molecule-only objectives.

3.5 Training Objective

The full objective is

$$\mathcal{L} = \mathcal{L}_{\mathrm{expr}} + \lambda_{\mathrm{cc}}\mathcal{L}_{\mathrm{cc}} - \lambda_{\mathrm{drug}}\mathcal{L}_{\mathrm{adv}}^{\mathrm{drug}} - \lambda_{\mathrm{cov}}\mathcal{L}_{\mathrm{adv}}^{\mathrm{cov}}, \tag{17}$$

Here, $\mathcal{L}_{\mathrm{adv}}^{\mathrm{drug}}$ is the loss of a drug adversary that predicts perturbation identity from the basal latent state $\mathbf{z}_i$ using multi-label binary cross-entropy. The term $\mathcal{L}_{\mathrm{adv}}^{\mathrm{cov}}$ sums covariate-adversary losses that predict nuisance covariates such as cell type from $\mathbf{z}_i$ using cross-entropy. The adversary networks are optimized to minimize these prediction losses, whereas the encoder-side objective includes them with negative signs, making $\mathbf{z}_i$ less informative about drug and covariate labels. In the main model, $\mathcal{L}_{\mathrm{cc}} = \mathcal{L}_{\mathrm{cc}}^{\mathrm{circ}}$ and gradients from this term are allowed to update the upstream perturbation pathway. The marker-score objective and stop-gradient variants are used as ablations to isolate the effect of closed-loop supervision. Thus, scCycleMol couples drug-conditioned perturbation prediction with treated-state cell-cycle supervision: the drug embedding predicts the expression shift, while the circular phase head provides biological feedback on whether that shift induces the correct cell-cycle state.

4 Experiments

Our experiments establish three findings. First, scCycleMol improves OOD expression metrics over included conditional perturbation reference baselines, especially on differentially expressed genes. Second, closed-loop cell-cycle supervision adds a strong biological fidelity signal: it raises phase accuracy by 0.54–0.62 absolute points relative to matched no-cell-cycle models while keeping expression reconstruction nearly unchanged. Third, circular phase supervision and pretraining source control complementary parts of the trade-off: LINCS pretraining favors expression metrics, whereas Tahoe pretraining gives the highest cell-cycle accuracy.

4.1 Datasets and Splits

SciPlex3. SciPlex3 is the primary single-cell drug perturbation benchmark [16]. We use a processed 24-hour benchmark with 635,541 cells, 186 non-control molecule-level perturbation units, 188 valid label-level compound embeddings, three cancer cell lines, four nonzero treatment doses plus DMSO control, and 5,080 modeled genes to evaluate expression prediction, DE-gene response prediction, cell-cycle fidelity, and out-of-distribution generalization. The preprocessing contract defines model-facing perturbation identities at the molecule level from RDKit-standardized SMILES [11]; original product labels are retained as traceability metadata rather than used as drug identities. Following the ChemCPA SciPlex3 convention [6], we use a chemCPA-compatible OOD fine-tuning split. The train and test partitions support internal validation and model selection, whereas the OOD partition is reserved for formal unseen-molecule evaluation.

LINCS pretraining. LINCS L1000 provides broader chemical coverage and is used as a pretraining source for compatible perturbation representations [17]. We evaluate whether LINCS initialization improves the downstream SciPlex3 expression and cell-cycle metrics after fine-tuning. Before formal SciPlex3 OOD evaluation, the L1000 pretraining input is filtered to remove molecules held out in the SciPlex3 OOD partition, while source perturbation identifiers are retained only for traceability.

Tahoe pretraining. Tahoe-100M is used as a large-scale single-cell pretraining source [23]. It provides a complementary pretraining regime with substantially larger single-cell context, allowing us to compare whether large-scale single-cell pretraining emphasizes the same metrics as LINCS pretraining. The Tahoe training input follows the same OOD holdout policy: source compound identifiers remain available for audit, but molecule-level identities control training identity and leakage checks. For both external resources, holdout filtering is performed against standardized molecule-level identities rather than source-specific compound labels.

4.2 Baselines and Model Variants

We compare against reference rows for a simple perturbation baseline, ChemCPA, and ChemCPA with LINCS pretraining. For scCycleMol, we evaluate variants that isolate the main design choices:

- **No cell-cycle supervision**: the perturbation model is trained without $\mathcal{L}_{\text{cc}}$.
- **Marker-score supervision**: the model uses the S/G2M marker-score auxiliary objective.
- **Circular phase supervision**: the model uses the learnable cell-cycle head and circular phase loss.
- **Closed-loop circular supervision**: circular phase loss is allowed to update the upstream perturbation pathway.

Model	Pretraining	E[r^2] all	E[r^2] DEGs	Δ DEGs	Median r^2 all	Median r^2 DEGs	Phase acc.	Δ phase
Baseline	–	0.5000	0.2900	–	0.4900	0.1200	–	–
ChemCPA	–	0.5100	0.3200	–	0.4700	0.2400	–	–
ChemCPA	LINCS	0.6800	0.5400	–	0.7500	0.6400	–	–
scCycleMol, no cell-cycle loss	–	0.8819	0.6315	0.0000	0.9294	0.7405	0.1918	0.0000
scCycleMol, closed-loop circular	–	0.8806	0.6318	+0.0003	0.9308	0.7449	0.7349	**+0.5431**
scCycleMol, circular	–	0.8829	0.6516	+0.0201	0.9164	0.7080	0.8872	**+0.6954**
scCycleMol, no cell-cycle loss	LINCS	0.9065	0.6695	0.0000	0.9369	0.7609	0.1928	0.0000
scCycleMol, closed-loop circular	LINCS	0.9085	0.6808	+0.0113	**0.9375**	**0.7774**	0.8098	**+0.6170**
scCycleMol, circular	LINCS	**0.9093**	**0.6843**	+0.0148	0.9369	0.7595	0.9006	**+0.7078**
scCycleMol, no cell-cycle loss	Tahoe	0.8801	0.6255	0.0000	0.9281	0.7293	0.1965	0.0000
scCycleMol, closed-loop circular	Tahoe	0.8824	0.6348	+0.0093	0.9309	0.7439	0.8122	**+0.6157**
scCycleMol, circular	Tahoe	0.8807	0.6312	+0.0057	0.9315	0.7344	**0.9609**	**+0.7644**

Table 1: Out-of-distribution expression and cell-cycle results on SciPlex3. The first three rows are included as conditional perturbation reference baselines. Δ DEGs and Δ phase are computed relative to the matched no-cell-cycle scCycleMol variant within the same pretraining group. Positive deltas are colored green. Closed-loop and circular supervision greatly improve phase fidelity while preserving expression reconstruction.

We also evaluate embedding-source variants in the ablation section, comparing ChemCPA embeddings with our molecular embeddings under LINCS and SciPlex pretraining.

4.3 Evaluation Metrics

Expression fidelity is measured with expected R^2 over all modeled genes, expected R^2 over differentially expressed genes, and the corresponding median R^2 values across perturbation conditions. The DE-gene metric is important because it focuses evaluation on genes most responsive to the perturbation. Cell-cycle fidelity is measured by phase classification accuracy from the predicted circular phase representation. Unless otherwise stated, the main tables report OOD results because this partition tests generalization to held-out molecules absent from SciPlex3 train/test and from the external pretraining inputs.

4.4 Main Results

Table 1 shows that scCycleMol substantially improves expression metrics over the included perturbation reference baselines. The best scCycleMol variant reaches 0.9093 expected all-gene r^2 and 0.6843 expected DE-gene r^2, compared with reference values of 0.6800 and 0.5400 for LINCS-pretrained ChemCPA. Cell-cycle supervision changes a different axis of performance: without $\mathcal{L}_{\text{cc}}$, phase accuracy remains near 0.19 on the OOD split, whereas closed-loop circular supervision raises it to 0.73–0.81 and direct circular supervision reaches as high as 0.9609. These gains do not come from sacrificing expression reconstruction, since the closed-loop variants keep expected all-gene r^2 within 0.003 of the corresponding no-cell-cycle setting.

4.5 Dose-Resolved OOD Performance

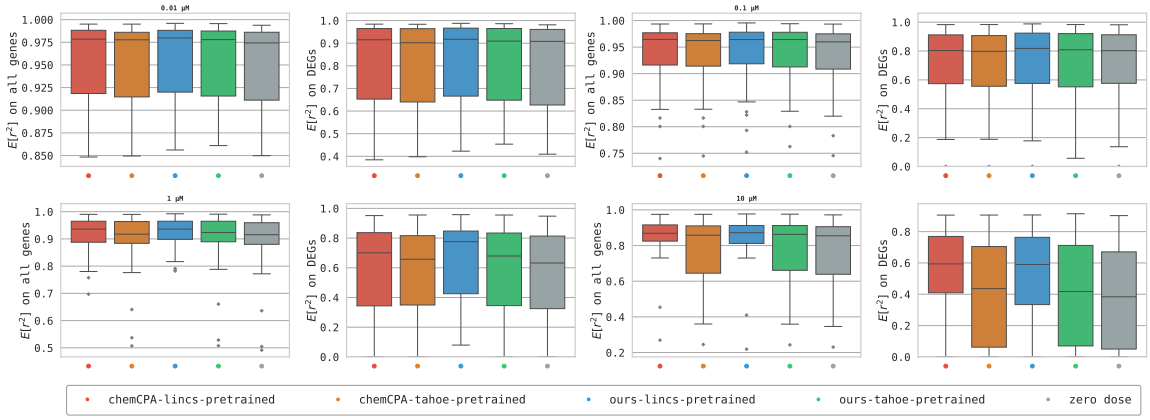


Figure 2: Dose-resolved OOD r^2 distributions on SciPlex3. Each box summarizes cell-line–molecule combinations at one treatment dose for all modeled genes or DE genes. The zero-dose baseline predicts treated expression from matched DMSO controls.

Figure 2 complements the aggregate OOD results by showing how performance varies across doses. Across the all-gene panels, pretrained models and the zero-dose control have high median r^2 at low doses, indicating that much of the transcriptome remains close to the control state when perturbation effects are weak. The DE-gene panels are more discriminative: their wider boxes and lower tails expose cell-line–molecule combinations where drug-responsive genes are substantially harder to predict. Within matched pretraining settings, scCycleMol remains competitive with ChemCPA across doses, and the LINCS-pretrained scCycleMol variant has especially strong DE-gene medians in the mid-dose panels. The dose-resolved view therefore supports the main conclusion that OOD expression performance is not only high on average, but also robust across the dose range where perturbation signal strength changes.

4.6 Pretraining Effects

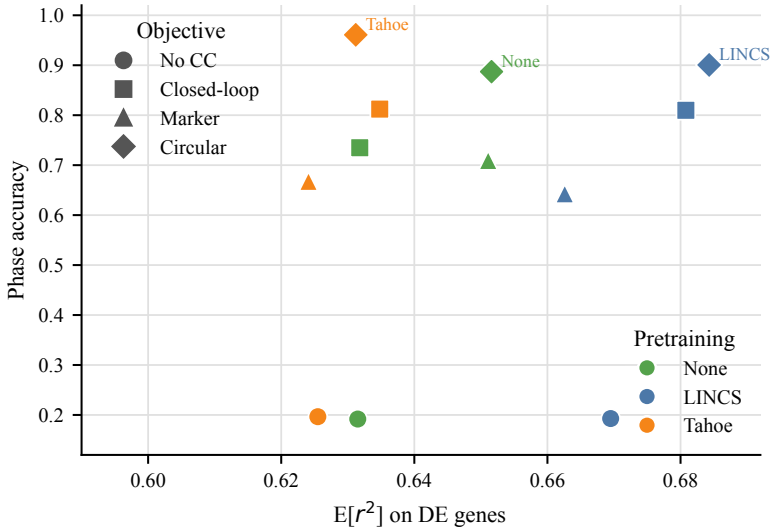


Figure 3: Expression–cell-cycle trade-off across objectives and pretraining sources. Points farther right have better DE-gene expression prediction, while points higher have better phase fidelity. LINCS pretraining moves models toward stronger expression metrics, while Tahoe pretraining gives the highest phase accuracy.

As shown in Figure 3, pretraining source changes the expression–cell-cycle trade-off. LINCS pretraining gives the highest expression metrics, especially on DE genes, while Tahoe pretraining gives the highest phase accuracy. We therefore treat pretraining source as an experimental factor rather than a universally dominant choice.

4.7 Implementation Details

All model variants use the same chemCPA-compatible train/test/OOD partitioning policy and the same evaluation protocol. For fair comparison, the cell-cycle objective weight λ_{cc} is fixed within each ablation group before OOD evaluation. The marker-score and circular objectives are compared as alternative auxiliary losses, while stop-gradient and closed-loop variants isolate whether the cell-cycle loss is allowed to update the upstream perturbation pathway.

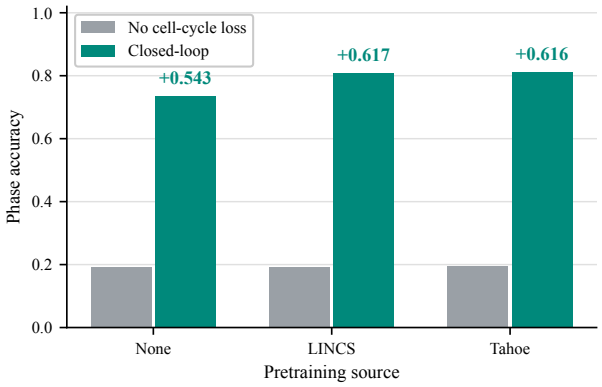
5 Analysis and Ablations

The main results show that scCycleMol improves both expression prediction and cell-cycle fidelity. We next isolate which design choices support these gains.

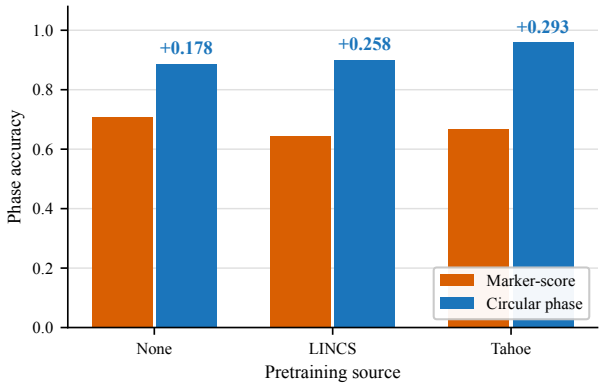
5.1 Closed-Loop Cell-Cycle Supervision

Closed-loop supervision tests whether cell-cycle loss is merely an auxiliary readout or whether it provides useful feedback to the perturbation pathway. We compare each no-cell-cycle model with its closed-loop counterpart, where the circular phase loss backpropagates through the predicted treated expression into the upstream decoder and drug-conditioned perturbation pathway.

Figure 4a supports the closed-loop design. Across all three pretraining settings, phase accuracy increases sharply: from 0.1918 to 0.7349 without pretraining, from 0.1928 to 0.8098 with LINCS pretraining, and



(a) Closed-loop supervision.



(b) Marker-score versus circular phase supervision.

Figure 4: Cell-cycle supervision ablations on the OOD split. (a) Allowing the circular phase loss to update the perturbation pathway improves phase accuracy by 0.54–0.62 absolute points across pretraining settings. (b) Circular phase supervision consistently improves phase accuracy over marker-score supervision, especially with LINCS or Tahoe pretraining.

from 0.1965 to 0.8122 with Tahoe pretraining. The corresponding changes in expected all-gene r^2 are small, ranging from -0.0013 to +0.0023. Thus, the biological-state feedback improves cell-cycle fidelity without introducing a large reconstruction penalty.

5.2 Marker-Score versus Circular Phase Supervision

The marker-score objective is interpretable because it directly supervises S and G2/M expression-derived scores, but it depends on a fixed marker subset. The circular objective instead supervises a learnable head on the full predicted expression profile. Figure 4b compares these alternatives.

Circular phase supervision consistently gives higher phase accuracy than marker-score supervision. The difference is largest with pretraining: LINCS increases from 0.6427 to 0.9006, and Tahoe increases from 0.6676 to 0.9609. This supports the use of a learnable full-expression cell-cycle head rather than a fixed marker-score rule as the main supervision mechanism.

5.3 Embedding Source Ablation

We also examine whether the learned drug embedding improves OOD perturbation prediction beyond the corresponding ChemCPA embedding. Table 2 compares ChemCPA and scCycleMol embeddings under LINCS and SciPlex pretraining.

Embedding setting	Mean R^2 all	Mean R^2 DEGs	Var. R^2 all	Var. R^2 DEGs
LINCS + ChemCPA embedding	0.8756	0.5869	0.0361	0.0201
LINCS + Ours embedding	**0.8841**	0.6245	0.0510	0.0220
SciPlex + ChemCPA embedding	0.8375	**0.8011**	0.7249	0.7230
SciPlex + Ours embedding	0.8386	**0.8011**	**0.7257**	**0.7255**

Table 2: Embedding-source ablation on the OOD split. Under LINCS pretraining, our embedding improves mean all-gene and DE-gene R^2 over the ChemCPA embedding. Under SciPlex pretraining, the two embeddings are nearly tied on mean DE-gene R^2, while our embedding slightly improves variance preservation.

The embedding ablation shows that representation gains depend on the pretraining source. With LINCS pretraining, our embedding improves expected all-gene R^2 by 0.0085 and expected DE-gene R^2 by 0.0376 over the ChemCPA embedding, with modest gains in both variance metrics. With SciPlex pretraining, the two embedding choices are effectively tied on DE-gene response, while our embedding slightly improves all-gene mean R^2 and both variance metrics. These results support using the proposed molecular embedding, but they also indicate that embedding choice should be interpreted together with the pretraining regime and the expression-versus-variance trade-off.

5.4 Expression–Cell-Cycle Trade-Off

The combined results suggest that expression reconstruction and cell-cycle fidelity should be reported together. LINCS pretraining gives the strongest expression reconstruction, while Tahoe pretraining gives the highest phase accuracy. Closed-loop supervision improves phase accuracy with little change in expression R^2, whereas circular supervision with a larger cell-cycle weight can maximize phase fidelity. This trade-off motivates reporting both DE-gene R^2 and phase accuracy as primary metrics rather than optimizing only one of them.

6 Conclusion

We presented scCycleMol, a cell-cycle-aware framework for single-cell drug perturbation prediction. The central idea is to predict drug-induced biological state rather than remove it: scCycleMol trains predicted treated expression to preserve S/G2M marker scores or circular G1/S/G2M phase targets. This framing is supported by a processed cell-cycle-aware SciPlex3 benchmark and by molecular representation and pretraining ablations within the same prediction task.

Experiments on the processed SciPlex3 benchmark show that this design improves OOD expression metrics over included reference baselines and sharply increases treated-state phase fidelity. Closed-loop circular supervision raises phase accuracy by 0.54–0.62 absolute points with little change in expression r^2, while circular phase supervision reaches the strongest phase fidelity under Tahoe pretraining. The embedding ablations show that drug representation gains depend on pretraining source, reinforcing the need to report expression response, variance preservation, and cell-cycle fidelity together. Limitations include reliance on reliable phase labels or marker scores, the coarse three-phase circular encoding, and the need to validate whether cell-cycle-aware supervision generalizes across broader perturbation datasets. Future work should extend the framework to continuous cell-cycle trajectories and broader cellular state variables beyond cell cycle.

References

[1] Constantin Ahlmann-Eltze, Wolfgang Huber, and Simon Anders. Deep-learning-based gene perturbation effect prediction does not yet outperform simple linear baselines. *Nature Methods*, 22(8):1657–1661, 2025. doi: 10.1038/s41592-025-02772-6.

[2] 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.

[3] Seyone Chithrananda, Gabriel Grand, and Bharath Ramsundar. ChemBERTa: Large-scale self-supervised pretraining for molecular property prediction, 2020. URL https://arxiv.org/abs/2010.09885. arXiv preprint arXiv:2010.09885.

[4] Adrián E. Granada, Alba Jiménez, Jacob Stewart-Ornstein, Nils Blüthgen, Simone Reber, Ashwini Jambhekar, and Galit Lahav. The effects of proliferation status and cell cycle phase on the responses of single cells to chemotherapy. *Molecular Biology of the Cell*, 31(8):845–857, 2020. doi: 10.1091/mbc.E19-09-0515.

[5] Sean M. Gross, Farnaz Mohammadi, Crystal Sanchez-Aguila, Paulina J. Zhan, Tiera A. Liby, Mark A. Dane, Aaron S. Meyer, and Laura M. Heiser. Analysis and modeling of cancer drug responses using cell cycle phase-specific rate effects. *Nature Communications*, 14(1):3450, 2023. doi: 10.1038/s41467-023-39122-z.

[6] Leon Hetzel, Simon Boehm, Niki Kilbertus, Stephan Günnemann, Mohammad Lotfollahi, and Fabian J. Theis. Predicting cellular responses to novel drug perturbations at a single-cell resolution. In *Advances in Neural Information Processing Systems*, volume 35, pages 26711–26722. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/aa933b5abc1be30baece1d230ec575a7-Paper-Conference.pdf.

[7] Zhaokang Liang, Shuyang Zhuang, Xiaoran Jiao, Weian Mao, Hao Chen, and Chunhua Shen. scppdm: A diffusion model for single-cell drug-response prediction. *arXiv preprint arXiv:2510.11726*, 2025.

[8] 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.

[9] Mohammad Lotfollahi, Anna Klimovskaia Susmelj, Carlo De Donno, Leon Hetzel, Yuge Ji, Ignacio L. Ibarra, Sanjay R. Srivatsan, Mohsen Naghipourfar, Riza M. Daza, Beth Martin, Jay Shendure, Jose L. McFaline-Figueroa, Pierre Boyeau, F. Alexander Wolf, Nafissa Yakubova, Stephan Günnemann, Cole Trapnell, David Lopez-Paz, and Fabian J. Theis. Predicting cellular responses to complex perturbations in high-throughput screens. *Molecular Systems Biology*, 19(6):e11517, 2023. doi: 10.15252/msb.202211517.

[10] 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(1):9256, 2024. doi: 10.1038/s41467-024-53457-1.

[11] RDKit Developers. RDKit: Open-source cheminformatics, 2026. URL https://www.rdkit.org. Accessed: 2026-05-16.

[12] Yu Rong, Yatao Bian, Tingyang Xu, Weiyang Xie, Ying Wei, Wenbing Huang, and Junzhou Huang. Self-supervised graph transformer on large-scale molecular data. In *Advances in Neural Information Processing Systems*, volume 33, pages 12559–12571. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/94aef38441efa3380a3bed3faf1f9d5d-Paper.pdf.

[13] Yusuf Roohani, Kexin Huang, and Jure Leskovec. Predicting transcriptional outcomes of novel multigene perturbations with GEARS. *Nature Biotechnology*, 42(6):927–935, 2024. doi: 10.1038/s41587-023-01905-6.

[14] Rahul Satija, Jeffrey A. Farrell, David Gennert, Alexander F. Schier, and Aviv Regev. Spatial reconstruction of single-cell gene expression data. *Nature Biotechnology*, 33(5):495–502, 2015. doi: 10.1038/nbt.3192.

[15] Peiting Shi, Ningfeng Que, Xianzhe Huang, Xiaofei Wang, and Jianzhong Jeff Xi. Statexdiff: Cell state-contextualized multimodal diffusion for single-cell perturbation prediction. *arXiv preprint arXiv:2605.16104*, 2026.

[16] 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.

[17] 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.

[18] Itay Tirosh, Benjamin Izar, Sanjay M. Prakadan, Marc H. Wadsworth, Daniel Treacy, John J. Trombetta, Asaf Rotem, Christopher Rodman, Christine Lian, George Murphy, Mohammad Fallahi-Sichani, Ken Dutton-Regester, Jia-Ren Lin, Ofir Cohen, Parin Shah, Diana Lu, Alex S. Genshaft, Travis K. Hughes, Carly G. K. Ziegler, Samuel W. Kazer, Aleth Gaillard, Kellie E. Kolb, Alexandra-Chloé Villani, Cory M. Johannessen, Aleksandr Y. Andreev, Eliezer M. Van Allen, Monica Bertagnolli, Peter K. Sorger, Ryan J. Sullivan, Keith T. Flaherty, Dennie T. Frederick, Judit Jané-Valbuena, Charles H. Yoon, Orit Rozenblatt-Rosen, Alex K. Shalek, Aviv Regev, and Levi A. Garraway. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science*, 352(6282):189–196, 2016. doi: 10.1126/science.aad0501.

[19] Zhiting Wei, Yiheng Wang, Yicheng Gao, Shuguang Wang, Ping Li, Duanmiao Si, Yuli Gao, Siqi Wu, Danlu Li, Kejing Dong, Xingbo Yang, Chen Tang, Shaliu Fu, Xiaohan Chen, Wannian Li, Yuzhou You, Chen Zhang, Aibin Liang, Guohui Chuai, and Qi Liu. Benchmarking algorithms for generalizable single-cell perturbation response prediction. *Nature Methods*, 23(2):451–464, 2026. doi: 10.1038/s41592-025-02980-0.

[20] F. Alexander Wolf, Philipp Angerer, and Fabian J. Theis. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biology*, 19(1):15, 2018. doi: 10.1186/s13059-017-1382-0.

[21] Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? In *International Conference on Learning Representations*, 2019. URL `https://openreview.net/forum?id=ryGs6iA5Km`.

[22] Hengshi Yu, Weizhou Qian, Yuxuan Song, and Joshua D. Welch. PerturbNet predicts single-cell responses to unseen chemical and genetic perturbations. *Molecular Systems Biology*, 21(8):960–982, 2025. doi: 10.1038/s44320-025-00131-3.

[23] Jesse Zhang, Airol A. Ubas, Richard de Borja, Valentine Svensson, Nicole Thomas, Neha Thakar, Ian Lai, Aidan Winters, Umair Khan, Matthew G. Jones, John D. Thompson, Vuong Tran, Joseph Pangallo, Efthymia Papalexi, Ajay Sapre, Hoai Nguyen, Oliver Sanderson, Maria Nigos, Olivia Kaplan, Sarah Schroeder, Bryan Hariadi, Simone Marrujo, Crina Curca Alec Salvino, Guillermo Gallareta Olivares, Ryan Koehler, Gary Geiss, Alexander Rosenberg, Charles Roco, Daniele Merico, Nima Alidoust, Hani Goodarzi, and Johnny Yu. Tahoe-100M: A giga-scale single-cell perturbation atlas for context-dependent gene function and cellular modeling. *bioRxiv*, page 2025.02.20.639398, 2025. doi: 10.1101/2025.02.20.639398. URL `https://www.biorxiv.org/content/10.1101/2025.02.20.639398`. Preprint.

[24] Ru Zhang, Yanmei Lin, Yijia Wu, Lei Deng, Hao Zhang, Mingzhi Liao, and Yuzhong Peng. MvMRL: a multi-view molecular representation learning method for molecular property prediction. *Briefings in Bioinformatics*, 25(4):bbae298, 2024. doi: 10.1093/bib/bbae298.

A Additional Experimental Details

A.1 Preprocessing Overview

We provide a reproducible Stage 1–2 preprocessing package for the three data resources used in this work: SciPlex3 [16], LINCS L1000 [17], and Tahoe-100M [23]. Stage 1 prepares dataset-specific expression matrices, metadata, feature vocabularies, and quality-control outputs. Stage 2 converts these outputs into model-ready training data, including compact expression matrices, cell and compound metadata, ChemBERTa drug embeddings, molecule-level perturbation identities, OOD holdout filtering, Top50 DEG annotations for SciPlex3, and validation manifests. The migrated preprocessing code preserves the legacy transformation logic where appropriate, while updating the formal data contract to use standardized-molecule identities rather than source-specific compound labels. Legacy-compatible outputs may be retained for auditability, but the formal paper protocol uses the molecule-aware, Top50-DEG, OOD-holdout-filtered assets described below. For model-facing identities, each dataset uses the standardized molecule SMILES emitted by its preprocessing path; external L1000 and Tahoe assets additionally use the shared molecule-identity helper, where the model SMILES is RDKit canonical isomeric SMILES. For cross-resource leakage control, the held-out SciPlex3 OOD molecules are matched more broadly using audit keys that include canonical and non-stereo SMILES, fragment/charge/tautomer parent SMILES, and InChIKey or connectivity keys. This broader audit matching is used only to remove potentially overlapping compounds from external pretraining inputs; source compound identifiers remain traceability fields.

The public release boundary is the preprocessing code, configuration files, lightweight documentation, and static manifest summaries. Raw datasets, third-party pretrained model weights, large processed outputs, local run directories, and per-run manifest JSON files are not redistributed. Users are expected to download raw datasets from the official sources and regenerate processed outputs locally. ChemBERTa embedding steps use a local copy of `seyonec/ChemBERTa-zinc-base-v1`; model weights are not included in the repository.

A.2 Dataset Statistics After Preprocessing

Dataset	Cells / samples	Model perturbation identities	Genes	Cell lines	Formal asset
SciPlex3	635,541 cells	186 non-control molecule units	5,080	3	molecule-aware Top50-DEG AnnData
Tahoe-100M	92,670,735 cells	351 non-control molecule units	5,037	49	OOD-holdout-filtered training shards
LINCS L1000	157,927 signatures	20,388 molecule IDs	978	21	OOD-holdout-filtered Stage-0 dataset

Table 3: Formal model-ready assets generated by the Stage 1–2 preprocessing pipelines. The perturbation-identity column reports model-facing molecule-level units rather than source compound labels; source identifiers are retained separately for traceability. SciPlex3 has 186 non-control molecule units, and Tahoe has 351 non-control molecule units, or 352 molecule identities when the control identity is included. The external L1000 and Tahoe assets are filtered against the SciPlex3 OOD molecule list using molecule audit keys.

A.3 SciPlex3 Processing

SciPlex3 processing starts from the decompressed GEO large-screen files for A549, MCF7, and K562. The required Stage 1 path performs basic quality control, hash-based doublet removal, highly variable gene selection, and cell-cycle scoring. The required Stage 2 path extracts training features, computes ChemBERTa embeddings for standardized SMILES strings, writes the compact training AnnData object, assigns molecule-aware splits, and annotates Top50 perturbation-responsive genes for formal DE-gene evaluation.

Step	Input	Output	Main operation
Basic filtering	799,317 cells, 58,347 genes	661,596 cells, 57,591 genes	Remove incomplete metadata, non-target time points, low UMI/gene cells, high mitochondrial fraction, and cleaned-up genes.
Doublet filtering	661,596 cells	635,541 cells	Remove 26,055 hash doublets.
HVG selection	57,591 genes	5,080 genes	Select HVGs with legacy-equivalent settings and force-include cell-cycle genes.
Cell-cycle scoring	635,541 cells	635,541 cells	Add S score, G2/M score, phase label, phase ID, and proliferation score.
Feature extraction	635,541 cells	188 non-control standardized SMILES	Add standardized SMILES, dose, cell-line, time, and control indicators.
ChemBERTa embedding	189 compound labels	188 valid embeddings + control sentinel	Encode valid label-level compounds with 768-dimensional ChemBERTa CLS embeddings; write a zero-vector sentinel for the control/invalid placeholder and attach the resulting per-cell matrix.
Training object	635,541 cells, 5,080 genes	635,541 cells, 5,080 genes	Write compact AnnData with 9 obs columns, `counts`, `log1p_norm`, `compound_chemberta`, and unified gene token IDs.
Molecule-aware split assignment	635,541 cells	564,501/45,862/25,178 train/test/OOD cells	Add the chemCPA-compatible OOD split over 186 non-control molecule units and an optional molecule-strict split for sensitivity analysis.
Top50 DEG annotation	635,541 cells, 5,080 genes	2,232 non-control groups	Compute the Top50 perturbation-responsive genes for each non-control cell-line–molecule–dose group.

Table 4: SciPlex3 preprocessing path under the molecule-aware Top50-DEG contract. The model-facing perturbation identity is the molecule-level condition; label-level embedding entries are retained only as molecular representation inputs and traceability links. Cell-cycle scores matched the legacy output within 2.9×10^{-7} maximum absolute difference, and phase labels matched exactly.

Molecule identity and repeated structures. Standardized SMILES define the molecule-level perturbation identities used for training and split assignment. Original SciPlex3 product labels, LINCS perturbation identifiers, and Tahoe compound identifiers are retained as traceability metadata rather than used as canonical drug identities. Repeated standardized structures are not averaged or dropped: sample rows remain distinct observations across cell line, dose, batch, and replicate, while shared structures map to the same molecule-level identity. For external leakage checks, the molecule-level identity is supplemented with broader parent, tautomer, stereochemistry-insensitive, and InChIKey-based audit keys before filtering L1000 or Tahoe.

Top50 DEG contract. SciPlex3 DE-gene evaluation uses the top 50 perturbation-responsive genes for each non-control cell-line–molecule–dose group, producing 2,232 groups with 50 genes each. Control reference groups are excluded from the DEG dictionary. Legacy full-gene DEG annotations are retained only for compatibility and are not used for paper-style DE-gene evaluation.

A.4 Tahoe-100M Processing

Tahoe-100M processing starts from public expression parquet shards and metadata tables. The pipeline validates shard continuity and metadata, computes gene statistics, selects HVGs, extracts sample/compound/cell-line/dose/phase features, creates training-ready expression shards, embeds compounds with ChemBERTa, and validates the resulting package. Normal cell lines are excluded from tumor-only statistics and training data. Compounds without a compatible single-compound SMILES representation are removed from the feature and training outputs. The formal Tahoe asset additionally removes cells whose compound rows match the SciPlex3 OOD molecule audit-key set; source compound identifiers remain available for audit, but molecule-level conditions control training identity and split assignment. The Tahoe HVG step uses a scalable Seurat-v3-inspired mean/variance procedure with forced cell-cycle genes rather than a full official Scanpy `seurat_v3` run on all cells.

Step	Input	Output	Main operation
Data validation	3,388 shards; 100,648,790 metadata rows	Validation reports	Check parquet continuity, metadata readability, barcode coverage, and sampled expression rows.
Gene statistics	62,710 genes	54,877 detected genes	Compute gene-level detection and expression statistics after excluding normal-cell summaries.
HVG selection	62,710 genes	5,037 HVGs	Use a scalable Seurat-v3-inspired HVG procedure, remove 40,337 genes by HVG prefilter, and force-add 37 cell-cycle genes.
Input features	1,344 samples	1,341 samples; 379 source compounds; 354 molecule identities; 49 cell lines	Remove three samples from an incompatible multi-compound drug representation and build feature vocabularies.
Training data	95,624,334 cells	93,698,301 cells	Remove normal-cell-line cells and incompatible compound rows; write HVG shards and cell metadata.
OOD holdout filtering	93,698,301 cells; 379 source compounds	92,670,735 cells; 376 source compounds; 351 non-control molecule units	Remove cells from compound rows matching the SciPlex3 OOD molecule audit-key set.
ChemBERTa embedding	379 source compounds	376 filtered embedding rows	Encode compatible source compounds before holdout filtering; copy the filtered embedding sidecar, retaining one zero-vector sentinel in the formal asset.
Validation	training-ready outputs	50/50 passed checks	Validate required fields, vocabularies, metadata, embeddings, expression shards, excluded-compound removal, and training manifest.

Table 5: Tahoe-100M preprocessing path. The formal OOD-holdout-filtered asset contains 92,670,735 cell metadata rows after removing 1,027,566 cells from three Tahoe compound IDs matched to SciPlex3 OOD molecules. After source-compound filtering, 376 Tahoe compound rows map to 351 non-control molecule-level split units, or 352 molecule identities when the control identity is included; molecule split counts are 281 train, 35 test, and 35 OOD non-control units.

A.5 LINCS L1000 Processing

L1000 processing creates a Stage 0 VAE dataset from treatment and control signatures. The pipeline filters signatures by perturbation type, tumor sample type, treatment duration, usable canonical SMILES, quality metrics, and required Stage 0 fields. It then extracts landmark-gene expression matrices for treatment and control signatures, computes ChemBERTa embeddings for compounds, and joins treatment signatures with matched expression, compound embeddings, and cell-line vehicle controls. The formal Stage 0 asset is filtered against the SciPlex3 OOD molecule audit-key set before pretraining. After filtering, it retains 20,714 source perturbation identifiers corresponding to 20,388 molecule-level identities. LINCS source perturbation identifiers are retained for traceability, while molecule-level identifiers define condition identity, split units, and repeated-structure weighting.

Step	Input	Output	Main operation
Load and filter	591,697 signatures	158,151 treatment signatures; 8,444 controls	Remove non-`trt_cp`, non-tumor, incompatible duration, unusable SMILES, low-quality, or incomplete signatures.
Treatment expression	158,151 signatures	$158{,}151 \times 978$ matrix	Extract L1000 landmark-gene expression for treatment signatures.
Control expression	8,444 signatures	$8{,}444 \times 978$ matrix	Extract matched control expression for vehicle signatures.
ChemBERTa embedding	20,721 source perturbation IDs	$20{,}721 \times 768$ matrix	Standardize SMILES and compute source-level compound embeddings before OOD holdout filtering.
Stage 0 dataset	158,151 treatment signatures	158,151 samples	Join expression, compound embeddings, metadata, control means, genes, condition summaries, and split units.
OOD holdout filtering	158,151 samples	157,927 samples; 20,388 molecule IDs	Remove 224 signatures from seven source perturbation identifiers matching the SciPlex3 OOD molecule audit-key set; split molecule IDs into 16,310 train, 2,039 test, and 2,039 OOD units.

Table 6: LINCS L1000 preprocessing path. The formal OOD-holdout-filtered Stage-0 asset contains 157,927 samples and 20,388 molecule-level identities; 20,714 source perturbation identifiers are retained only for traceability.

A.6 Gene Vocabulary and Manifests

All datasets use a shared lightweight gene vocabulary for stable token joins. The public key table contains stable Ensembl-to-token mappings, final-dataset membership flags, retired-ID replacement fields, and traceability annotations. Only `Unified_ENSG` and `Unified_Token_ID` are stable join keys; gene symbols are annotations and may reflect dataset-original, alias, or retired-ID names. Rows marked as traceability-only are retained for auditability and are not additional training genes.

Each accepted preprocessing step writes a machine-readable manifest recording inputs, outputs, filtering decisions, counts, configuration values, runtime, and hashes where practical. The public documentation includes a static cross-dataset manifest summary generated from the accepted manifests. The manuscript reports the corresponding data contracts rather than local run commands or site-specific paths.