scUGPT: Gene-wise Response Curves Along a Supplied Single-Cell Trajectory Axis

Anonymous authors
Paper under double-blind review

Abstract

A central goal in functional genomics is to predict not only how strongly a perturbation shifts a cell, but when each gene responds. Single-cell measurements are noisy and sparse, and a response is observed at only a few experimental time points. Most perturbation models predict gene expression at a measured time point, so they do not directly describe response onset or persistence. We present scUGPT, a trajectory-conditioned model that predicts gene-wise response curves along a supplied process axis. An analytic process kernel separates response centre, persistence, and late switching, distinguishing transient and maintained programmes in the parameterization. Perturbation-conditioned attention and a residual prediction path refine the predicted curve. Kernel parameters describe the primary path; timing is scored on the final output. On the fixed main GSE213069 split (seed 2026), anchor calibration improves full-curve reconstruction and the model reaches day-4-to-day-5 change correlation 0.261 ± 0.041, versus 0.156 ± 0.032 without the residual path. Across four additional condition splits, change correlation averages 0.007 ± 0.122. The results support a structured representation of response curves, with forecasting performance that depends on the condition split. Code is available at `https://osf.io/xn3fh/?view_only=dc7333b3d4d74a0aa1baaf9f964306bf`.

1 Introduction

Predicting how a cell responds to a genetic or chemical perturbation requires both magnitude and timing. A programme that is induced and then decays before harvest can look similar to one that remains active, even though the underlying regulation differs. Exhaustive time-resolved profiling is infeasible at screening scale, so models that place responses along an axis derived from existing measurements are useful. A response is observed only at the harvest times included in an experiment. Some immune genes activate early and return to baseline before a late harvest (Amit et al., 2009; Hagai et al., 2018), and hematopoiesis shows a similar order, with erythroid factors preceding globin and heme programs (Cantor & Orkin, 2002; Paul et al., 2015). At a single endpoint, transient, delayed, and sustained responses can therefore have similar levels. The measurements also make timing difficult: single-cell transcriptomes are sparse, zero-inflated, and noisy, and a time course contains only a few harvests, so derivatives are estimated from few intervals. Existing methods address parts of this problem. Perturbation predictors such as scGen, GEARS, scGPT, and neural optimal transport estimate counterfactual expression states or distributions at a measured harvest (Lotfollahi et al., 2019; Roohani et al., 2024; Cui et al., 2024; Bunne et al., 2023). These models suit pooled CRISPR screens with limited time coverage (Dixit et al., 2016; Norman et al., 2019; Replogle et al., 2022), but they do not expose response onset, persistence, or reversal. Pseudotime and RNA-velocity methods characterize progression among observed cells (Trapnell et al., 2014; Haghverdi et al., 2016; La Manno et al., 2018; Bergen et al., 2020; Lange et al., 2022), but they do not directly predict a gene-wise process curve for an unseen perturbation. Endpoint benchmarks also show that simple baselines remain competitive on endpoint metrics (Ahlmann-Eltze et al., 2025). These settings leave a gap between endpoint perturbation prediction and trajectory analysis. We introduce scUGPT, a trajectory-conditioned model with an explicit process representation. The analytic kernel parameterizes response centre, persistence, and late switching; onset and peak time are measured from the resulting curve. A single symmetric RBF couples rise and decay around one centre, so it cannot distinguish a transient response from a persistent one or represent a delayed component that reinforces or reverses the initial effect without relying on superposition. A parametric prior can also

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.

be underdetermined on a coarse grid. We therefore refine the kernel with perturbation-conditioned attention and a residual prediction path. The primary path retains explicit shape parameters, while the refinement accommodates asymmetric, persistent, delayed, and reversing kinetics on the supplied axis.

Our contributions are threefold.

First, we formulate perturbation response prediction as a structured process function along a supplied axis, with explicit parameters for response location, persistence, and late reversal. The estimand is the curve shape rather than a level at one harvest.

Second, we implement this formulation with a gene-wise analytic process kernel, perturbation-conditioned process attention, and a residual prediction path that represents structure outside the analytic prior.

Third, we evaluate the model on six public datasets spanning erythroid pseudotime, wall-clock immune stimulation, a four-day scCRISPR time course with held-out-condition reconstruction and one-day forecasting, and harvest-only Perturb-seq screens for unseen single perturbations.

2 Related Work

Perturbation prediction. scGen, GEARS, scGPT, and CellOT predict counterfactual expression states or distributions (Lotfollahi et al., 2019; Roohani et al., 2024; Cui et al., 2024; Bunne et al., 2023), but none exposes parameters for onset or persistence. A response curve along a supplied axis is a different estimand from an endpoint value, and a model that fits the endpoint need not place the peak correctly. We therefore treat the harvest-only comparison as a boundary case rather than the primary measure.

Trajectory and velocity models. Monocle, diffusion pseudotime, and Palantir infer developmental order from observed cells (Trapnell et al., 2014; Haghverdi et al., 2016; Setty et al., 2019); RNA velocity and CellRank estimate local direction or fate probabilities (La Manno et al., 2018; Bergen et al., 2020; Lange et al., 2022). scNODE combines a VAE with a neural ODE to predict missing single-cell time points (Zhang et al., 2024), whereas PRESCIENT learns a population-level stochastic landscape for time-series trajectories and interventions (Yeo et al., 2021). RNAForecaster forecasts future states from matched same-cell matrices at adjacent times (Erbe et al., 2023), and PerturbODE couples perturbation prediction with neural-ODE regulatory discovery (Lin et al., 2025). These methods either learn the trajectory itself or require matched observations across time, whereas scUGPT takes the coordinate as given and parameterizes each gene's response along it. This makes outputs comparable across perturbations that share an axis but not a cell state.

3 Method

3.1 Notation and Problem Setting

The supplied coordinate $\tau \in [0, 1]$ is an experimental time or an inferred trajectory position; it is fixed throughout training. A condition has a static gene-level input $\boldsymbol{\delta} \in \mathbb{R}^G$, a continuous control-expression profile $\mathbf{c}_0$, and optional categorical metadata c. These are distinct inputs. The target $\mathbf{y}(\tau)$ is the condition's gene-wise response along the axis. The neural model learns projections, metadata embeddings, kernel-parameter maps, attention, and residual weights.

The meaning of $\boldsymbol{\delta}$ depends on the protocol. GSE213069 reconstruction uses the condition's observed day-5 effect, including for held-out conditions; forecasting uses its observed day-4 effect. The control profile is averaged over days 2–4 for forecasting. Hagai adaptation supplies the six-hour effect. Harvest-only prediction instead uses a GO-Jaccard profile derived from perturbation identity, with no observed test effect as input. Paul fits the analytic kernel directly to preordered gene curves. These settings test, respectively, anchored reconstruction, future response conditional on an observed anchor, endpoint prediction, and curve representation.

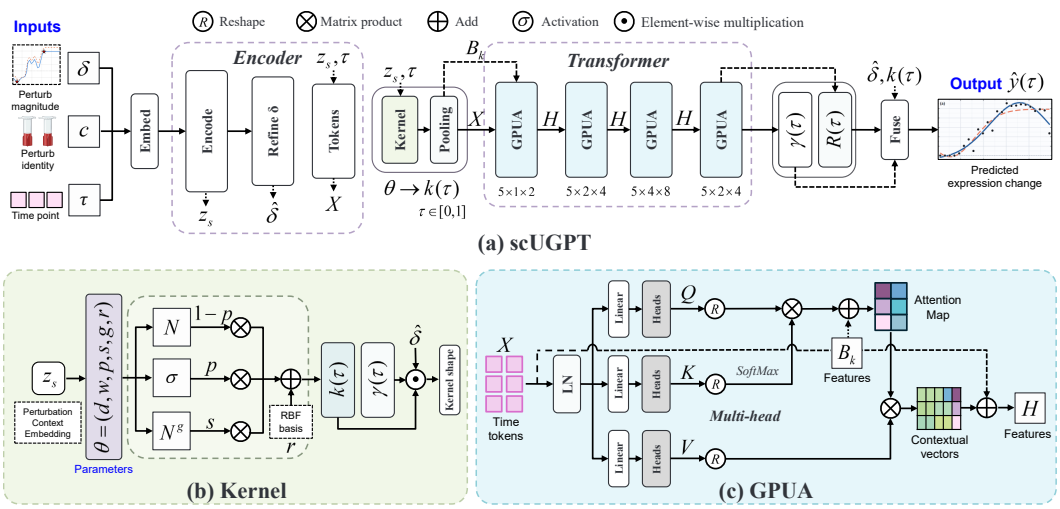


Figure 1: **Overview of scUGPT.** (a) Static profile $\boldsymbol{\delta}$, control-expression profile $\mathbf{c}_0$, optional metadata c, and coordinate τ define the inputs; their construction follows the evaluation protocol. A gene-wise process kernel is fused with a gated process-attention representation and a residual prediction. (b) The kernel combines a transient response, a persistent component, and a signed late component. (c) Gene-wise Process Unfolding Attention (GPUA) uses multi-head attention, residual connections, and feed-forward layers to couple perturbation context to temporal tokens, and the resulting gate modulates the refined perturbation effect before fusion with the residual.

3.2 Gene-Wise Process Kernel

The analytic process kernel parameterizes the primary path of scUGPT. For gene g it is

$$k_g(\tau) = (1-p_g)\exp\left[-\frac{1}{2}\left(\frac{\tau-d_g}{w_g}\right)^2\right] + p_g\,\sigma\left(\frac{\tau-d_g}{w_g}\right) + s_g\exp\left[-\frac{1}{2}\left(\frac{\tau-d_g-q_g}{\rho_w w_g}\right)^2\right] + r_g\sum_{j=1}^{J}\alpha_{gj}(z_s)\phi_j(\tau), \tag{1}$$

where $d_g \in [0,1]$ locates the response in process position, $w_g > 0$ is its width, and $p_g \in [0,1]$ interpolates between a transient bump and a maintained step. The gap $q_g > 0$ and signed amplitude $s_g \in (-1,1)$ place a second lobe after the first, allowing a late reversal or reinforcement without imposing symmetry. This lobe uses the fixed width ratio $\rho_w = 0.85$ relative to w_g. The ratio is a structural constant rather than a fitted parameter; its position $d_g + q_g$ and amplitude s_g remain learned. The final term is a perturbation-conditioned RBF dictionary with per-gene gain r_g and context z_s. The analytic Scalar RBF is the nested case $p_g = s_g = r_g = 0$ (Broomhead & Lowe, 1988). The Fixed RBF-basis baseline uses eight fixed centres, a learned scale, and a context-conditioned coefficient map, so it approximates the same shapes through superposition. Persistence distinguishes transient from maintained responses, and a negative s_g permits a late reversal. Together, the terms represent the four kinetic classes scored below: transient, persistent, biphasic, and delayed. The normalized static and control profiles and metadata are projected into a shared context z_s. The static profile is refined as $\widehat{\boldsymbol{\delta}} = \boldsymbol{\delta} + f_{\text{fill}}(\boldsymbol{\delta}) + f_{\text{refine}}(z_s)$, where the two maps fill unobserved entries and adjust the profile to the current context. Gene-wise Process Unfolding Attention (GPUA), a gated process-attention stack (Vaswani et al., 2017), maps the anchor and temporal tokens to a gate $\gamma_g(\tau)$ that controls how strongly each gene follows its kernel at each coordinate. The primary prediction is

$$\mathbf{h}(\tau) = \boldsymbol{\gamma}(\tau) \odot \mathbf{k}(\tau) \odot \widehat{\boldsymbol{\delta}}. \tag{2}$$

We pass the kernel path through dilated lag and program mixing, allowing one gene's curve to depend on other genes at nearby coordinates. The learned coupling agrees across the three evaluated seeds and tracks genes that respond together; Appendix A.4 tests whether it also encodes temporal order. For the GSE213069 neural runs, the model adds a bidirectional residual $\mathbf{r}(\tau)$ through a learned scalar gate λ, giving $\widehat{\mathbf{y}}(\tau) = \mathbf{h}_{\text{lag}}(\tau) + \lambda\mathbf{r}(\tau)$. The parameters describe kernel shape before attention, lag, gene mixing, and residual correction. Consequently, d_g is a kernel centre, not the onset or peak

of the final curve; timing metrics are computed from $\widehat{\mathbf{y}}$, and held-out parameter checks appear in Appendix D.

3.3 Objectives and Training

The neural runs use normalized Huber losses,

$$\mathcal{L} = \mathcal{L}_{\text{point}} + 0.5\mathcal{L}_{\text{shape}} + 0.25\mathcal{L}_{\text{amp}} + \kappa\mathcal{L}_{\text{primary}}, \tag{3}$$

where $\mathcal{L}_{\text{point}}$ and $\mathcal{L}_{\text{shape}}$ are Huber losses on normalized levels and adjacent first differences, $\mathcal{L}_{\text{amp}}$ is a Huber loss on the normalized amplitude span, and $\mathcal{L}_{\text{primary}}$ is a Huber loss between the normalized primary-path output after lag and program mixing and the normalized target curve. The primary-path term constrains the analytic path while the residual captures the remainder. The training implementation uses $\kappa = 0.45$ in the small-sample GSE branch (fewer than 48 training trajectories) and 0.2 otherwise. It omits this term when either the kernel or residual path is disabled, including the No residual and baseline rows. Appendix B.4 sweeps $\kappa \in \{0, 0.05, 0.2, 0.45, 1\}$ on GSE213069: day-5 MSE varies by about 1.3% across the grid, with the smallest reported mean at $\kappa = 1$. AdamW selects checkpoints by validation full-curve MSE on the available time grid, with a maximum of 200 epochs and patience 12; the future-day branch has a three-day training budget. Timing scores are computed after checkpoint selection.

3.4 Evaluation Protocol

We report each estimand separately. *Rate of change* uses MSE and Pearson correlation of adjacent first differences. *Timing* uses onset, defined as the coordinate where a curve first reaches 10% of its peak amplitude, gene-wise peak time, and derivative zero-crossings. Turning-point MAE is conditional on a valid observed and predicted zero-crossing; missing predicted crossings are not imputed. A constant forecast has no unique peak or turning point, so those metrics are unavailable for that row. *Direction* is sign agreement for first differences on above-median changes. *Harvest-only perturbation prediction* uses Pearson-of-Δ and top-20 DE overlap at one harvest time and does not evaluate dynamics. For the two real GSE213069 protocols we additionally report a level-calibrated readout, denoted scUGPT-cal, which is a fixed convex combination of the raw scUGPT curve and the supplied static anchor,

$$\widehat{y}_{\text{cal}} = \alpha\widehat{y}_{\text{scUGPT}} + (1-\alpha)\delta_{\text{anchor}}. \tag{4}$$

We select $\alpha \in \{0, .25, .5, .75, 1\}$ by mean validation level MSE on the fixed condition split and freeze it before reading the test targets, giving $\alpha = 0.75$ for reconstruction and $\alpha = 0.5$ for future-day forecasting. Because the anchor is constant over time, any $\alpha > 0$ preserves the sign and Pearson correlation of the model's first differences; calibration rescales difference amplitudes and can change both level MSE and first-difference MSE.

4 Experiments

Table notation. Bold blue marks the best displayed mean within each comparison; ties at the reported precision are marked together. The marking does not imply statistical significance.

4.1 Experimental Setup

Datasets. We evaluate six public resources, each assigned to a distinct estimand. Paul et al. (Paul et al., 2015) provides densely sampled erythroid pseudotime for fitting and scoring gene-wise rate curves across thousands of genes. Hagai et al. (Hagai et al., 2018) provides experimental hours under immune stimulation, with experimental hours as the coordinate. Ishikawa et al. (Ishikawa et al., 2023) contributes GSE213069, a four-day scCRISPR time course with perturbation identity and repeated measurements; it is the only dataset supporting held-out-condition reconstruction and one-day forecasting. Adamson, Norman, and Replogle (Adamson et al., 2016; Norman et al., 2019; Replogle et al., 2022) provide three large harvest-only Perturb-seq screens for unseen single perturbations. We exclude single-time-point screens from process-axis claims and screens without perturbation labels from timing claims.

Real trajectories. For Paul et al. (Paul et al., 2015), curve values use a 10^4-normalized, `log1p`-transformed copy. On a separate copy we compute the supplied axis with Scanpy's Zheng17 recipe (20-PC PCA, 10-neighbour graph, diffusion map), root DPT in the stem-cell cluster, and bin curves into 24 positions with at least five cells per occupied bin and six occupied bins, interpolating gaps. The rate-of-change block in Table 1 uses independent per-gene seven-parameter analytic process-kernel fits and the four-parameter Scalar RBF smoother on erythroid curves. After bin eligibility, it retains curves with $\max_t |y_t - y_0| \geq 0.05$, giving $n = 3{,}080$. Operational timing and class scores use all genes with at least six occupied bins, 3,451 curves per lineage; monocyte bins are omitted. No perturbation is held out. Hagai et al. (Hagai et al., 2018) contributes experimental hours $0, 2, 4, 6$ for leave-one-donor and leave-one-stimulus analyses, and each panel states whether its axis is wall-clock time or diffusion pseudotime. Paul is a descriptive representation diagnostic. We additionally evaluate held-out bins against the seven-parameter Two-RBF and Fixed RBF-basis (7-coef.) comparators, score turning points on a common-valid subset, vary the supplied DPT construction, and test known-parameter recovery (Appendix B.1). These checks do not hold out perturbations or biological replicates. For GSE213069, the released matrices contain four collection days (days 2–5 after transduction), two guide RNAs for each of 23 target genes, and four control guides. Cells are assigned to their dominant detected guide, pooled at the target-gene level, and normalized by library size followed by `log1p`; AAVS1 and non-cutting guides define the control mean. For both protocols, 96 control-variable genes are selected from days 2–4 only, keeping the feature set free of test-day information. In condition-held-out reconstruction, day-5 effects anchor the held-out conditions while day 5 remains a target. In the future-day protocol, day 5 is excluded from training and checkpoint selection, and validation day-5 labels are used only for blend selection. The same leakage-controlled bundle is used for every method and seed, and its SHA-256 prefix and suffix are recorded in Appendix A.

Harvest-only Perturb-seq. Adamson CRISPRi, Norman CRISPRa, and Replogle K562 essential are evaluated with the GEARS simulation split at seed 1 (Adamson et al., 2016; Norman et al., 2019; Replogle et al., 2022; Roohani et al., 2024). For an unseen gene, the input is its GO-Jaccard profile against measured genes; the observed test Δx is used only as the target. These datasets contain one harvest time, so they evaluate endpoint extrapolation rather than process unfolding.

Baselines. We compare against Fixed RBF-basis; the no-residual variant, which retains process attention and the kernel; GRU; Neural ODE; a vanilla Transformer; MLP-time; scGen-style MLP; a static broadcast that repeats the supplied anchor across the axis; and ridge and training-mean reference baselines. The scGen-style MLP is a deterministic encoder–decoder comparator inspired by scGen, rather than the original variational implementation. Every learned comparator trains on the same bundle with the same splits, feature set, and early-stopping rule.

Hardware and configuration. All GSE213069 neural runs use hidden dimension 128, four attention heads, three layers, batch size 8, and learning rate 10^{-3}. The small-sample branch uses weight decay 3×10^{-4} and residual dropout 0.35. Training, calibration, and evaluation for the revision ran on a single NVIDIA RTX 4090 GPU with 25 CPU cores and 90 GB of host memory, under Python 3.10, PyTorch 2.1.2 with CUDA 12.1, and NumPy and SciPy for the analytic kernel fits. No run required more than one GPU.

4.2 Process-Axis Accuracy

On Paul erythroid pseudotime, the in-sample analytic kernel reaches Δ-Pearson 0.534 versus 0.459 for the four-parameter Scalar RBF (Table 1). This comparison assesses representation with unequal parameter counts. The matched-capacity held-out-bin check gives MSE 0.12186 versus 0.18270 for the seven-parameter Two-RBF model on the original axis. The seven-coefficient Fixed RBF-basis performs better than the process kernel under the 20-neighbour axis, showing that the readout depends on the supplied axis and basis choice. In the common-valid turning-point check, both models and the observation have crossings for 2,394 of 3,080 active curves; mean error is 0.08224 versus 0.20781, and median error is 0.02920 versus 0.19356. Appendix B.1 reports eligibility, convergence, and axis dependence. The difference is largest on curves with a change of direction. On non-monotonic erythroid genes, class accuracy rises from 0.659 to 0.831, and on neutrophils it rises from 0.638 to 0.812 (Table 2). Figure 2 shows the corresponding fits. The signed late lobe preserves the decay of

Table 1: **Core process-axis results on Paul erythroid pseudotime.** Rate metrics use all $n = 3{,}080$ eligible genes. Turning-point statistics use each method's own valid crossings, so they are not a common-subset comparison. The kernel column is an independent per-gene analytic fit with no neural component.

Metric	Process Kernel (Analytic)	Scalar RBF
Δ-Pearson ↑	**0.534**	0.459
Δ-MSE ↓	**0.0091**	0.0099
Median turning-point error ↓	**0.037**	0.223
Valid turning-point curves	2,768	2,496
Genes won by Δ-Pearson ↑	**84.3%**	15.7%
Genes won by Δ-MSE ↑	**85.7%**	14.3%

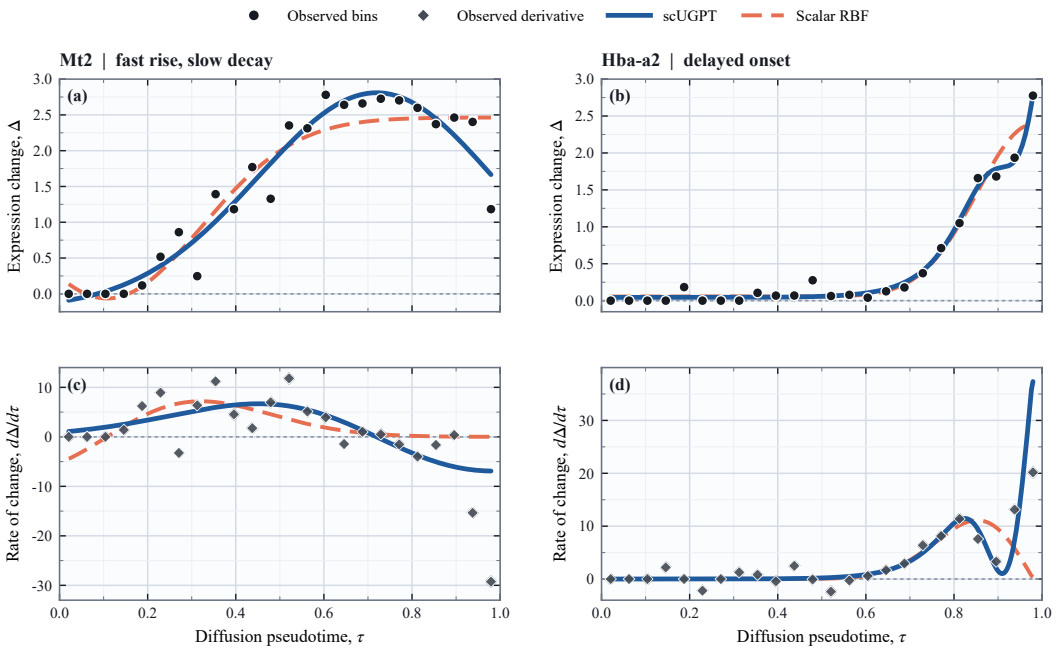


Figure 2: Paul erythroid DPT examples selected by observed shape rather than model error. *Mt2* exhibits a fast rise followed by slower decay, and *Hba-a2* exhibits delayed globin activation. The upper row overlays both fitted trajectories on the observed bins, and the lower row compares their derivatives. Black circles denote observed bins, dark-gray diamonds denote interval-wise observed derivative estimates, blue solid lines the analytic process-kernel fit (labelled SCUGPT in the panel), and orange dashed lines Scalar RBF. Observed derivatives are left unconnected because adjacent differences are defined on discrete intervals rather than as a continuous curve.

Mt2 and the delayed pulse of *Hba-a2*, whereas a symmetric RBF tends to flatten both into a single broad bump. A symmetric smoother has no term for a reversal, so it uses one width to represent both the rise and the fall and cannot maintain a plateau.

4.3 Real Time-Series Condition Hold-Out

GSE213069 tests the model on a measured perturbation time course rather than a pseudotime fit. We hold out entire target conditions, supply each held-out condition's day-5 effect as an endpoint anchor for the reconstruction protocol, train on all four observed days for the remaining targets, and repeat training over seeds 42, 43, and 44 on the fixed reference split (seed 2026). Gene selection is frozen from days 2–4. The static broadcast provides a reference for this endpoint-assisted setting. The calibrated readout SCUGPT-cal reaches full-curve MSE 0.00602 ± 0.00021 and full-curve Pearson 0.782 ± 0.009, while uncalibrated SCUGPT reaches MSE 0.00738 ± 0.00026 in the independently rerun three-seed bundle (Table 3). The shape column isolates changes along the process axis, whereas the level metrics summarize reconstruction under the supplied anchor. The future-day protocol evaluates extrapolation without day-5 labels during training and checkpoint selection. The

Table 2: In-sample analytic kernel fits on Paul diffusion pseudotime, process kernel / Scalar RBF. The n column is the total number of curves; active-curve counts are 2,736 and 2,860 for erythroid and neutrophil. Turning-point MAE is conditional on valid crossings. Class accuracy uses curve-derived labels. Monocyte bins were too sparse and are omitted.

Lineage	Subset	n	Δ-P ↑	Turn. MAE ↓	Peak MAE ↓	Class ↑
Erythroid	all	3451	**0.608** / 0.558	**0.097** / 0.235	**0.180** / 0.342	**0.698** / 0.398
Erythroid	non-monotonic	1964	**0.577** / 0.513	**0.061** / 0.135	**0.090** / 0.196	**0.831** / 0.659
Neutrophil	all	3451	**0.643** / 0.550	**0.064** / 0.244	**0.088** / 0.316	**0.812** / 0.638

Each pair is analytic process-kernel fit / Scalar RBF smoother.

Table 3: GSE213069 under both protocols on the fixed main split (seed 2026), using the same leakage-controlled bundle and the same 96 genes selected from days 2–4. Left: condition hold-out, where the held-out condition's last observed day is a static anchor and the full four-day curve is the target (seeds 42, 43, 44). Right: future day, where day 5 is excluded from training and checkpoint selection (seeds 7, 21, 84). Pearson and Day-5 Pearson measure expression levels; Shape and Δ-Pearson measure adjacent first differences. Only scUGPT-cal uses the additional anchor blend.

	Reconstruction (endpoint-assisted)			Future day (strict)		
Method	MSE ↓	Pearson ↑	Shape ↑	MSE ↓	Pearson ↑	Δ-r ↑
scUGPT-cal	**0.00602 ± 0.00021**	**0.782 ± 0.009**	0.258 ± 0.074	**0.00799 ± 0.00014**	0.703 ± 0.005	**0.261 ± 0.041**
scUGPT	0.00738 ± 0.00026	0.723 ± 0.012	0.258 ± 0.074	0.00909 ± 0.00035	0.654 ± 0.021	**0.261 ± 0.041**
No residual	0.00771 ± 0.00056	0.704 ± 0.024	0.228 ± 0.101	0.01006 ± 0.00045	0.600 ± 0.024	0.156 ± 0.032
Fixed RBF-basis	0.01028 ± 0.00004	0.568 ± 0.002	**0.409 ± 0.005**	0.01663 ± 0.00126	0.323 ± 0.047	0.025 ± 0.042
GRU	0.00967 ± 0.00011	0.605 ± 0.007	0.340 ± 0.004	0.01370 ± 0.00040	0.541 ± 0.003	0.010 ± 0.016
Neural ODE	0.00935 ± 0.00015	0.620 ± 0.008	0.363 ± 0.006	0.01562 ± 0.00026	0.521 ± 0.009	−0.003 ± 0.069
Transformer	0.01029 ± 0.00053	0.583 ± 0.026	0.144 ± 0.067	0.01317 ± 0.00040	0.410 ± 0.038	0.055 ± 0.020
MLP-time	0.01042 ± 0.00016	0.565 ± 0.011	0.100 ± 0.005	0.01229 ± 0.00051	0.477 ± 0.029	−0.075 ± 0.017
scGen-style MLP	0.01056 ± 0.00015	0.557 ± 0.011	0.104 ± 0.010	0.01573 ± 0.00087	0.380 ± 0.011	−0.096 ± 0.012
Static broadcast	0.00725 ± 0.00000	0.763 ± 0.000	0.000 ± 0.000	0.00902 ± 0.00000	**0.707 ± 0.000**	0.000 ± 0.000

reconstruction metrics separate level and derivative behavior. No residual reaches level scores close to the full model, while Fixed RBF-basis reaches the highest derivative correlation, 0.409 ± 0.005, on this four-point grid. The future-day protocol evaluates the same components under a different estimand. Table 5 further tests the kernel and residual paths. A zero-anchor information-ablation is reported in Appendix B.3; it isolates the contribution of the supplied endpoint while keeping the target, feature selection, checkpoint rule, and seeds fixed.

Future-day extrapolation. To test extrapolation rather than interpolation, we remove day 5 from model training and checkpoint selection and reserve validation day-5 labels for the validation-based blend selection. The static input is the held-out condition's observed day-4 effect and the control input is averaged over days 2–4. The main condition split uses seed 2026, with 13 training targets, 5 validation targets, and 5 held-out targets, namely `ID1`, `SOX2`, `TRIM24`, `TRIM25`, and `ZIC2`; test targets are excluded from checkpoint selection. All feature selection is restricted to days 2–4. On the fixed main split (seed 2026), scUGPT reaches the highest day-4-to-day-5 change correlation among the tested common-interface models: 0.261 ± 0.041, versus 0.156 ± 0.032 for No residual and 0.025 ± 0.042 for Fixed RBF-basis (Table 3). Raw scUGPT MSE is 0.00909 ± 0.00035, close to the static baseline's 0.00902, while validation-calibrated scUGPT-cal reaches 0.00799 ± 0.00014. Calibration uses validation day-5 labels and is reported separately from the uncalibrated forecast. Training seeds are optimization replicates, not biological samples. Averaging per-seed predictions before scoring gives an ensemble derivative correlation of 0.298, higher than the mean per-seed value of 0.261 in Table 3; the two quantities differ because the ensemble correlation is computed once on averaged predictions rather than averaged over seeds. Across four additional condition splits, derivative Pearson averages 0.007 ± 0.122, and static broadcast has lower mean day-5 MSE. Appendix B.2 gives the condition-level errors and split stability; test conditions are the resampling units.

4.4 Wall-Clock Adaptation and Harvest-Only Screens

Paul pseudotime and GSE213069 use an inferred or experimental process coordinate supplied to the model. Hagai supplies experimental hours, allowing an endpoint-assisted, in-fold adaptation diagnostic. Leave-one-donor averages all available LPS and PIC curves, where scUGPT reaches

Table 4: Hagai 2018 in-fold adaptation fit at hours $0, 2, 4, 6$. LOD averages the LPS and PIC curves in leave-one-donor folds and LOS both held-LPS and held-PIC directions. The held-out curve's six-hour effect is part of the static input and the same fold supplies checkpoint selection and scoring. Values are mean $\pm$ standard deviation.

Method	LOD Δ-Pearson	LOS Δ-Pearson	LOS Pearson
scUGPT	**0.874 ± 0.016**	**0.522 ± 0.178**	**0.788 ± 0.054**
Fixed RBF-basis	0.867 ± 0.006	0.231 ± 0.056	0.438 ± 0.025
scGen-style MLP	0.501 ± 0.032	0.040 ± 0.033	0.148 ± 0.059
MLP-time	0.560 ± 0.071	0.033 ± 0.019	0.199 ± 0.036
GEARS	0.000 ± 0.000	0.000 ± 0.000	0.071 ± 0.024

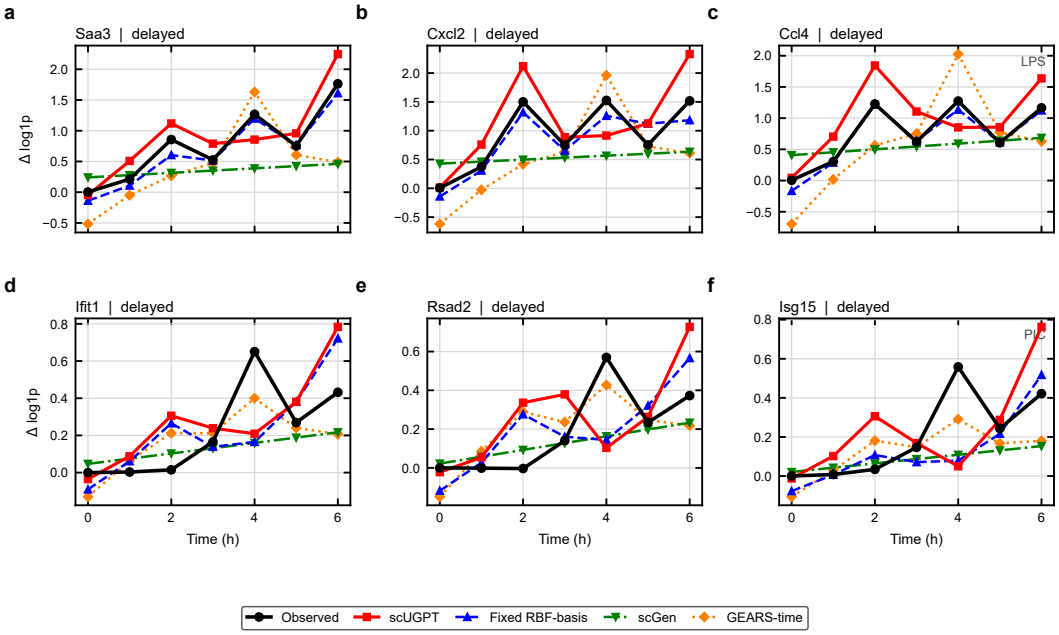


Figure 3: Hagai in-fold adaptation on experimental hours. The scored fold also supplies checkpoint selection and its six-hour endpoint is an input. Panels a–c use LPS and d–f use PIC. Each panel shows the observation and the locally implemented methods trained for this diagnostic: scUGPT, Fixed RBF-basis, scGen-style MLP, and a time-conditioned GEARS variant. Each panel uses the three largest-amplitude responders to that stimulus in the held-out donor; the label records the observed kinetic behaviour. Leave-one-donor mean MSE is 0.127 for scUGPT, 0.139 for Fixed RBF-basis, and 0.28 for the local comparison models.

Δ-Pearson 0.874 ± 0.016 versus 0.867 ± 0.006 for Fixed RBF-basis. Leave-one-stimulus averages the held-LPS and held-PIC directions, where scUGPT reaches 0.522 ± 0.178 versus 0.231 ± 0.056 for Fixed RBF-basis and 0.040 ± 0.033 for scGen-style MLP (Table 4). The static input includes the held-out curve's six-hour effect, and checkpoint selection uses the fold later scored. These values measure in-fold adaptation fit, not generalization to an independently scored fold. A process model and a harvest-only screen answer different estimands, so we report the latter as an endpoint diagnostic under the GEARS unseen-single split. scUGPT reaches a Norman top-20 DE overlap of 0.361 with a 48.6% win rate across hold-outs (Appendix Table 14). Complete top-20 overlap results are in Table 14, and Pearson-of-Δ results are in Table 13. These results describe single-harvest behavior separately from the process-axis evaluation.

4.5 Ablations

Table 5 compares the full model with two single-path removals. No residual retains the attention-modulated kernel, lag, and gene mixing. No process kernel retains the attention-driven residual, with the primary output and kernel-derived attention bias set to zero. Both omit the primary-path auxiliary loss. Removing the kernel changes reconstruction MSE from 0.00738 to 0.01200; removing the residual changes future-day derivative correlation from 0.261 to 0.156. The archived joint-removal

row has no matching forward configuration in the recovered source and is recorded separately in Appendix A.3; it is excluded from component attribution.

Table 9 fixes one kernel term at a time and retrains the remaining terms. The preferred value depends on the metric: persistence favours different reconstruction and forecasting settings, while zero late amplitude or dictionary gain remains competitive.

Table 5: Structural ablation on GSE213069. The two single-path variants use the full model's splits, seeds, training limits, and checkpoint rule. Forward paths and the archived joint-removal result are specified in Appendix A.3. Values are mean ± standard deviation over three seeds: 42, 43, and 44 for reconstruction, and 7, 21, and 84 for future-day prediction. The no-residual row retains attention and the process kernel and is the same variant reported in Table 3.

	Reconstruction			Future day		
Variant	MSE ↓	Pearson ↑	Shape ↑	MSE ↓	Pearson ↑	Δ-r ↑
Full model	**0.00738 ± 0.00026**	**0.723 ± 0.012**	**0.258 ± 0.074**	**0.00909 ± 0.00035**	**0.654 ± 0.021**	**0.261 ± 0.041**
No residual	0.00771 ± 0.00056	0.704 ± 0.024	0.228 ± 0.101	0.01006 ± 0.00045	0.600 ± 0.024	0.156 ± 0.032
No kernel	0.01200 ± 0.00025	0.470 ± 0.019	0.182 ± 0.079	0.01547 ± 0.00023	0.192 ± 0.043	0.069 ± 0.030

5 Discussion and Limitations

The fixed main GSE split supports future-day change prediction against the tested common-interface comparators. Across four additional condition splits, derivative Pearson averages 0.007 ± 0.122 and static broadcast has lower mean day-5 MSE (Appendix B.2). Endpoint-assisted reconstruction includes the supplied day-5 anchor in its four-day target, and the calibrated row includes an additional validation-selected blend. Paul fits describe observed trajectories conditional on an axis inferred from all cells. Hagai reports in-fold adaptation fit using the six-hour endpoint. On held-out GSE conditions, kernel-parameter stability is uneven and kernel exchanges yield uncertain error changes (Appendix D). Kernel parameters are shape descriptors rather than identifiable biophysical rates; the results do not establish a split-independent forecasting advantage.

AI Use Statement

We used generative AI tools (DeepSeek Harness, Claude, and OpenAI Codex) for model and evaluation-code implementation, development of the experimental methodology, literature retrieval and summarisation, manuscript organisation and editing, interpretation of results, and figure creation and revision. The analyses use public single-cell datasets; generative AI was not used to generate the experimental data. The authors are responsible for the manuscript, its claims, and the accompanying code and figures.

References

Britt Adamson, Thomas M. Norman, Marco Jost, Min Y. Cho, James K. Nuñez, Yuwen Chen, Jacqueline E. Villalta, Luke A. Gilbert, Max A. Horlbeck, Marco Y. Hein, Ryan A. Pak, Andrew N. Gray, Carol A. Gross, Atray Dixit, Oren Parnas, Aviv Regev, and Jonathan S. Weissman. A multiplexed single-cell CRISPR screening platform enables systematic dissection of the unfolded protein response. *Cell*, 167(7):1867–1882.e21, 2016. doi: 10.1016/j.cell.2016.11.048.

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: 1657–1661, 2025. doi: 10.1038/s41592-025-02772-6.

Ido Amit, Manuel Garber, Nicolas Chevrier, et al. Unbiased reconstruction of a mammalian transcriptional network mediating pathogen responses. *Science*, 326(5950):257–263, 2009. doi: 10.1126/science.1179050.

Volker Bergen, Marius Lange, Stefan Peidli, et al. Generalizing RNA velocity to transient cell states through dynamical modeling. *Nature Biotechnology*, 38:1408–1414, 2020. doi: 10.1038/s41587-020-0591-3.

David S. Broomhead and David Lowe. Multivariable functional interpolation and adaptive networks. *Complex Systems*, 2(3):321–355, 1988. URL https://www.complex-systems.com/abstracts/v02_i03_a05/.

Charlotte Bunne, Stefan G. Stark, Gabriele Gut, et al. Learning single-cell perturbation responses using neural optimal transport. *Nature Methods*, 20:1759–1768, 2023. doi: 10.1038/s41592-023-01969-x.

Alan B. Cantor and Stuart H. Orkin. Transcriptional regulation of erythropoiesis: an affair involving multiple partners. *Oncogene*, 21(21):3368–3376, 2002. doi: 10.1038/sj.onc.1205326.

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.

Atray Dixit, Oren Parnas, Biyu Li, et al. Perturb-seq: dissecting molecular circuits with scalable single-cell RNA profiling of pooled genetic screens. *Cell*, 167(7):1853–1866.e17, 2016. doi: 10.1016/j.cell.2016.11.038.

Rossin Erbe, Genevieve Stein-O'Brien, and Elana J. Fertig. Transcriptomic forecasting with neural ordinary differential equations. *Patterns*, 4(8):100793, 2023. doi: 10.1016/j.patter.2023.100793.

Tzachi Hagai, Xi Chen, Ricardo J. Miragaia, et al. Gene expression variability across cells and species shapes innate immunity. *Nature*, 563:197–202, 2018. doi: 10.1038/s41586-018-0657-2.

Laleh Haghverdi, Maren Buettner, F. Alexander Wolf, et al. Diffusion pseudotime robustly reconstructs lineage branching. *Nature Methods*, 13:845–848, 2016. doi: 10.1038/nmeth.3971.

Masato Ishikawa, Seiichi Sugino, Yoshie Masuda, Yusuke Tarumoto, Yusuke Seto, Nobuko Taniyama, Fumi Wagai, Yuhei Yamauchi, Yasuhiro Kojima, Hisanori Kiryu, Kosuke Yusa, Mototsugu Eiraku, and Atsushi Mochizuki. RENGE infers gene regulatory networks using time-series single-cell RNA-seq data with CRISPR perturbations. *Communications Biology*, 6:1290, 2023. doi: 10.1038/s42003-023-05594-4.

Gioele La Manno, Ruslan Soldatov, Amit Zeisel, et al. RNA velocity of single cells. *Nature*, 560:494–498, 2018. doi: 10.1038/s41586-018-0414-6.

Marius Lange, Volker Bergen, Michal Klein, et al. CellRank for directed single-cell fate mapping. *Nature Methods*, 19:159–170, 2022. doi: 10.1038/s41592-021-01346-6.

Zaikang Lin, Sei Chang, Aaron Zweig, Minseo Kang, Fabian J. Theis, Elham Azizi, and David A. Knowles. Interpretable neural ODEs for gene regulatory network discovery under perturbations. *arXiv preprint arXiv:2501.02409*, 2025. doi: 10.48550/arXiv.2501.02409. URL https://arxiv.org/abs/2501.02409.

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

Thomas M. Norman, Max A. Horlbeck, Joseph M. Replogle, et al. Exploring genetic interaction manifolds constructed from rich single-cell phenotypes. *Science*, 365(6455):786–793, 2019. doi: 10.1126/science.aax4438.

Franziska Paul, Ya'ara Arkin, Amir Giladi, et al. Transcriptional heterogeneity and lineage commitment in myeloid progenitors. *Cell*, 163(7):1663–1677, 2015. doi: 10.1016/j.cell.2015.11.013.

Joseph M. Replogle, Reuben A. Saunders, Angela N. Pogson, et al. Mapping information-rich genotype-phenotype landscapes with genome-scale Perturb-seq. *Cell*, 185(14):2559–2575.e28, 2022. doi: 10.1016/j.cell.2022.05.013.

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

Manu Setty, Vaidotas Kiseliovas, Jacob Levine, et al. Characterization of cell fate probabilities in single-cell data with Palantir. *Nature Biotechnology*, 37:451–460, 2019. doi: 10.1038/s41587-019-0068-4.

Damian Szklarczyk, Rebecca Kirsch, Mikaela Koutrouli, et al. The STRING database in 2023: protein–protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic Acids Research*, 51(D1):D638–D646, 2023. doi: 10.1093/nar/gkac1000.

Cole Trapnell, Davide Cacchiarelli, Jonna Grimsby, et al. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nature Biotechnology*, 32: 381–386, 2014. doi: 10.1038/nbt.2859.

Ashish Vaswani, Noam Shazeer, Niki Parmar, et al. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pp. 5998–6008, 2017. URL `https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html`.

Grace Hui Ting Yeo, Sachit D. Saksena, and David K. Gifford. Generative modeling of single-cell time series with PRESCIENT enables prediction of cell trajectories with interventions. *Nature Communications*, 12:3222, 2021. doi: 10.1038/s41467-021-23518-w.

Jiaqi Zhang, Erica Larschan, Jeremy Bigness, and Ritambhara Singh. scNODE: generative model for temporal single cell transcriptomic data prediction. *Bioinformatics*, 40(Supplement_2):ii146–ii154, 2024. doi: 10.1093/bioinformatics/btae393.

A Protocol Details

A.1 Separation of Estimands

Table 6 maps the real-data evaluations to their sources. A separate synthetic known-parameter recovery diagnostic assesses analytic-kernel identifiability; it is not a benchmark of unseen biological perturbations.

Table 6: Data provenance and estimands.

Setting	Source	Protocol and target
Paul 2015	Real scRNA-seq	Erythroid DPT analytic kernel fit ($n = 3{,}080$ for the rate block; 3,451 curves per lineage in the operational block); neutrophil DPT operational; no unseen knockout.
Hagai 2018	Real scRNA-seq	Wall-clock leave-one-donor and leave-one-stimulus in-fold adaptation diagnostics; the coordinate is experimental hour.
GSE213069	Real scCRISPR	Four collection days (days 2–5), 23 target-level perturbations, 96 control-variable genes selected from days 2–4; endpoint-assisted condition-held-out reconstruction and day-4-to-day-5 extrapolation.
Perturb-seq	Real screens	Adamson, Norman, and Replogle unseen-single harvest; top-20 DE overlap and Pearson-of-Δ; no time-course target.

GSE213069 preprocessing manifest. The four processed GEO matrices and guide-name workbook were downloaded from GEO accession GSE213069 and archived with the paper's code. The final target-level bundle has 23 perturbation conditions, one hiPSC lineage, four time bins, and 96 expression genes selected from control cells on days 2–4 only. Its SHA-256 prefix/suffix is `fa939be5...d0cb100`; the full checksum is stored in both experiment manifests and the local bundle is included in the release directory. The endpoint-assisted reconstruction run observes day 5 for training conditions and supplies the held-out condition's day-5 anchor. The strict forecast uses day 4 and excludes day 5 from every training and early-stopping tensor. Both manifests record the condition split, cell-assignment rule, seeds, convergence configuration, and best-checkpoint criterion.

Implementation provenance. The GSE213069 runs use the archived remote source snapshot in `data/source_snapshot/`; `MANIFEST.tsv` records the SHA-256 hash of each model, metric, and training script used to produce the reported checkpoints. Fixed RBF-basis, GRU, Neural ODE,

static broadcast, and the scGen-style MLP use local implementations of the common trajectory interface. The scGen-style MLP has deterministic encoders and a decoder; it omits the variational posterior and KL objective of the original scGen. Comparisons refer to these specified implementations. The GSE configuration uses hidden dimension 128, four attention heads, three layers, batch size 8, and learning rate 10^{-3}. The small-sample branch uses weight decay 3×10^{-4} and residual dropout 0.35. All GSE neural runs use a 200-epoch cap, patience 12, and validation-MSE checkpoint selection. The reruns reported here were executed on a single NVIDIA RTX 4090 with CUDA 12.1, PyTorch 2.1.2, and Python 3.10; the manifests record the device and package versions. The Paul rate artifact and its independent amplitude-filtered script are archived under `data/paul15/` and `data/source_snapshot/`.

Additional checks. Appendix B.1 gives the held-out-bin rule, common-crossing coverage, parameter-matched comparisons, and recovery protocol. Appendix B.2 reports condition-level GSE uncertainty. The zero-anchor run retrains every method with zero static inputs for training, validation, and test; it is an information-ablation diagnostic, not an identity-conditioned zero-shot predictor.

A.2 Training Procedure

Algorithm 1 (Table 7) restates the procedure from the Method section. The kernel parameters, attention stack, residual path, and gate are optimized jointly, while the coordinate τ remains fixed.

Table 7: Training and prediction for the neural scUGPT model. The analytic Paul fits use a separate per-gene fitting procedure.

Algorithm 1: trajectory-conditioned gene-wise response prediction

Inputs: static profile $\boldsymbol{\delta}$, control expression $\mathbf{c}_0$, optional metadata c, supplied grid $\{\tau_t\}$; target curve $\mathbf{Y}$ during training.

Predict$(\boldsymbol{\delta}, \mathbf{c}_0, c, \{\tau_t\})$:

Normalize static and control profiles using the fitted training statistics.
$z_s \leftarrow \mathrm{LayerNorm}(f_\delta(\boldsymbol{\delta}) + f_0(\mathbf{c}_0) + f_c(c))$.
Predict constrained kernel parameters $(d_g, w_g, p_g, q_g, s_g, r_g)$ from z_s.
Compute $k_g(\tau_t)$ from equation 1 and its amplitude-weighted gene pool.
$\widehat{\boldsymbol{\delta}} \leftarrow \boldsymbol{\delta} + f_{\mathrm{fill}}(\boldsymbol{\delta}) + 0.25 \tanh f_{\mathrm{refine}}(z_s)$.
Form the static token and temporal tokens $z_s + f_\tau(\tau_t)$.
Apply the attention stack with the kernel-derived bias to the joint token sequence.
Map temporal tokens to gate $\boldsymbol{\gamma}(\tau_t)$ and residual $\mathbf{r}(\tau_t)$.
$\mathbf{h}(\tau_t) \leftarrow \boldsymbol{\gamma}(\tau_t) \odot \mathbf{k}(\tau_t) \odot \widehat{\boldsymbol{\delta}}$.
Apply the lag stack, gene mixing, lag hop, and second gene mixing to obtain $\mathbf{h}_{\mathrm{lag}}$.
$\widehat{\mathbf{y}}(\tau_t) \leftarrow \mathbf{h}_{\mathrm{lag}}(\tau_t) + \sigma(\lambda_{\mathrm{raw}})\mathbf{r}(\tau_t)$.
return the final curve and the internal kernel parameters.

Train:

For each minibatch, compute predictions at all available training coordinates.
Minimize the level, first-difference, amplitude, and primary-path losses in equation 3 with AdamW.
Stop after at most 200 epochs or 12 epochs without validation-MSE improvement.
Select the checkpoint on the protocol's validation grid; score timing after selection.
For future-day forecasting, training and checkpoint selection use days 2–4.
Select any anchor blend separately on validation labels, as stated in the protocol.

A.3 Structural-Ablation Definitions and Provenance

Write $P = \mathbf{h}_{\mathrm{lag}}$ and $R = \lambda \mathbf{r}$ for the two output paths. The full model predicts $P + R$ and uses all terms in equation 3. The no-residual variant predicts P; the no-kernel variant predicts R, also removing the kernel-derived attention bias. Both retain the level, shape, and amplitude losses and omit $\mathcal{L}_{\mathrm{primary}}$.

The earlier results archive also contains a row labelled "w/o residual and kernel". In the recovered forward implementation, disabling both paths yields a zero output, which does not explain that row's nonzero correlations. Its exact forward configuration has not been recovered. Table 8 preserves the reported values without using them to infer component substitution.

Table 8: Archived joint-removal result with unresolved forward configuration. These values are excluded from the structural comparison in Table 5.

Protocol	MSE	Pearson	Shape / Δ-Pearson
Reconstruction	0.00754 ± 0.00004	0.710 ± 0.004	0.309 ± 0.009
Future day	0.00966 ± 0.00011	0.623 ± 0.004	0.244 ± 0.006

A.4 Cross-Gene Coupling

The program-mixing matrices are the only part of the model where one gene's prediction can depend on another gene's curve. We extract the symmetric part of both mixing weights for each of the three seeds of the strict future run and normalize each gene by its total coupling. The normalized matrices agree at Spearman $\rho = 0.70$, 0.63, and 0.70 for the three seed pairs. Coupling is also higher for pairs whose observed curves move together: over all $4{,}560$ pairs, the strongest coupling decile has mean co-response 0.134 versus 0.066 for the remaining pairs (nominal pair-level Mann–Whitney $p = 2.2 \times 10^{-4}$; pairs share genes). Figure 5 shows four such pairs with coherent biological themes.

The coupling does not encode temporal order. The lagged correlation between the two genes' response rates is unrelated to learned coupling (Spearman -0.007, $p = 0.62$), and the strongest decile does not differ from the rest on that measure ($p = 0.92$). We therefore interpret the mixing as capturing genes that respond together, without assigning which gene acts first. As an external check, we compared the coupling with STRING network embeddings for the 88 of 96 genes with a mapping (Szklarczyk et al., 2023). The association is weak and does not survive control for co-expression (partial Spearman 0.013, $p = 0.43$), so the coupling reflects co-variation within this experiment rather than known functional association.

B Additional Quantitative Results

Table 9: Kernel-parameter sweep on GSE213069. Each row fixes one term and retrains the remaining terms with the splits, seeds, and training limits of Table 5. Persistence and dictionary gain are varied over $[0, 1]$; late amplitude is varied over $[-1, 1]$. Reconstruction uses seeds 42, 43, and 44; future-day prediction uses seeds 7, 21, and 84.

Term	Value	Recon. MSE ↓	Shape ↑	Future MSE ↓	Future Δ-r ↑
Persistence p	0.0	0.00851 ± 0.00005	0.292 ± 0.008	0.01460 ± 0.00017	0.357 ± 0.019
	0.2	0.00797 ± 0.00008	0.291 ± 0.008	0.01162 ± 0.00022	**0.364 ± 0.022**
	0.4	0.00772 ± 0.00025	0.277 ± 0.017	0.00971 ± 0.00009	0.336 ± 0.028
	0.6	0.00722 ± 0.00019	0.300 ± 0.017	**0.00920 ± 0.00036**	0.108 ± 0.063
	0.8	0.00704 ± 0.00017	**0.321 ± 0.033**	0.01016 ± 0.00085	-0.323 ± 0.049
	1.0	**0.00700 ± 0.00023**	0.318 ± 0.034	0.01334 ± 0.00216	-0.371 ± 0.032
Late amplitude s	-1.0	0.00875 ± 0.00016	0.278 ± 0.027	0.01297 ± 0.00043	0.130 ± 0.022
	-0.5	0.00781 ± 0.00034	**0.293 ± 0.035**	0.00982 ± 0.00016	0.073 ± 0.029
	0.0	**0.00733 ± 0.00034**	0.250 ± 0.082	**0.00885 ± 0.00038**	**0.251 ± 0.085**
	0.5	**0.00733 ± 0.00015**	0.283 ± 0.019	0.01036 ± 0.00071	0.185 ± 0.058
	1.0	0.00743 ± 0.00017	0.269 ± 0.028	0.01277 ± 0.00121	0.121 ± 0.046
Dictionary gain r	0.0	**0.00721 ± 0.00015**	**0.301 ± 0.019**	**0.00874 ± 0.00021**	**0.305 ± 0.027**
	0.25	0.00757 ± 0.00029	0.205 ± 0.073	0.00931 ± 0.00031	0.275 ± 0.028
	0.5	0.00782 ± 0.00020	0.166 ± 0.039	0.00956 ± 0.00027	0.256 ± 0.057
	0.75	0.00791 ± 0.00027	0.175 ± 0.045	0.00968 ± 0.00016	0.248 ± 0.062
	1.0	0.00810 ± 0.00039	0.153 ± 0.066	0.00989 ± 0.00021	0.245 ± 0.062

B.1 Paul held-out-bin and recovery checks

Table 10 reports the held-out-bin diagnostic under the original DPT axis and five perturbations of the supplied representation. Each row uses the same seven-parameter process and seven-parameter Two-RBF comparator, with bins held out only after the axis is constructed. The 16-bin setting is unavailable because it does not leave the required number of training and test bins. In the known-parameter synthetic check, median curve MSE is 1.96×10^{-20} without noise and 1.34×10^{-4} with Gaussian noise standard deviation 0.02; median absolute errors for the late amplitude and gap parameters are 0.148 and 0.045.

Table 10: Paul held-out-bin checks. Values are held-out MSE; lower is better. The process and Two-RBF columns are fit and scored on the same eligible genes and bins.

Axis	Occupied bins	Process	Two-RBF
Original DPT	11	**0.12186**	0.18270
20-neighbour DPT	21	**0.10563**	0.14330
10-PC DPT	16	**0.11746**	0.16934
30-PC DPT	13	**0.09340**	0.15799
Alternate root cell	11	**0.09509**	0.12715
32-bin discretization	14	**0.11978**	0.17911

Table 11: Condition-level day-5 MSE for the fixed main future-day split (seed 2026).

Condition	scUGPT	Fixed RBF-basis	Difference
ID1	**0.004876**	0.012484	-0.007608
SOX2	**0.015922**	0.031674	-0.015752
TRIM24	**0.002729**	0.007599	-0.004870
TRIM25	**0.003523**	0.008073	-0.004550
ZIC2	**0.018378**	0.023343	-0.004965

B.2 Main-Split Uncertainty and Split Stability

The fixed main split uses condition seed 2026. Its strict day-5 comparison is summarized by condition below. These are mean gene-level MSEs over the three fixed seeds; the exact two-sided sign-flip check over the five condition-level paired differences is $p = 0.0625$.

Split stability. We repeat the strict protocol for condition-split seeds 2027–2030 with training seeds 7, 21, and 84 fixed. These runs assess sensitivity to condition allocation; they are not used to select a new main split. The archived manifests identify seed 2026 as the fixed main split but do not establish that it was chosen before test-result inspection. Averaged over the four splits, scUGPT obtains day-5 MSE 0.01359 ± 0.00519 and derivative Pearson 0.007 ± 0.122, while static broadcast obtains 0.00970 ± 0.00246 and has no derivative residual by construction. The near-zero mean change correlation and higher MSE show that the main-split forecasting advantage does not persist across these additional splits.

Table 12: Condition-split stability under the future-day protocol. Each neural entry is mean $\pm$ standard deviation over training seeds 7, 21, and 84. Static broadcast has zero predicted change. All four additional splits are retained.

Split seed	Raw scUGPT MSE	Δ-Pearson	Static MSE
2027	0.01079 ± 0.00012	0.078 ± 0.043	**0.01074**
2028	0.01089 ± 0.00059	0.022 ± 0.016	**0.00619**
2029	0.02137 ± 0.00011	-0.170 ± 0.028	**0.01000**
2030	**0.01131 ± 0.00048**	0.096 ± 0.030	0.01187

B.3 Endpoint-Anchor Sensitivity

The endpoint-assisted reconstruction protocol supplies the held-out condition's day-5 effect as a static input. To isolate this input, we reran the same split with that input replaced by a zero vector while keeping the target, feature selection, checkpoint rule, and seeds unchanged. scUGPT obtains full-curve MSE 0.01424 ± 0.00014 and Pearson 0.281 ± 0.018, compared with 0.01417 ± 0.00004 and 0.301 ± 0.000 for Fixed RBF-basis; static broadcast has MSE 0.01549 and zero Pearson. This information-ablation diagnostic is separate from the strict future-day forecast.

B.4 Objective-Weight Sensitivity

The primary-path weight in equation 3 is the one objective coefficient that is not fixed by the loss definition, so we swept it under the strict three-way GSE213069 split with identical condition splits across three seeds. Validation MSE is the checkpoint-selection metric, so choosing κ by it uses no test information. The day-5 columns report held-out prediction performance; their values were not used for checkpoint selection. Across the tested weights, day-5 MSE ranges from 0.008840 to 0.008957, a

Table 13: Harvest-only Pearson-of-Δ on the same unseen-single splits as Table 14. This is an endpoint diagnostic rather than the operational follow-up metric.

Method	Adamson	Norman	Replogle
scUGPT	0.790	0.483	0.475
Ridge	0.823	0.413	**0.486**
scGen-style MLP	**0.838**	0.499	0.474
Train mean	0.711	0.498	0.373
GEARS-graph adapter	0.705	**0.506**	0.373

Table 14: Harvest-only top-20 DE overlap on unseen single perturbations under the GEARS simulation split at seed 1. Win rate is the fraction of hold-outs on which a method has the best top-20 overlap. Ties are counted for each tied method, so win rates need not sum to 100%. This is an endpoint diagnostic rather than a trajectory metric.

Method	Adamson	Norman	Norman win (%)	Replogle
scUGPT	0.486	**0.361**	**48.6**	0.253
Ridge	0.525	0.282	32.4	**0.283**
scGen-style MLP	**0.552**	0.335	24.3	0.250
GEARS-graph	0.436	0.336	32.4	0.143
Train mean	0.441	0.334	24.3	0.143
Held-out singles	22	37	37	273

relative range of about 1.3%, while validation MSE ranges from 0.004766 to 0.004821. The largest tested weight has the lowest mean error in both columns. These measurements describe the tested grid; they do not establish equivalence between objective weights.

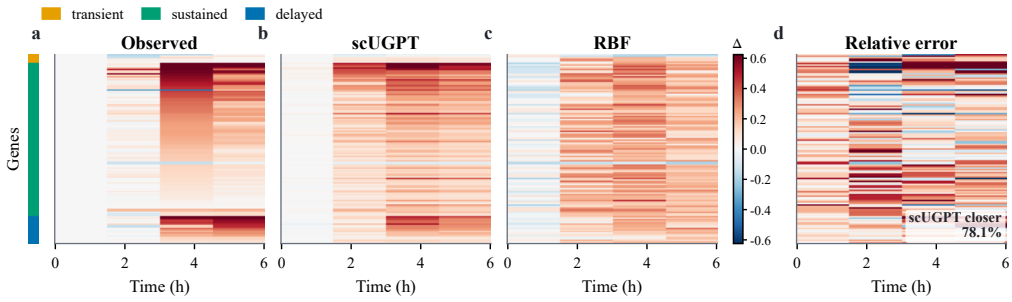
Table 15: GSE213069 objective-weight sweep, three seeds on the strict three-way split. Validation MSE is computed on days 2–4 and is the metric used for checkpoint selection; the day-5 column is held out. Lower MSE and higher Pearson are better.

κ	Validation MSE ↓	Day-5 MSE ↓	Day-5 Pearson ↑
0	0.004821 ± 0.000066	0.008957 ± 0.000092	0.664 ± 0.006
0.05	0.004811 ± 0.000067	0.008951 ± 0.000093	0.664 ± 0.006
0.2	0.004796 ± 0.000073	0.008940 ± 0.000094	0.664 ± 0.005
0.45	0.004778 ± 0.000092	0.008894 ± 0.000093	0.666 ± 0.006
1	$\mathbf{0.004766 \pm 0.000108}$	$\mathbf{0.008840 \pm 0.000083}$	$\mathbf{0.668 \pm 0.006}$

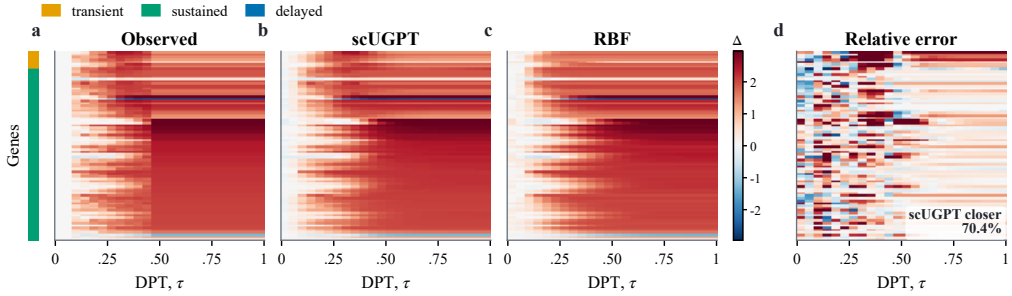
B.5 Architecture Sensitivity

We retrained an attention-causal-mask variant under both GSE213069 protocols with identical splits and convergence budgets. Transformer attention is causal in process order, but the subsequent symmetric lag convolutions remain bidirectional, so the variant is not an end-to-end causal decoder. It is nearly tied on reconstruction (MSE 0.00736 ± 0.00027, level Pearson 0.722 ± 0.011, shape correlation 0.259 ± 0.073) but lower on the future derivative score (0.232 ± 0.082 versus 0.261 ± 0.041 for the bidirectional model). The bidirectional decoder remains the primary model, and this variant is reported as an architecture ablation.

C Additional Figures



(a) Hagai PIC in-fold adaptation, 144 genes.



(b) Paul erythroid DPT, 78 high-amplitude genes; the fitted model is the per-gene analytic process kernel.

Figure 4: Program-level views. Each row shows the observation, the fitted model (labelled scUGPT), Scalar RBF, and the signed error advantage (red: scUGPT closer; blue: Scalar RBF closer). Genes are ordered from the held-out observation (Hagai) or observed DPT trajectory (Paul), not by model error. Hagai uses the endpoint-assisted neural model, with the same fold for checkpoint selection and scoring; Paul uses the per-gene analytic kernel. The fraction of time–gene cells with lower error is 0.781 for Hagai PIC and 0.704 for Paul.

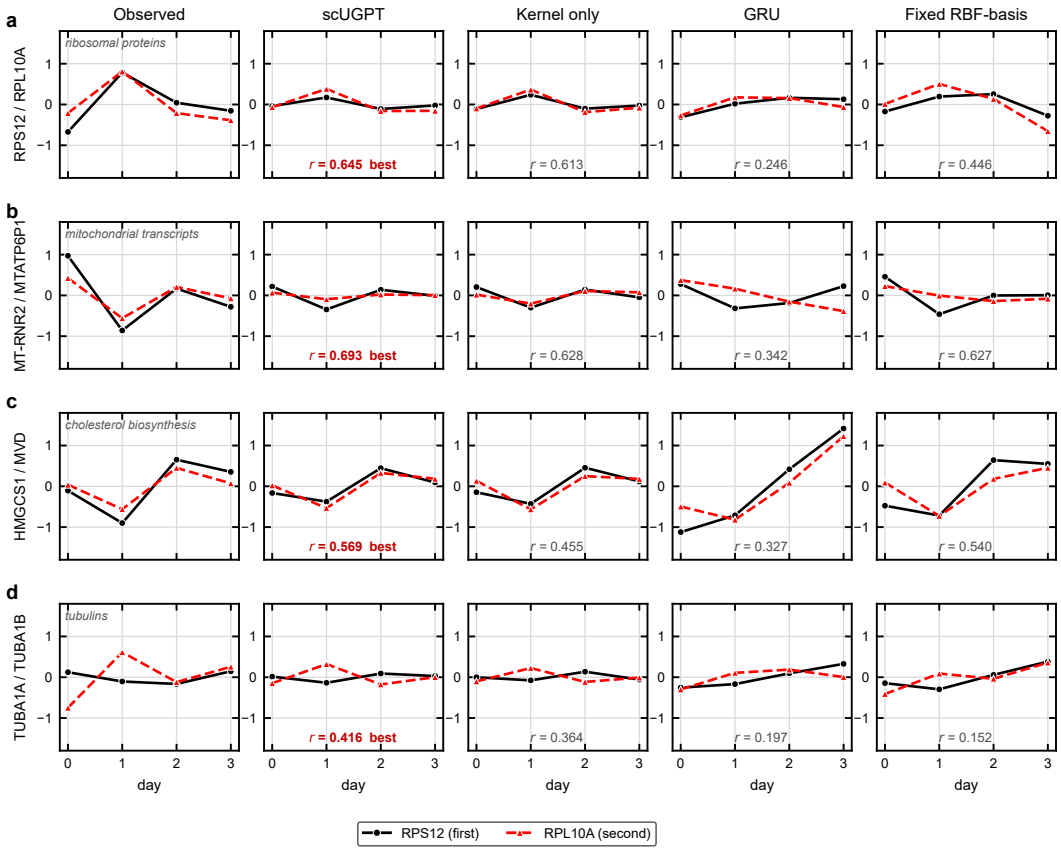


Figure 5: Behaviour of strongly coupled gene pairs under held-out perturbations, for four pairs with a coherent biological theme. Columns are the observation and four methods; rows are the pairs. Each pair is scaled by the spread of its observed curve and centred on its own mean, and curves are averaged over the five held-out conditions, so the panels show whether the two genes move together rather than whether their absolute levels match. The number in each panel is the correlation between the predicted pair curve and the observed pair curve, averaged over those conditions. The panel labelled Kernel only is the no-residual variant, which retains attention and the process kernel. Reported in Section A.4.

D Held-Out Kernel Falsification Diagnostic

We test whether the kernel's gene-wise descriptors remain reproducible and condition-specific on perturbations excluded from training and checkpoint selection. This post-hoc diagnostic uses all five test perturbations and all 96 genes in each of the five GSE future-day splits, with the three saved training checkpoints per split. All 15 checkpoints reproduce the archived raw predictions to maximum absolute error below 2×10^{-6}. We add no fitting or calibration.

Parameter stability. Within each held-out condition, we rank the 96 genes separately for each kernel parameter and compute Spearman correlation for each of the three checkpoint pairs. Table 16 averages these correlations over pairs and conditions. Instability would challenge a reproducible gene-wise interpretation, even when curve predictions are similar. These are descriptive rank correlations, not estimates based on independent biological replicates.

Centre ranks are highly stable, but the implementation initializes their shared component from $d^{(0)} = \mathrm{sigmoid}(\mathrm{linspace}(-1, 1, 96))$ in gene-index order. The learned centres correlate with this template at 0.962–0.970 across splits. Cross-seed stability of the deviations is only 0.199–0.309. This subtraction controls for the shared centre template, not the full randomly initialized conditional modulator. Width is moderately stable; persistence and the late-component parameters are less stable. Thus the high raw centre correlation alone does not establish learned kinetic ordering.

Table 16: Cross-seed gene-rank stability on held-out perturbations. Parameters are centre d, width w, persistence p, late gap q, signed late amplitude s, and dictionary gain r. The last column ranks $d - d^{(0)}$, the deviation from the shared centre template.

Split seed	d	w	p	q	s	r	$d - d^{(0)}$
2026	0.949	0.520	0.227	0.094	0.236	0.307	0.309
2027	0.952	0.303	0.203	0.036	0.128	0.120	0.245
2028	0.947	0.346	0.176	0.023	0.149	0.090	0.221
2029	0.952	0.321	0.176	0.071	0.088	0.080	0.222
2030	0.955	0.360	0.171	0.060	0.116	0.084	0.199

Condition exchange. We exchange the complete kernel output among the five held-out conditions, preserving gene order, anchors, all other inputs, and all weights. This intervenes on the actual kernel tensor used by the forward pass. We enumerate all $5! = 120$ assignments, including the original assignment. If condition-specific kernels carry essential predictive information, assigning them to other conditions should increase forecast error.

For each condition we average day-5 MSE across checkpoints, then compare the mean over the 119 nonidentity assignments with the original assignment. Table 17 reports the mean difference and a percentile 95% interval from 10,000 resamples of the five condition-level differences. These intervals describe variation among the five evaluated conditions. The final column is the fraction of all 120 assignments with mean MSE no larger than the original, including the original assignment; it is an unadjusted, descriptive lower-tail permutation probability under exchangeable kernel assignments.

Table 17: Kernel-exchange diagnostic on held-out perturbations. Positive ΔMSE means that exchanging kernels increases raw forecast error. MSE differences and intervals are in units of 10^{-3}. No checkpoint is retrained.

Split seed	ΔMSE $\times 10^3$	95% condition interval	Permutation p
2026	0.161	$[-0.332, 0.896]$	0.283
2027	0.875	$[-0.096, 2.647]$	0.108
2028	0.162	$[-0.176, 0.674]$	0.558
2029	0.096	$[-0.122, 0.409]$	0.367
2030	0.598	$[-0.145, 1.918]$	0.142

Every interval includes zero, and none of the original assignments is unusually accurate at the nominal 0.05 level. Together with uneven parameter stability, this limits the evidence for condition-specific kinetic interpretation on these held-out perturbations. The kernel remains an explicit curve parameterization; the current diagnostic does not establish uniquely recoverable biological timing parameters. With only five test conditions per split and a residual path that can compensate for the kernel, small exchange effects are not evidence of equivalence or of an absent biological signal.