CELLSHIFT: BENCHMARKING SINGLE-CELL TEMPORAL MODELS ON UNSEEN BIOLOGY

Anonymous authors
Paper under double-blind review

ABSTRACT

Single-cell temporal modeling is the task of predicting the gene expression of individual cells at unobserved time points, either after the last measurement (forecasting) or between measurements (interpolation). Machine-learning models such as PRESCIENT, scNODE and CellTempo learn these tasks from single-cell RNA sequencing time courses, which measure thousands of genes in each of thousands of cells. Existing evaluations, however, hold out time points within the cell types and conditions seen during training, so it remains unclear whether these models learn dynamics that transfer to unseen biology. This paper introduces CellShift, a benchmark spanning seven time-course datasets, with three extended to test generalization across held-out cell fates, lineages, and starting populations. It evaluates statistical baselines, neural dynamics models, pretrained temporal models, and language-model agents under a shared evaluator with explicit tracking of information access. On lineage-traced blood development (the LARRY *in vitro* dataset), PRESCIENT and a population-time MLP outperform simple copy baselines when forecasting the whole day-6 population. However, when evaluated on held-out fates, no fitted model without access to target-day cells improves upon copying the latest observed snapshot by more than 5%, even when initialized from the held-out cells. Even on identical target cells, holding the fate out raises the error of nearly every fitted model. At the clone level, models trained on day-6 states assign at most 28% of predicted cells to a held-out fate that comprises 93 to 95% of the observed population, highlighting the difficulty of generalizing to unseen cell fates. Agents that choose among a fixed set of numerical tools perform on par with a selector that uses no language model.

1 INTRODUCTION

The *state* of a cell is characterized by the expression levels of its roughly 20,000 protein-coding genes, which single-cell RNA sequencing measures in thousands of individual cells at a time. Time-course experiments repeat this measurement at successive time points to follow how cell populations change during development, disease, and treatment. Models that predict these changes over time are a central component of the proposed *virtual cell*, a computational model that simulates cellular behavior (Bunne et al., 2024; Roohani et al., 2025), which could predict how a cell population will develop, or how a drug will alter it, before the experiment is performed (Van de Sande et al., 2023). Such uses require generalization to cells whose type was absent from training, the property this paper evaluates.

Most current virtual-cell models predict a *static endpoint*. Given cells before a treatment, they predict a single snapshot of the cells after the treatment (Lotfollahi et al., 2023; Roohani et al., 2024; Adduri et al., 2026), and on these benchmarks, simple linear baselines match deep models (Ahlmann-Eltze et al., 2025; Heidari et al., 2026). A single before/after pair also carries little information about the intervening process: macrophages exposed to different immune stimuli, for example, are distinguished more reliably by their expression time courses than by any single time point (Sheu et al., 2024). Time-course data, in which the same system is measured at several time points, allow the dynamics themselves to be learned.

Sequencing destroys the cell it measures, so a time course is a sequence of *snapshots* of different cells from the same population. Lineage tracing reconnects them: a heritable DNA barcode marks

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

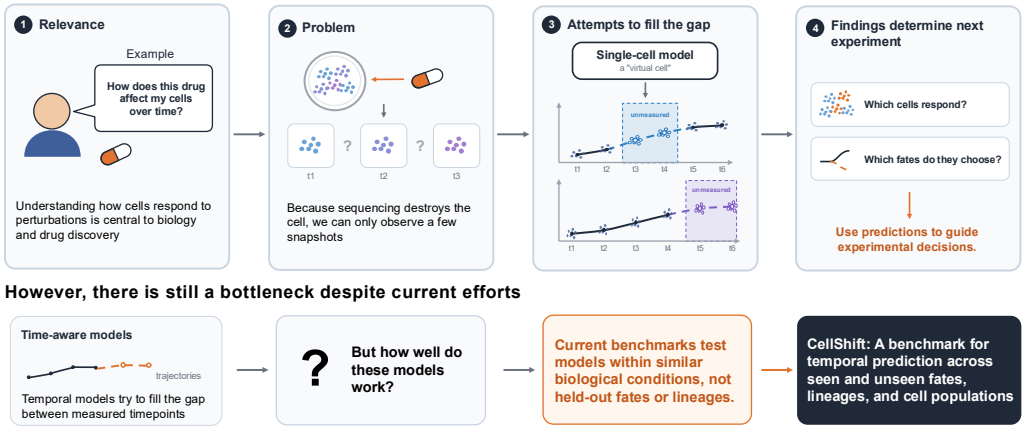

Figure 1: ***Problem***. Characterizing how cells change over time is fundamental to biology and drug discovery, but destructive sequencing and costly experiments limit what can be observed. Temporal single-cell models can predict unmeasured time points, yet existing benchmarks evaluate them within the biological settings seen in training, so their ability to generalize to unseen biology is unclear. CellShift addresses this gap by testing models on held-out cell fates, lineages and starting populations.

all descendants of one ancestor cell, a *clone* (Weinreb et al., 2020; Biddy et al., 2018). Temporal models are trained on such sequences of cell states, or *trajectories*, and predict the distribution of cell states at a later time (*forecasting*) or at a missing intermediate time (*interpolation*) (Yeo et al., 2021; Zhang et al., 2024; Tong et al., 2024; Luo et al., 2026); this paper refers to both tasks as *temporal modeling*.

If a model captures gene-regulatory dynamics that are shared across cell types, it should also predict cells absent from its training data, such as cells of an unseen fate or lineage. This capability is rarely tested by existing temporal benchmarks, in which time points are held out while the same cell types and conditions appear in both training and test data (Osakwe et al., 2026; Zhou et al., 2026). Under this design, a model can score well by recalling the trajectories of familiar biology, but a high score does not demonstrate that transferable dynamics have been learned.

Language models, together with agents that use them to select and combine analysis tools, have also been proposed for single-cell analysis and prediction (Levine et al., 2024; Gao et al., 2025; Yuan et al., 2026). Their prior knowledge of gene function could, in principle, guide the choice of a forecasting method when the data alone are ambiguous, but such agents have not been compared with dedicated models on temporal tasks.

This paper introduces **CellShift**, a benchmark that assesses whether temporal single-cell models generalize to biology absent from their training data (Figure 1). Whereas earlier temporal single-cell benchmarks hold out time points (Osakwe et al., 2026; Zhou et al., 2026), CellShift also holds out cell fates, a lineage and starting populations. The contributions are twofold.

1. **A temporal benchmark for generalization to unseen biology.** Seven time-course datasets cover blood development, reprogramming and embryogenesis; beyond held-out time points, three carry tasks that hold out cell fates, a lineage or a starting population. One evaluator scores every method, and every result is tagged with its information regime (Section 3.2).
2. **An evaluation of language-model agents.** To the best of the authors' knowledge, this is the first evaluation of agents on single-cell temporal modeling. The agents choose among a fixed library of numerical tools and are compared with a *zero-call selector* that searches the

same library on validation data without a language model, which isolates the contribution of the language model itself.

Four findings emerge, numbered as in Section 5: (1) fitted models (methods with parameters fit to training data) outperform copy baselines (methods that copy or average observed cells) on the whole day-6 population of LARRY *in vitro* but on held-out fates none beats copying the held-out cells' latest snapshot by more than 5%, even when started from those cells; (2) simple copy baselines are strong across the benchmark; (3) a low average error can conceal an unrealistic predicted population; and (4) agents perform on par with the zero-call selector. Current models have not yet learned dynamics that transfer to unseen fates.

2 Related Work

Models of single-cell dynamics learn how a cell population moves from one time point to the next. Optimal-transport methods infer transitions between consecutive samples (Schiebinger et al., 2019; Klein et al., 2025), RNA velocity and fate mapping infer directions of change from one snapshot (La Manno et al., 2018; Lange et al., 2022), and neural dynamics models such as TrajectoryNet (Tong et al., 2020), PRESCIENT (Yeo et al., 2021), MIOFlow (Huguet et al., 2022), scNODE (Zhang et al., 2024), flow matching with minibatch optimal transport (OT-CFM; Tong et al., 2024) and scIMF (Jiang et al., 2025) learn a continuous-time model from which later populations can be generated. Sequence models such as CellTempo (Luo et al., 2026) and MaxToki (Gómez Ortega et al., 2026) continue a series of cell states autoregressively. These methods are typically evaluated by holding out intermediate or final time points of a time course whose cell types and conditions also appear in training. Benchmarks of temporal single-cell prediction follow the same design: scTimeBench (Osakwe et al., 2026) and an evaluation of zero-shot foundation-model embeddings (Zhou et al., 2026) hold out time points within a single biological setting, and both find that simple baselines are difficult to beat (Appendix Table 3). On the LARRY data used here, the original study and CoSpar predict the eventual fate of a clone from its early cells (Weinreb et al., 2020; Wang et al., 2022), and Chreode reports zero-shot clonal fate scores (Qiu et al., 2026); these tasks classify fates seen in training.

Generalization to unseen biology has primarily been studied in static perturbation-response models, which predict responses to gene knockouts or drugs at a single time point rather than across time. CPA (Lotfollahi et al., 2023), GEARS (Roohani et al., 2024), State (Adduri et al., 2026), and single-cell foundation models, such as scGPT, (Cui et al., 2024) are evaluated on held-out perturbations. PerturBench additionally test on held-out perturbations and cell contexts (Wu et al., 2024). The Virtual Cell Challenge evaluates perturbation effects in a held-out cell type (Roohani et al., 2025; Arc Virtual Cell Initiative Team & Goodarzi, 2026). In this setting, deliberately simple linear baselines match or outperform deep learning models (Ahlmann-Eltze et al., 2025), and the choice of metric can hide failures such as disrupted dependencies between genes (Heidari et al., 2026). CellShift brings the held-out-biology design of these benchmarks to temporal prediction.

Language models have also been applied to single-cell data, both as foundation models that represent a cell as a sentence of gene names (Levine et al., 2024) and as agents that reason over data while calling analysis tools (Gao et al., 2025). Applied to perturbation response, such reasoning produces plausible explanations yet often underperforms a simple baseline (Yuan et al., 2026). In general time-series forecasting, language models can forecast zero-shot (Gruver et al., 2023), although ablations indicate that the language model often contributes little (Tan et al., 2024). To the best of the authors' knowledge, language-model agents have not previously been compared with dynamics models on single-cell temporal modeling under biological shift.

3 Benchmark

3.1 Data and tasks

CellShift draws on seven temporal datasets (Table 1, Figure 2; details in Appendix C). Splits are fixed before any reported model was scored, and lineage-traced data are split by clone, so that all cells of a clone fall on the same side. Each task withholds one of the following:

Table 1: Temporal datasets in CellShift, grouped by the tasks scored on each. I denotes interpolation of a held-out time and F a forecast of a later time. Lineage-traced datasets are split by clone. Appendix Table 18 gives the cell and clone counts of every task. For Farrell, the ortholog mapping keeps 9,061 of the 18,791 genes, which scales its RMSE values but not comparisons within the dataset.

Dataset	Biological system and sampled times	Held-out times (in-domain)	Held-out biology (shift)
LARRY *in vitro* Weinreb et al., 2020	Mouse blood progenitors differentiating in culture, with clonal barcodes; days 2, 4, 6	I (day 4); F (day 6)	Cross-fate: neutrophil, monocyte, basophil; starting population: LK, LSK
LARRY cytokine Weinreb et al., 2020	The same system under five cytokine conditions; days 2, 4, 6	I (day 4); F (day 6), overall and within the full cocktail	Cross-condition: full cytokine cocktail (for completeness)
CellTag Biddy et al., 2018	Mouse fibroblasts reprogrammed toward endoderm progenitors, with clonal barcodes; eight time points from day 0 to 28	I (day 9); F (days 21, 28)	Cross-fate: dead-end and successful reprogramming (I and F); replicate transfer
Schiebinger Schiebinger et al., 2019	Mouse fibroblasts reprogrammed to induced pluripotent stem cells; every half day from day 0 to 18	I (days 5, 10, 15); F (days 16, 17, 18)	None
Pijuan-Sala Pijuan-Sala et al., 2019	Mouse gastrulation atlas with annotated lineages; nine stages, E6.5 to E8.5	I (E7.0, E7.5, E8.0); F (E8.5 from E6.5–E7.5)	Cross-lineage: mesoderm, with a matched in-domain control
Farrell Farrell et al., 2018	Zebrafish embryogenesis; 12 stages up to 12 hpf	I (4.7, 6.0, 8.0 hpf); F (10, 11, 12 hpf)	None
Moon Moon et al., 2019	Human embryoid bodies differentiating from stem cells; five windows, days 0 to 27	I (three interior windows)	None

- **A time point (in-domain).** Cells at a held-out time are predicted within a biological setting seen in training.
- **A cell fate (cross-fate).** All clones that give rise to one fate, such as LARRY basophils or CellTag dead-end cells, are removed from training.
- **A lineage (cross-lineage).** One developmental lineage, Pijuan-Sala mesoderm, is removed from training.
- **A treatment condition (cross-condition).** One cytokine condition of LARRY is removed from training; since no evaluated model reads a condition label, this task is reported for completeness only.
- **A starting population (cross-population).** LARRY clones that start from one progenitor population (LK or LSK) are removed from training.
- **A replicate (cross-replicate).** One of the two CellTag replicates is predicted from the other.

Two static perturbation-response datasets, Norman (Norman et al., 2019) and Dixit (Dixit et al., 2016), are evaluated in a separate track (Appendix A, Table 22), because a single endpoint response is not a temporal forecast.

3.2 Task definition and information regimes

Let p_t^g denote the distribution of expression vectors $x \in \mathbb{R}^G$ ($G = 18{,}791$ genes) of the cells of group g at time t, where a group is a set of clones, a fate, a lineage, a condition or a whole population. A task specifies a target group g^*, a target time t^*, a set of context times $\mathcal{S}$ at which cells of g^* are observed, and a training pool $\mathcal{P}$ of cells from other groups or times. Forecasting uses $\mathcal{S} \subset \{t < t^*\}$, and interpolation also admits times after t^*; the method outputs a set of cells or a mean whose

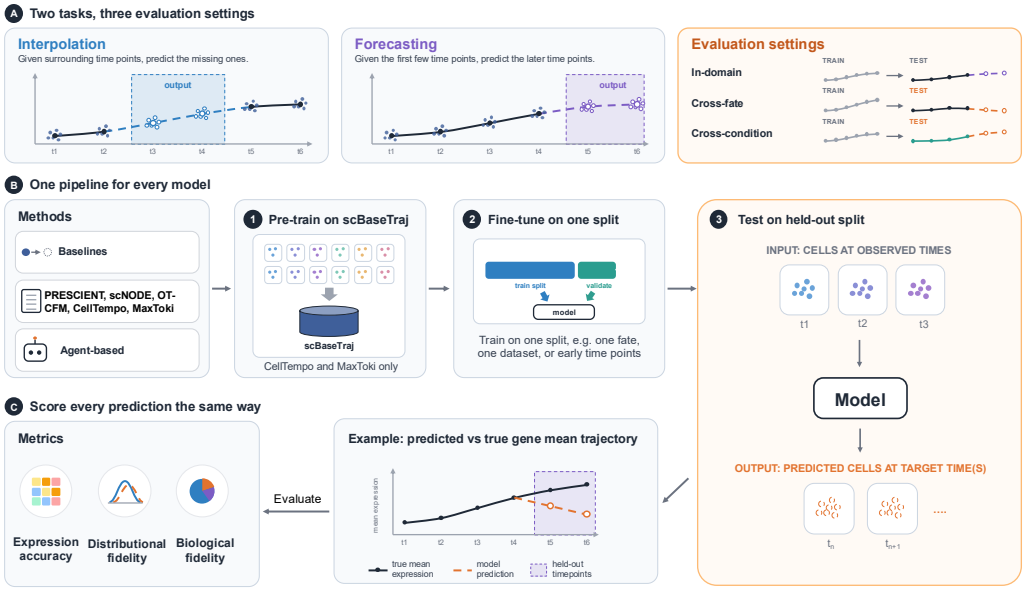

Figure 2: ***Benchmark and evaluation framework***. (a) Tasks are evaluated **in-domain** and under held-out biology, e.g. **cross-fate**; lineage and starting-population shifts are also held out, and cross-condition is reported for completeness only. (b) Each method uses only its permitted observations; only CellTempo and MaxToki are pretrained. (c) One evaluator scores every prediction, and each result carries its information regime.

distribution $\hat{p}$ approximates $p^{g^*}_{t^*}$. *Copy baselines* copy or average observed cells, and *fitted models* are all methods with parameters fit to training data. A method's *regime* is *pre-target* when $\mathcal{P}$ holds no cell at t^* and *target-day* when $\mathcal{P}$ holds other groups' cells at t^*; target-day results are marked ^{t} and never counted as wins over pre-target methods. A method is *group-conditioned* when its prediction depends on the observed context of g^*. Copy baselines that read the test context and OT-CFM on LARRY are group-conditioned; PRESCIENT, scNODE and the regression baselines are not in their reference configuration, and group-conditioned variants of PRESCIENT, scNODE and OT-CFM are evaluated separately (Section 4). A score counts as evidence of generalization only when the compared methods predict the same target in the same regime. Because the held-out fate cohorts are defined by the clones' day-6 fate, their early cells are already biased toward that fate, which favors group-conditioned copies. Matched in-domain controls on identical target cells exist for the held-out mesoderm of Pijuan-Sala and, on a half split of their clones, for the three LARRY fates.

3.3 Evaluation

With $p = p^{g^*}_{t^*}$, the primary expression metric is $\mathrm{RMSE} = \|\mathbb{E}_{\hat{p}}[x] - \mathbb{E}_{p}[x]\|_2/\sqrt{G}$, the error of the pseudobulk mean, with the Pearson correlation of the two means alongside. Distributional quality is measured by the marginal $W_1 = \frac{1}{50}\sum_{k=1}^{50} W_1(\hat{p}^{(k)}, p^{(k)})$, where $p^{(k)}$ is the distribution of the k-th coordinate of a training-fitted PCA; coordinates are not whitened, so leading components weigh more, and joint structure across components is ignored. A mean prediction is a point mass, so RMSE favors mean predictors and marginal W_1 favors methods that output cells. Fate proportions come from a 15-nearest-neighbor classifier fitted in the same PCA space on annotated cells at the target time that belong to no split of the task. Preprocessing and PCA are fitted on training data only, and MaxToki, which outputs gene ranks without expression values, is scored on a separate rank track.

Table 2: Method families in CellShift. Training data is what a method may use. Population-level methods are pre-target, except OT-CFM on the cross-fate and starting-population tasks: its implementation supports only held-out times, so there it trains on the other groups' full trajectories and is marked ^{t}. Group-conditioned variants of PRESCIENT, scNODE and OT-CFM start from the held-out group's own cells and are pre-target (Appendix Table 16).

Family	Methods	Training data	Output
Copy baselines	Persistence, nearest prior, test-context copy, neighbor interpolation, history and source means	None (copies observed cells)	Cells or mean
Regression baselines	Linear time, PCA-ridge, population-time MLP	Training means over permitted times	Mean
Dynamics models	PRESCIENT (Yeo et al., 2021), scNODE (Zhang et al., 2024), OT-CFM (Tong et al., 2024)	Training cells over permitted times	Cells
Pretrained models	CellTempo (Luo et al., 2026), MaxToki (Gómez Ortega et al., 2026)	Pretraining corpus, optional fine-tuning	Cells; gene ranks
Agents	Zero-call selector, single agents, recursive agents, specialists	Tools applied to the observed context	Mean

4 MODELS

Table 2 summarizes the five method families. Appendix Table 4 lists every method with its output, starting population and regime, Appendix D gives the settings of each, and Appendix G lists the models considered but not run. Models are selected on training and validation data before the shared evaluator scores the hidden target, and a task is marked not applicable when it lacks an input that a method needs, such as clone histories or a condition descriptor.

Baselines. Copy baselines do not model dynamics. They either copy the latest observed cells, the nearest earlier timepoint, or the observed cells from the held-out group (*test-context copy*), or average the two observed timepoints surrounding an interpolation target. Regression baselines fit either a per-gene trend or a regularized map from time to mean expression. The population-time MLP selects its size and weight decay by leave-one-time-out cross-validation. None of the regression baselines reads the test context, so on a held-out group they predict the training mean at the target time.

Clone-level track. On LARRY, a separate track predicts each test clone's day-6 cells from its own early cells with models trained on early and day-6 cells of training clones, including nearest neighbors in frozen scGPT embeddings (Cui et al., 2024). These models are group-conditioned but target-day, so the track is reported separately (Appendix Table 7).

Dynamics models. PRESCIENT simulates stochastic dynamics under a learned potential, scNODE evolves encoded cells with a neural ODE, and OT-CFM learns a velocity field from optimal-transport pairings. PRESCIENT and scNODE simulate from all observed cells at the task's origin time (on LARRY, day 2 pooled over every clone), so they do not know which group they predict; group-conditioned variants start them instead from the held-out group's own day-2 or day-4 cells, and a pre-target OT-CFM variant trains without day-6 cells. OT-CFM starts from the latest context time, on LARRY from the held-out group's own day-4 cells. PRESCIENT and OT-CFM select checkpoints and step caps on validation data, or on a held-out training time where a task has no validation split.

Pretrained models. CellTempo generates later VQ-VAE tokens autoregressively (released and locally trained checkpoints; no interpolation), and MaxToki generates ranked gene tokens, scored with rank metrics against permutation nulls; pretraining overlap with the target studies has not been ruled out.

Agents. The agents receive population-mean trajectories with gene symbols but no other biological annotation, and submit a plan that combines, with parameters and weights, twelve numerical

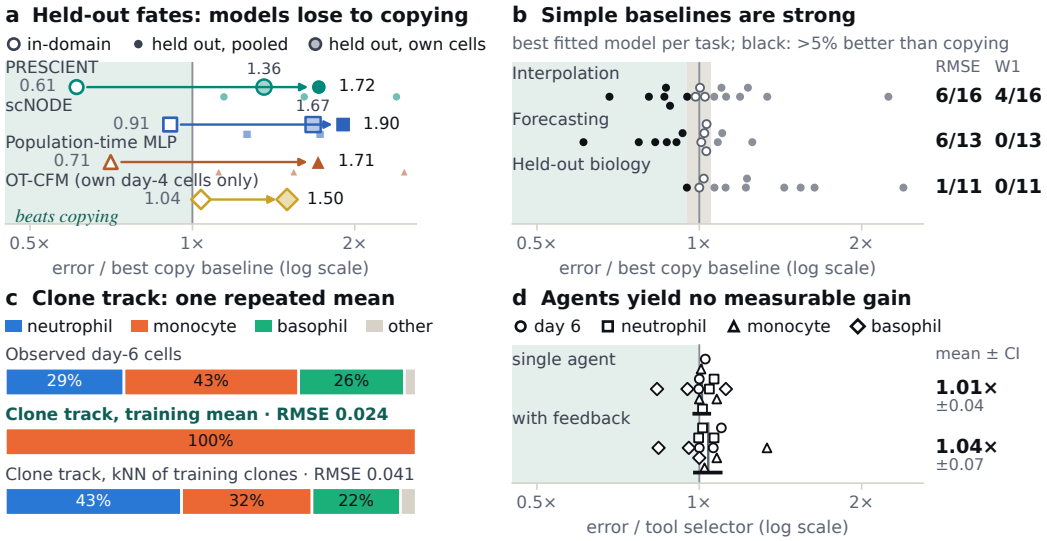

Figure 3: The four findings (pre-target unless labeled). (a) Error relative to the best copy baseline on LARRY: open, the in-domain forecast of the whole day-6 population; filled, each held-out fate (faint) and their mean, relative to the test-context copy on that fate's own targets; filled, pooled runs; tinted, runs started from the held-out clones' own day-4 cells (OT-CFM trained without day-6 cells, and only in this form). (b) Best fitted model per task among the runs of Appendix Table 5, excluding ^{t} results and the group-conditioned variants; open dots are ties within 5%. (c) Cell types of the day-6 prediction in the target-day clone track; the training mean repeats one vector. (d) Agent runs relative to the zero-call selector; marker shape gives the task, and the bar a 95% interval of the mean.

tools (persistence, recent means, damped linear trends, offsets from a reference population, ridge and low-rank corrections, and blends), which the host executes. The zero-call selector grid-searches the same tools and parameters on validation episodes built from training data and serves as the reference. Single agents run with or without aggregate validation feedback. Recursive agents split the task into sub-calls (Zhang et al., 2025), parallel specialists predict gene modules or size-matched random gene groups, and a recall control reruns an agent on original, anonymized and noise inputs. All calls use DeepSeek V4.1 Flash through OpenRouter in the OpenCode harness. Appendix E describes the harness and reproduces every system prompt.

5 Results

Scores below are pseudobulk RMSE on the hidden test cells unless another metric is named. Appendix Table 5 gives the complete matrix, Appendix Tables 11 and 12 give it under the Wasserstein distance and the Pearson correlation, and Figure 3 summarizes the four findings. On LARRY *in vitro*, values are means and standard deviations over three seeds. Elsewhere they are single runs. A fitted model counts as beating a copy baseline only by a margin of more than 5%, the single decision rule used throughout.

5.1 On held-out fates, fitted models without group context lose to copying

Finding 1. Fitted models beat the best copy baseline when forecasting the whole LARRY day-6 population. On held-out fates, no pre-target fitted model beats copying the held-out clones' day-4 cells: pooled models are 1.1 to 2.7 times worse, and models started from those cells tie on neutrophils and are 1.1 to 1.9 times worse otherwise.

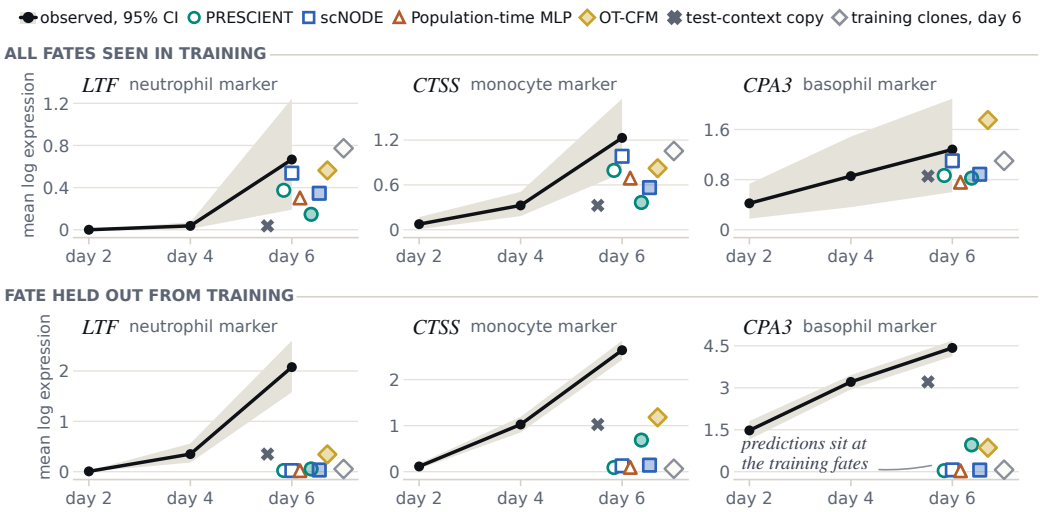

Figure 4: Fate markers on LARRY. Black: observed mean log expression of the test clones, with a 95% bootstrap interval over clones (61 clones on top; 160, 159 and 88 below). Open markers: predicted day-6 means over three seeds of the pooled runs; tinted markers: the same models started from the test clones' own day-4 cells (OT-CFM trained without day-6 cells). Cross: the test-context copy, which predicts the day-4 mean. Diamond: training clones at day 6. Bottom row: the column's fate is held out. Vertical scales differ. In the top row the copy misses widely; under shift the fitted models fall further.

Whole population. On the LARRY day-6 forecast of all test clones, PRESCIENT (0.0519 ± 0.0116, seeds 0.0415 to 0.0643) and the MLP (0.0600 ± 0.0016) beat the best copy baseline (0.0850), and scNODE (0.0773 ± 0.0141) is 9% below it.

Held-out fates. The test-context copy scores 0.1119, 0.0918 and 0.0657 on held-out neutrophils, monocytes and basophils, and pooled PRESCIENT, scNODE and the MLP are 1.1 to 2.7 times worse (on different target cells than the in-domain task). On a matched half split of the held-out clones, holding the fate out raises the error of every fitted model by 7 to 88%, except PRESCIENT started from day-4 cells on neutrophils (-1%), while the test-context copies are unchanged by construction (Appendix Table 15). Held-out clones are selected by their day-6 fate, so their day-4 cells are partly committed, which favors the copy (Section 3.2). OT-CFM^t, trained on the other fates' day-6 cells, beats the copy on neutrophils only. Started from the clones' own day-4 cells, PRESCIENT, scNODE and OT-CFM (trained without day-6 cells) score 0.101 to 0.160 (Appendix Table 16): they beat a copy of the day-2 cells (0.19 to 0.21) but not the day-4 copy by more than 5%; on neutrophils, the closest case, bias-corrected bootstrap intervals make every such comparison a tie (Appendix F.5). In held-out basophil clones (Figure 4), *CPA3* rises from 3.2 at day 4 to 4.4, yet the pooled models predict 0.04 to 0.06, near the training clones' 0.07, and models started from the clones' own day-4 cells lower it to 0.06 to 0.96: they move held-out cells toward the training fates.

Starting population and lineage. The starting-population shifts are the mildest in the benchmark: LK and LSK target cells lie 0.079 (RMSE) from the day-6 training cells, against 0.135 to 0.150 for the held-out fates (Appendix Table 17). On LSK, the pooled MLP (0.0712) is 5% below the best copy baseline (0.0751), and PRESCIENT started from the LSK clones' own day-2 cells reaches 0.0615, the one case of transfer in the benchmark; on LK, the best conditioned model (0.1119) trails the copy (0.0941). On Pijuan-Sala, removing mesoderm raises the regression baselines' error by 29 to 38% on the same 1,162 target cells, while the test-context copy is unchanged and beats every fitted model in both tasks (Appendix Table 14).

5.2 Simple baselines are strong

Finding 2. A parameter-free average of the two neighboring snapshots is the best interpolator on 9 of 16 tasks, by more than 5% on 7, and no fitted model beats copying the latest population on the three Schiebinger forecasts (Figure 3b).

Neighbor interpolation wins on all Moon and Schiebinger times and CellTag day 9 (Appendix Table 6). Among forecasts, the MLP beats persistence on all three Farrell stages by 7 to 17%, PRESCIENT wins on Pijuan-Sala E8.5 (0.0588 against 0.0648), the source-population mean on CellTag, and persistence on Schiebinger, where PRESCIENT ties it within 3%.

5.3 Low error can conceal an unrealistic population

Finding 3. In the target-day clone-level track, the lowest-error prediction of the LARRY day-6 cells repeats one mean expression vector. Under marginal W_1, fitted models beat copying on only 4 of 40 tasks (Figure 3b,c).

The clone-level training mean scores 0.0243, but its one repeated vector is a monocyte, while the observed cells are 29% neutrophils, 43% monocytes and 26% basophils. Under W_1, which penalizes point masses, fitted models win on 4 of 16 interpolation tasks and none of the 24 others (4 of 35 among cell-output methods).

Pretrained models. The released CellTempo checkpoint's error sits near its tokenizer's reconstruction error, and the locally trained checkpoint collapses to three output vectors; neither enters a count or figure (Appendix Tables 9 and 10). MaxToki's ranked genes pass 0 of the released checkpoint's change-based tests and 5 of 106 (4.7%) for a checkpoint pretrained from scratch, the rate expected by chance (Appendix Table 24).

5.4 Agents yield no measurable gain

Finding 4. Across four LARRY tasks, agents that choose among a fixed set of numerical tools perform on par with the zero-call selector (Figure 3d).

Validation feedback doubles the calls without reducing the error (Appendix Table 8). Noise inputs raise the error from 0.1051 to 0.1415 and anonymized inputs leave it unchanged (0.1045), so the agents use the data; repeat standard deviations reach 15%, so smaller differences are not interpreted.

6 Discussion and limitations

Current temporal models do not transfer to unseen fates. Even started from the held-out clones' own cells, the dynamics models move them toward the training fates and do not beat copying those cells by more than 5% on any held-out fate (on neutrophils, a tie under bootstrap intervals). On identical target cells, holding a fate out raises the error of nearly every fitted model by 7 to 88%, while the copy is unaffected; the one case of transfer (PRESCIENT on LSK) occurs on a mild shift. A model should therefore start from the context of the group it predicts, and the copy of that context, not a pooled baseline, is the bar to beat.

Agents add no measurable value over tool selection. Agents that choose among the zero-call selector's tools match it on four LARRY tasks, so a language model's prior knowledge adds nothing measurable when the numerical tools are fixed.

Limitations. Matched controls cover the LARRY fates and one lineage, no task with known dynamics shows that a held-out fate is reachable, and the CellTag fate tasks cannot be split by lineage (Appendix C). Pre-target LARRY models extrapolate from day 4 to day 6, test sets are small (61 clones in-domain), results outside LARRY *in vitro* are single runs, and PRESCIENT seeds vary by up to 70%. The agents, which match a selector that already searches their tools, use one language model and a fixed tool library.

AI USE STATEMENT

Generative AI tools were used to help write code, plan and analyze experiments, interpret results, make figures, search the literature, and draft and edit the manuscript. Separately, an LLM agent is one of the methods evaluated in this paper. Its model, tools, prompts and token use are reported in Section 4 and Appendix E. Scripts rebuild every results table and results figure from the recorded evaluator scores, and every reference was checked against its source page. The authors reviewed all AI-assisted work and take full responsibility for the text, claims, code and results of this paper.

ETHICS STATEMENT

This work collects no new experimental data. It uses published, processed expression matrices from LARRY and its cytokine arm, CellTag, Schiebinger, Pijuan-Sala, Moon embryoid bodies and Farrell. Local CellTempo and MaxToki pretraining runs also use the published scBaseTraj collection. The static perturbation screens of Norman and Dixit, used in the appendix, come from published experiments in a human cell line and mouse dendritic cells. The lineage-traced and developmental data behind the main results come from model organisms. The embryoid bodies of Moon derive from human pluripotent stem cells, for which consent and approval were obtained by the original study. No attempt is made to recover donor identities, and scBaseTraj is not redistributed. Because scBaseTraj aggregates published studies, a target study may overlap with pretraining data, so transfer results are labeled with this possible overlap and are not described as zero-shot. No participants or annotators were recruited for this work.

REPRODUCIBILITY STATEMENT

Section 3 and Appendix C describe every dataset, preprocessing step and split rule. Appendix Table 4 lists every method, Appendix D gives the model settings and seeds, and Appendix E gives the agent model, tools, prompts and budgets. The task manifests, model settings, seeds, and evaluation code are available in an anonymized code repository at `https://anonymous.4open.science/r/cellshift-review-13D5`. The processed benchmark datasets, evaluator scores, and prediction files with checksums are hosted at `https://anonymous-hf.com/a/ztx9z312wue5/`. One evaluator with fixed PCA and sampling seeds scores every table cell, and scripts rebuild every results table and figure from the recorded evaluator scores.

REFERENCES

Abhinav K. Adduri, Dhruv Gautam, Beatrice Bevilacqua, Mohsen Naghipourfar, Alishba Imran, Rohan Shah, Noam Teyssier, Rishi Verma, Christopher Carpenter, Basak Eraslan, et al. Predicting cellular responses to perturbation across diverse contexts with State. *Cell*, 189(19):5914–5931.e20, 2026. doi: 10.1016/j.cell.2026.07.052.

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.

Arc Virtual Cell Initiative Team and Hani Goodarzi. Virtual Cell Challenge 2026: Benchmarking zero-shot generalization across cellular contexts. *Cell*, 189(19):5827–5830, 2026. doi: 10.1016/j.cell.2026.08.004.

Brent A. Biddy, Wenjun Kong, Kenji Kamimoto, Chuner Guo, Sarah E. Waye, Tao Sun, and Samantha A. Morris. Single-cell mapping of lineage and identity in direct reprogramming. *Nature*, 564 (7735):219–224, 2018. doi: 10.1038/s41586-018-0744-4.

Charlotte Bunne, Yusuf Roohani, Yanay Rosen, Ankit Gupta, Xikun Zhang, Marcel Roed, Theo Alexandrov, Mohammed AlQuraishi, Patricia Brennan, Daniel B. Burkhardt, et al. How to build the virtual cell with artificial intelligence: Priorities and opportunities. *Cell*, 187(25):7045–7063, 2024. doi: 10.1016/j.cell.2024.11.015.

Haotian Cui, Chloe Wang, Hassaan Maan, Kuan Pang, Fengning Luo, Nan Duan, and Bo Wang. scGPT: toward building a foundation model for single-cell multi-omics using generative AI. *Nature Methods*, 21(8):1470–1480, 2024. doi: 10.1038/s41592-024-02201-0.

Atray Dixit, Oren Parnas, Biyu Li, Jenny Chen, Charles P. Fulco, Livnat Jerby-Arnon, Nemanja D. Marjanovic, Danielle Dionne, Tyler Burks, Raktima Raychowdhury, Britt Adamson, Thomas M. Norman, Eric S. Lander, Jonathan S. Weissman, Nir Friedman, and Aviv Regev. 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.

Jeffrey A. Farrell, Yiqun Wang, Samantha J. Riesenfeld, Karthik Shekhar, Aviv Regev, and Alexander F. Schier. Single-cell reconstruction of developmental trajectories during zebrafish embryogenesis. *Science*, 360(6392):eaar3131, 2018. doi: 10.1126/science.aar3131.

Yiming Gao, Zhen Wang, Jefferson Chen, Mark Antkowiak, Mengzhou Hu, JungHo Kong, Dexter Pratt, Jieyuan Liu, Enze Ma, Zhiting Hu, and Eric P. Xing. scPilot: Large language model reasoning toward automated single-cell analysis and discovery. In *Advances in Neural Information Processing Systems*, volume 38, pp. 1307–1344, 2025. arXiv:2602.11609.

Javier Gómez Ortega, Rangarajan D. Nadadur, Akira Kunitomi, Steven Kothen-Hill, Julian U. G. Wagner, Sarp D. Kurtoglu, Bumjoon Kim, Madigan M. Reid, Thomas Lu, Kaho Washizu, et al. Temporal AI model predicts drivers of cell state trajectories across human aging. *bioRxiv*, 2026. doi: 10.64898/2026.03.30.715396.

Nate Gruver, Marc Finzi, Shikai Qiu, and Andrew Gordon Wilson. Large language models are zero-shot time series forecasters. In *Advances in Neural Information Processing Systems*, volume 36, pp. 19622–19635, 2023.

Mahshid Heidari, Mina Karimpour, Sumana Srivatsa, and Hesam Montazeri. Evaluating single-cell perturbation response models is far from straightforward. *bioRxiv*, 2026. doi: 10.64898/2026.02.14.705879.

Guillaume Huguet, Daniel Sumner Magruder, Alexander Tong, Oluwadamilola Fasina, Manik Kuchroo, Guy Wolf, and Smita Krishnaswamy. Manifold interpolating optimal-transport flows for trajectory inference. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Qi Jiang, Lei Zhang, Longquan Li, and Lin Wan. Learning collective multi-cellular dynamics from temporal scRNA-seq via a transformer-enhanced neural SDE. *arXiv preprint arXiv:2505.16492*, 2025.

Dominik Klein, Giovanni Palla, Marius Lange, Michal Klein, Zoe Piran, Manuel Gander, Laetitia Meng-Papaxanthos, Michael Sterr, Lama Saber, Changying Jing, et al. Mapping cells through time and space with moscot. *Nature*, 638:1065–1075, 2025.

Gioele La Manno, Ruslan Soldatov, Amit Zeisel, Emelie Braun, Hannah Hochgerner, Viktor Petukhov, Katja Lidschreiber, Maria E. Kastriti, Peter Lönnerberg, Alessandro Furlan, et al. RNA velocity of single cells. *Nature*, 560(7719):494–498, 2018. doi: 10.1038/s41586-018-0414-6.

Marius Lange, Volker Bergen, Michal Klein, Manu Setty, Bernhard Reuter, Mostafa Bakhti, Heiko Lickert, Meshal Ansari, Janine Schniering, Herbert B. Schiller, Dana Pe'er, and Fabian J. Theis. CellRank for directed single-cell fate mapping. *Nature Methods*, 19:159–170, 2022.

Daniel Levine, Syed A. Rizvi, Sacha Lévy, Nazreen Pallikkavaliyaveetil, David Zhang, Xingyu Chen, Sina Ghadermarzi, Ruiming Wu, Zihe Zheng, Ivan Vrkic, Anna Zhong, Daphne Raskin, Insu Han, Antonio Henrique De Oliveira Fonseca, Josue Ortega Caro, Amin Karbasi, Rahul Madhav Dhodapkar, and David Van Dijk. Cell2Sentence: Teaching large language models the language of biology. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 27299–27325. PMLR, 2024.

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.

Erpai Luo, Haoxiang Gao, Haiyang Bian, Shuguang Peng, Yaru Li, Chen Li, Minsheng Hao, Mo Chen, Yuli She, Lei Wei, Kai Liu, and Xuegong Zhang. Forecasting cell state futures from static snapshots. *bioRxiv*, 2026. doi: 10.64898/2026.02.08.704720. Version 2, posted 31 July 2026.

Kevin R. Moon, David van Dijk, Zheng Wang, Scott Gigante, Daniel B. Burkhardt, William S. Chen, Kristina Yim, Antonia van den Elzen, Matthew J. Hirn, Ronald R. Coifman, Natalia B. Ivanova, Guy Wolf, and Smita Krishnaswamy. Visualizing structure and transitions in high-dimensional biological data. *Nature Biotechnology*, 37(12):1482–1492, 2019. doi: 10.1038/s41587-019-0336-3.

Thomas M. Norman, Max A. Horlbeck, Joseph M. Replogle, Alex Y. Ge, Albert Xu, Marco Jost, Luke A. Gilbert, and Jonathan S. Weissman. Exploring genetic interaction manifolds constructed from rich single-cell phenotypes. *Science*, 365(6455):786–793, 2019. doi: 10.1126/science.aax4438.

Adrien Osakwe, Eric H. Huang, and Yue Li. scTimeBench: a streamlined benchmarking platform for single-cell time-series analysis. *Bioinformatics*, 42(Supplement_2):btag414, 2026. doi: 10.1093/bioinformatics/btag414.

Blanca Pijuan-Sala, Jonathan A. Griffiths, Carolina Guibentif, Tom W. Hiscock, Wajid Jawaid, Fernando J. Calero-Nieto, Carla Mulas, Ximena Ibarra-Soria, Richard C. V. Tyser, Debbie Lee Lian Ho, Wolf Reik, Shankar Srinivas, Benjamin D. Simons, Jennifer Nichols, John C. Marioni, and Berthold Göttgens. A single-cell molecular map of mouse gastrulation and early organogenesis. *Nature*, 566(7745):490–495, 2019. doi: 10.1038/s41586-019-0933-9.

Mufan Qiu, Genhui Zheng, Yinuo Xu, Ruichen Zhang, Ying Ding, Qi Long, and Tianlong Chen. Chreode: A cell world model for one-step temporal dynamics and perturbation prediction. *arXiv preprint arXiv:2605.28111*, 2026.

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.

Yusuf H. Roohani, Tony J. Hua, Po-Yuan Tung, Lexi R. Bounds, Feiqiao B. Yu, Alexander Dobin, Noam Teyssier, Abhinav Adduri, Alden Woodrow, Brian S. Plosky, et al. Virtual Cell Challenge: Toward a Turing test for the virtual cell. *Cell*, 188(13):3370–3374, 2025. doi: 10.1016/j.cell.2025.06.008.

Geoffrey Schiebinger, Jian Shu, Marcin Tabaka, Brian Cleary, Vidya Subramanian, Aryeh Solomon, Joshua Gould, Siyan Liu, Stacie Lin, Peter Berube, Lia Lee, Jenny Chen, Justin Brumbaugh, Philippe Rigollet, Konrad Hochedlinger, Rudolf Jaenisch, Aviv Regev, and Eric S. Lander. Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming. *Cell*, 176(4):928–943.e22, 2019. doi: 10.1016/j.cell.2019.01.006.

Katherine M. Sheu, Aditya Pimplaskar, and Alexander Hoffmann. Single-cell stimulus-response gene expression trajectories reveal the stimulus specificities of dynamic responses by single macrophages. *Molecular Cell*, 84(21):4095–4110.e6, 2024. doi: 10.1016/j.molcel.2024.09.023.

Mingtian Tan, Mike A. Merrill, Vinayak Gupta, Tim Althoff, and Thomas Hartvigsen. Are language models actually useful for time series forecasting? In *Advances in Neural Information Processing Systems*, volume 37, pp. 60162–60191, 2024.

Alexander Tong, Jessie Huang, Guy Wolf, David Van Dijk, and Smita Krishnaswamy. TrajectoryNet: A dynamic optimal transport network for modeling cellular dynamics. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 9526–9536. PMLR, 2020.

Alexander Tong, Kilian Fatras, Nikolay Malkin, Guillaume Huguet, Yanlei Zhang, Jarrid Rector-Brooks, Guy Wolf, and Yoshua Bengio. Improving and generalizing flow-based generative models with minibatch optimal transport. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=CD9Snc73AW. arXiv:2302.00482.

Bram Van de Sande, Joon Sang Lee, Euphemia Mutasa-Gottgens, Bart Naughton, Wendi Bacon, Jonathan Manning, et al. Applications of single-cell RNA sequencing in drug discovery and development. *Nature Reviews Drug Discovery*, 22(6):496–520, 2023. doi: 10.1038/s41573-023-00688-4.

Shou-Wen Wang, Michael J. Herriges, Kilian Hurley, Darrell N. Kotton, and Allon M. Klein. CoSpar identifies early cell fate biases from single-cell transcriptomic and lineage information. *Nature Biotechnology*, 40:1066–1074, 2022.

Caleb Weinreb, Alejo Rodriguez-Fraticelli, Fernando D. Camargo, and Allon M. Klein. Lineage tracing on transcriptional landscapes links state to fate during differentiation. *Science*, 367(6479): eaaw3381, 2020. doi: 10.1126/science.aaw3381.

Yan Wu, Esther Wershof, Sebastian M. Schmon, Marcel Nassar, Błażej Osiński, Ridvan Eksi, Zichao Yan, Rory Stark, Kun Zhang, and Thore Graepel. PerturBench: Benchmarking machine learning models for cellular perturbation analysis. *arXiv preprint arXiv:2408.10609*, 2024.

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(1):3222, 2021. doi: 10.1038/s41467-021-23518-w.

Xinyu Yuan, Xixian Liu, Jianan Zhao, Yashi Zhang, Hongyu Guo, and Jian Tang. Plausibility is not prediction: Contrastive evidence for LLM-based cellular perturbation reasoning. *arXiv preprint arXiv:2606.01042*, 2026.

Alex L. Zhang, Tim Kraska, and Omar Khattab. Recursive language models. *arXiv preprint arXiv:2512.24601*, 2025.

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.

Xueya Zhou, Zihan Wang, Yue Ling, Qinxue Tian, Zhenyi Zhang, Yongge Li, Luonan Chen, and Peijie Zhou. Benchmarking zero-shot single-cell foundation model embeddings for cellular dynamics reconstruction. *bioRxiv*, 2026. doi: 10.64898/2026.03.10.710748.

A Datasets, models and full results

Static perturbation track. The Norman screen (Norman et al., 2019) uses single and paired CRISPR activation in 91,168 K562 cells. The Dixit screens (Dixit et al., 2016) use Perturb-seq in 195,990 cells: bone marrow-derived dendritic cells before and 3 hours after LPS stimulation, and K562 cells at 7 and 13 days. Both are mapped to the same gene axis as the temporal datasets. Both supply endpoint-response tasks. Norman holds out gene combinations with zero or one unseen gene. Dixit holds out an LPS response, a shift in perturbation duration and a perturbation. These tasks have no time axis in the sense of Section 3, so they are scored with the same evaluator but kept out of every temporal comparison. On Dixit, where no test perturbation appears in training, linear time, PCA-ridge, the source-population mean and the population MLP reach the same RMSE to four decimals on all three tasks. Persistence has the lowest Wasserstein distance on all three, and Table 22 gives every score.

Full results. Table 5 gives every population-level model on every temporal task, and Table 6 summarizes the interpolation tasks. Table 7 gives the clone-level LARRY track, and Table 8 every agent condition, including the recursive and specialist agents and the recall control. Tables 11 and 12 give the matrix of Table 5 under the Wasserstein distance and the Pearson correlation. Under the Wasserstein distance, the best fitted model beats every copy baseline on 4 of 16 interpolation tasks and on none of the 24 other tasks. The Pearson correlation picks the same best method as RMSE on 33 of the 40 tasks.

Table 3: Temporal single-cell benchmarks, and one perturbation challenge with a held-out cell type, compared with this work.

Benchmark	Held out	Shift in biology
scTimeBench (Osakwe et al., 2026)	time points within one setting	none
Virtual Cell Challenge (Roohani et al., 2025)	perturbation effects in a new cell type	cell type
Zebrafish course (Jiang et al., 2025)	intermediate and last two stages	none
Zero-shot embeddings (Zhou et al., 2026)	past, intermediate and future states	none
CellShift (this work)	intermediate and later times	fate, lineage, condition and starting population

Table 4: Methods evaluated in CellShift. Start is the population a method propagates or copies. Regime follows Section 3.2: pre-target methods train on no cell from the target day, and target-day methods train on other groups' cells at the target day.

Category	Method	Output	Start	Regime
Copy baselines	Persistence; nearest prior; test-context copy	Cells	Latest, nearest earlier or held-out group's context	Pre-target
	Neighbor interpolation	Pseudobulk	Both neighboring context times	Pre-target
	Observed-history mean; source-population mean	Pseudobulk	All context or source time	Pre-target
Regression baselines	Linear time; PCA-ridge; population-time MLP	Pseudobulk	Training means over permitted times	Pre-target
Dynamics	OT-CFM (Tong et al., 2024)	Cells	Latest context time	Pre-target; target-day on cross-fate and starting-population tasks
	PRESCIENT (Yeo et al., 2021)	Cells	Pooled population at the origin (LARRY day 2)	Pre-target
	scNODE (Zhang et al., 2024)	Cells	Pooled population at the origin (LARRY day 2)	Pre-target
Clone-level (LARRY)	Training mean; mean shift; PCA-ridge; PCA-kNN; kNN-copy; clone-paired MLP; scNODE; OT-CFM	Cells per clone	Test clone's early cells	Target-day
	scGPT embedding + kNN (Cui et al., 2024)	Cells per clone	Test clone's early cells	Target-day
	Linear trend through the test clones' own pseudobulk	Cells per clone	Test clone's own cells	Pre-target
Pretrained	CellTempo, released, from scratch, and local pretraining with fine-tuning (Luo et al., 2026)	Tokens decoded to cells	Latest context	Pre-target (possible pretraining overlap)
	MaxToki, released and from scratch (Gómez Ortega et al., 2026)	Gene ranks	Latest context	Pre-target (rank track, possible pretraining overlap)
Agents	Zero-call selector; single agent, with and without feedback	Pseudobulk	Test observed context	Pre-target
	Recursive and delegation disabled	Pseudobulk	Training split	Target-day
	Specialists (gene modules or random gene groups); recall control	Pseudobulk	Test observed context	Pre-target

Table 5: Pseudobulk RMSE (↓) of every population-level method on every temporal task. **Bold** marks the best value in each row. On LARRY *in vitro*, ± gives the standard deviation over three seeds; all other cells are single runs. A blank cell marks a method that is not defined for the task (for example, test-context copy without test context), and n/a a task that the method's implementation does not support. Nearest prior copies the closest earlier observed time and equals persistence on forecasting tasks. Neighbor interp. averages the two context times around the target, weighted linearly in time. PRESCIENT and scNODE simulate from the pooled population at the task's origin (day 2 on LARRY, also on the day 2+4→6 tasks); OT-CFM and the copy baselines start from the latest context time. ^{t}Target-day regime: OT-CFM trains on the other populations' target-day cells. OT-CFM selects its step cap on validation data, or on a held-out training time where a task has no validation split. The scNODE forecasts on Farrell and CellTag and the diverged OT-CFM forecast on Pijuan-Sala E8.5 are treated as failed runs; no failed run is ever the best of a row, so none changes a count.

		Copy and regression baselines									Dynamics models		
Task	Target	Persist.	Nearest prior	Test ctx.	Neighbor interp.	Obs. mean	Src. mean	Linear	PCA-ridge	MLP (pop.)	OT-CFM	PRESCIENT	scNODE
LARRY in vitro													
Interp.	day 4	0.1014	0.1103	0.1183	0.0537	0.0754	0.0985	0.0538	0.0518	0.0524$_{\pm 0.0010}$	**0.0466**$_{\pm 0.0017}$	0.0501$_{\pm 0.0071}$	0.0569$_{\pm 0.0059}$
Forecast	day 6 (2+4)	0.0874		0.1076		0.1080	0.0850	0.0886	0.0800	0.0600$_{\pm 0.0016}$	0.0882$_{\pm 0.0068}$	**0.0519**$_{\pm 0.0116}$	0.0773$_{\pm 0.0141}$
Cross-fate	neutrophil	0.1168		0.1119		0.1338	0.1329	0.1553	0.1466	0.1252$_{\pm 0.0012}$	**0.0918**$_{\pm 0.0009}^{t}$	0.1281$_{\pm 0.0049}$	0.1413$_{\pm 0.0053}$
Cross-fate	monocyte	0.1189		**0.0918**		0.1354	0.1483	0.1653	0.1570	0.1415$_{\pm 0.0007}$	0.0992$_{\pm 0.0029}^{t}$	0.1479$_{\pm 0.0017}$	0.1583$_{\pm 0.0026}$
Cross-fate	basophil	0.1371		**0.0657**		0.1468	0.1597	0.1897	0.1875	0.1624$_{\pm 0.0008}$	0.1365$_{\pm 0.0039}^{t}$	0.1571$_{\pm 0.0042}$	0.1789$_{\pm 0.0084}$
Start.-pop.	LK	0.1025		0.0941		0.1220	0.1336	0.1253	0.1153	0.1181$_{\pm 0.0003}$	**0.0789**$_{\pm 0.0009}^{t}$	0.1323$_{\pm 0.0067}$	0.1376$_{\pm 0.0097}$
Start.-pop.	LSK	0.0772		0.1020		0.0969	0.0751	0.1320	0.1343	**0.0712**$_{\pm 0.0014}$	0.0822$_{\pm 0.0014}^{t}$	0.0933$_{\pm 0.0086}$	0.1144$_{\pm 0.0043}$
LARRY cytokine													
Interp.	day 4	0.1142	0.1245	0.1286	0.0536	0.0723	0.1146	0.0545	0.0539	0.0541	**0.0527**	0.0535	0.0638
Forecast	day 6 (2+4)	0.1097		0.1120		0.1397	0.1079	0.0967	0.0966	0.0849	0.0998	0.1167	**0.0834**
Interp., full cocktail	day 4	0.1606	0.1770	0.2035	**0.1338**	0.1365	0.1626	0.1347	0.1343	0.1343	n/a	0.1342	0.1454
Forecast, full cocktail	day 6 (2+4)	**0.1397**		0.1664		0.1622	0.1470	0.2058	0.1988	0.1752	n/a	0.1989	0.2485
Cross-condition	full cocktail	0.0840		0.0982		0.0868	**0.0831**	**0.0831**	**0.0831**	**0.0831**	n/a	n/a	n/a
Schiebinger													
Interp.	day 5	0.1155	0.0296		**0.0150**	0.0730	0.1155	0.0649	0.0639	0.0459	0.0337	0.0376	0.0701
Interp.	day 10	0.0976	0.0455		**0.0400**	0.0801	0.0976	0.0803	0.0803	0.0683	0.0560	0.0449	0.0807
Interp.	day 15	0.0767	0.0459		**0.0423**	0.0667	0.0767	0.0492	0.0502	0.0433	0.0490	0.0445	0.0720
Forecast	day 16	**0.0815**				0.1018	**0.0815**	0.0931	0.0929	0.1269	0.1192	0.0822	0.1136
Forecast	day 17	**0.0461**				0.0809	**0.0461**	0.0888	0.0910	0.1088	0.1327	0.0475	0.1014
Forecast	day 18	**0.0704**				0.0906	**0.0704**	0.0963	0.0985	0.1371	0.2036	0.0719	0.1160
Farrell													
Interp.	4.7 hpf	0.2068	0.1015		0.0657	0.1298	0.2068	0.0793	0.0784	0.0552	0.0559	**0.0448**	0.0662
Interp.	6.0 hpf	0.1661	0.0990		0.0839	0.0939	0.1661	0.0782	0.0779	0.0745	0.0843	**0.0676**	0.1706
Interp.	8.0 hpf	0.1159	0.0508		**0.0213**	0.0453	0.1159	0.0557	0.0557	0.0396	0.0285	0.0317	0.2959
Forecast	10 hpf	0.0740				0.1274	0.0740	0.1033	0.1040	**0.0613**	0.2134	0.0669	0.5393
Forecast	11 hpf	0.1083				0.1547	0.1083	0.1472	0.1476	**0.0942**	0.2998	0.0979	0.6357
Forecast	12 hpf	0.1159				0.1605	0.1159	0.1765	0.1767	**0.1079**	0.3766	0.1082	0.7196
Pijuan-Sala													
Interp.	E7.0	0.0730	0.0436		0.0335	0.0433	0.0747	**0.0289**	0.0317	0.0307	0.0385	0.0320	0.0476
Interp.	E7.5	0.0648	0.0498		0.0296	0.0321	0.0648	0.0273	0.0280	**0.0261**	0.0407	0.0371	0.0340
Interp.	E8.0	0.0440	0.0386		0.0316	0.0276	0.0440	0.0268	**0.0262**	0.0276	0.0397	0.0401	0.0382
Forecast	E8.5	0.0648				0.0674	0.0648	0.0799	0.0769	0.0697	0.7901	**0.0588**	0.0764
Cross-lineage	E8.5 mesoderm	0.0637		**0.0505**		0.0832	0.0714	0.0768	0.0766	0.0725	n/a	n/a	n/a
Matched control	E8.5 mesoderm	0.0569		**0.0505**		0.0749	0.0565	0.0564	0.0557	0.0563	n/a	n/a	n/a
CellTag													
Interp.	day 9	0.1278	0.0527	0.0493	**0.0369**	0.0613	0.1274	0.0508	0.0504	0.0534	0.0406	0.0450	0.0618
Forecast	day 21	0.0840		0.0957		0.1075	**0.0810**	0.1242	0.1253	0.0876	0.1060	0.1208	0.5134
Forecast	day 28	0.1149		0.1293		0.1341	**0.1111**	0.1668	0.1649	0.1145	0.2944	0.1391	0.7864
Cross-fate fc.	dead end to succ.	**0.1177**		0.1484		0.1234	0.1200	0.1641	0.1636	0.1255	n/a	n/a	n/a
Cross-fate fc.	succ. to dead end	**0.1202**		0.1707		0.1454	0.1206	0.1513	0.1471	0.1226	n/a	n/a	n/a
Cross-fate int.	dead end to succ.	0.1737	0.0528	**0.0504**	0.0668	0.1093	0.1726	0.0836	0.0836	0.0811	n/a	0.0615	0.1077
Cross-fate int.	succ. to dead end	0.1622	0.0424	**0.0385**	0.0594	0.1037	0.1623	0.0895	0.0900	0.0833	n/a	0.0630	0.1067
Moon													
Interp.	day 7.5	0.0827	0.0683		**0.0329**	0.0442	0.0827	0.0427	0.0442	0.0442	0.0423	0.0404	0.0503
Interp.	day 13.5	0.0534	0.0475		**0.0243**	0.0288	0.0534	0.0380	0.0382	0.0310	0.0289	0.0321	0.0428
Interp.	day 19.5	0.0435	0.0292		**0.0253**	0.0401	0.0435	0.0316	0.0316	0.0479	0.0270	0.0291	0.0391

Table 6: Interpolation tasks, common evaluator, RMSE. Neighbor interpolates linearly between the two observed populations around the target. Best control is the best of persistence, the nearest-prior and test-context copies, the observed and source means, linear time, PCA-ridge and the population-time MLP. Best dynamics is the best of OT-CFM, PRESCIENT and scNODE. **Bold** marks the best of the three columns. ‡OT-CFM does not support the task.

Dataset	Target	Neighbor	Best control	Best dynamics
LARRY *in vitro*	day 4	0.0537	0.0518 PCA-ridge	**0.0466** OT-CFM
LARRY cytokine	day 4	0.0536	0.0539 PCA-ridge	**0.0527** OT-CFM
LARRY cytokine	day 4, full cocktail	**0.1338**	0.1343 PCA-ridge	0.1342 PRESCIENT‡
Schiebinger	day 5	**0.0150**	0.0296 nearest prior	0.0337 OT-CFM
Schiebinger	day 10	**0.0400**	0.0455 nearest prior	0.0449 PRESCIENT
Schiebinger	day 15	**0.0423**	0.0433 MLP	0.0445 PRESCIENT
Farrell	4.7 hpf	0.0657	0.0552 MLP	**0.0448** PRESCIENT
Farrell	6.0 hpf	0.0839	0.0745 MLP	**0.0676** PRESCIENT
Farrell	8.0 hpf	**0.0213**	0.0396 MLP	0.0285 OT-CFM
Pijuan-Sala	E7.0	0.0335	**0.0289** linear	0.0320 PRESCIENT
Pijuan-Sala	E7.5	0.0296	**0.0261** MLP	0.0340 scNODE
Pijuan-Sala	E8.0	0.0316	**0.0262** PCA-ridge	0.0382 scNODE
CellTag	day 9	**0.0369**	0.0493 test context	0.0406 OT-CFM
Moon	day 7.5	**0.0329**	0.0427 linear	0.0404 PRESCIENT
Moon	day 13.5	**0.0243**	0.0288 observed mean	0.0289 OT-CFM
Moon	day 19.5	**0.0253**	0.0292 nearest prior	0.0270 OT-CFM

Across the six tasks in Table 9, the released CellTempo checkpoint reaches a Pearson correlation between 0.72 and 0.82, yet its RMSE is 2.2 to 3.3 times that of persistence. This error lies close to the encoding and decoding error of the tokenizer itself, about 0.28 on LARRY cells, which points to a mismatch in preprocessing or data distribution between the pretraining corpus and these measurements. Its left-to-right decoder cannot perform interpolation tasks. The locally trained checkpoint collapses to three distinct expression vectors across 17 predictions (Table 10), so its scores reflect a failed training run.

Table 7: LARRY *in vitro* forecasts that start from each test clone's own early cells, pseudobulk RMSE (↓). **Bold** marks the best value in each column. ± gives the standard deviation over three seeds. A dash marks a model that was not run on the task. The training-population mean predicts one expression vector per task. The zero-call selector (Table 8) extrapolates the test clones' pseudobulk.

	Forecast		Starting population		Cross-fate		
Method	day 6 (2+4)	day 4 (2)	LK	LSK	neutrophil	monocyte	basophil
Persistence	0.1076	0.1183	0.0941	0.1020	0.1119	0.0918	**0.0657**
Mean shift	0.0689	0.0606	0.0975	0.0859	0.0934	0.0753	0.0668
Training mean	**0.0243**	0.0369	0.0789	0.0794	0.1374	0.1352	0.1495
PCA-ridge	0.0540	0.0381	0.0652	0.0673	0.1423	0.1266	0.1479
PCA-kNN	0.0563	0.0419	0.0830	0.0763	0.1485	0.1333	0.1474
scGPT + kNN	0.0527	0.0424	0.0864	0.0798	0.1556	0.1362	0.1518
scNODE	0.0643±0.0124	0.0493±0.0046	0.0850±0.0060	0.0897±0.0046	0.1261±0.0042	0.1434±0.0020	0.1427±0.0011
kNN-copy	0.0414±0.0008	0.0572±0.0008	0.0727±0.0012	0.0823±0.0028	0.1425±0.0015	0.1381±0.0005	0.1523±0.0007
MLP (clone)	0.0312±0.0005	0.0400±0.0007	**0.0577**±0.0016	**0.0597**±0.0023	0.1268±0.0012	0.1243±0.0012	0.1291±0.0031
Linear	0.1116	–	0.0998	0.0912	0.1080	0.0900	0.0992
OT-CFM (clone)	0.0497±0.0015	**0.0336**±0.0017	0.0786±0.0011	0.0792±0.0013	**0.0916**±0.0011	0.0987±0.0023	0.1424±0.0006
Zero-call selector	0.0946	–	–	–	0.0918	**0.0687**	0.0673

Table 8: Agentic conditions on LARRY *in vitro*, pseudobulk RMSE (↓), mean ± standard deviation over three repeats. The zero-call selector is deterministic and was run once. Calls and tokens (thousands, input and output) are means per run. The recall control runs under neutral run-folder names. [r]Target-day regime: the recursive runs build their context from the training split, which contains training clones' target-day cells, so they are not comparable with the selector and the single agents.

			Held-out target			
Condition	Calls	Tokens (k)	Day 6 (2+4)	Neutrophil	Monocyte	Basophil
Zero-call selector	0.0	0.0 / 0.0	0.0946	0.0918	0.0687	0.0673
Single agent	11.4	54.1 / 9.9	0.0953±0.0015	0.0955±0.0024	0.0707±0.0029	0.0652±0.0097
Single agent with validation feedback	22.0	102.0 / 25.0	0.0997±0.0047	0.0941±0.0032	0.0787±0.0115	0.0627±0.0056
Delegation disabled[r]	3.4	37.2 / 4.3	0.0457±0.0372	0.1596±0.0384	0.1352±0.0000	0.1509±0.0024
Recursive[r]	11.2	98.5 / 28.9	0.0625±0.0379	0.1369±0.0005	0.1489±0.0152	0.1622±0.0235
Specialists, biological groups	74.3	529.5 / 51.1	0.1165±0.0114	–	–	–
Specialists, random groups	73.3	410.4 / 31.1	0.1214±0.0028	–	–	–
Recall control, original inputs	16.0	37.5 / 16.1	0.1051±0.0044	–	–	–
Recall control, anonymized	20.7	47.4 / 29.6	0.1045±0.0053	–	–	–
Recall control, noise inputs	13.3	43.0 / 15.4	0.1415±0.0023	–	–	–

Table 9: Released CellTempo checkpoint, re-scored once with its released tokenizer, common evaluator, single run, stochastic top-k decoding chosen on training and validation data. Pretraining overlap with the target studies is unresolved. The fine-tuned neutrophil row is a fresh six-epoch fine-tune. The last column divides RMSE by that of persistence on the same task.

Dataset	Task	RMSE	Pearson	× persistence
LARRY *in vitro*	Day 2+4 → 6 forecast	0.2908	0.7334	3.33
LARRY *in vitro*	Held-out neutrophil, fine-tuned	0.3156	0.7226	2.70
CellTag	Forecast → day 21	0.2342	0.8217	2.79
CellTag	Forecast → day 28	0.2516	0.7835	2.19
CellTag	Dead end → successful	0.2554	0.7771	2.17
CellTag	Successful → dead end	0.2622	0.7682	2.18

Table 10: CellTempo checkpoints trained locally on the scBaseTraj collection, common evaluator, seed 1729, greedy decoding. Top, the released architecture trained from random initialization. Bottom, that checkpoint fine-tuned on each LARRY training split and selected on its validation split. Fine-tuning left the day 2+4 $\rightarrow$ 6 scores unchanged to four decimals, and the fine-tuned checkpoint collapses to three distinct expression vectors across its 17 predictions, so identical targets receive identical scores. These rows record a failed training run and support no claim about the architecture. Greedy decoding was the default of the local pipeline; the released checkpoint in Table 9 uses stochastic decoding selected on validation data. The last column divides RMSE by that of persistence on the same task. [f] Forward-only control on an interpolation target, not an interpolation.

Task	RMSE	Pearson	$\times$ persistence
From random initialization			
LARRY *in vitro* day 2+4 $\rightarrow$ 6 forecast	0.3648	0.5770	4.17
Local pretraining, fine-tuned per task			
LARRY *in vitro* day 2+4 $\rightarrow$ 6 forecast	0.3648	0.5770	4.17
LARRY *in vitro* day 2 $\rightarrow$ 4 forecast	0.3630	0.5740	3.29
LARRY *in vitro* day 2 $\rightarrow$ 6 forecast	0.3648	0.5770	1.95
LARRY *in vitro* day-4 target, interpolation split[f]	0.3630	0.5740	3.58
LARRY *in vitro* cross-fate neutrophil	0.3764	0.5418	3.22
LARRY *in vitro* cross-fate monocyte	0.3731	0.5579	3.14
LARRY *in vitro* cross-fate basophil	0.3795	0.5472	2.77
LARRY *in vitro* starting-population LK	0.3679	0.5672	3.59
LARRY *in vitro* starting-population LSK	0.3650	0.5775	4.73
LARRY cytokine day 2+4 $\rightarrow$ 6 forecast	0.3687	0.5705	3.36
LARRY cytokine day 2 $\rightarrow$ 4 forecast	0.3531	0.6021	2.84
LARRY cytokine day 2 $\rightarrow$ 6 forecast	0.3687	0.5705	1.75
LARRY cytokine day-4 target, interpolation split[f]	0.3531	0.6021	3.09
LARRY cytokine full-cocktail day 2+4 $\rightarrow$ 6	0.3729	0.5595	2.67
LARRY cytokine full-cocktail day 2 $\rightarrow$ 4	0.3702	0.5773	2.09
LARRY cytokine full-cocktail day 2 $\rightarrow$ 6	0.3729	0.5595	1.86
LARRY cytokine full-cocktail day-4 target, interpolation split[f]	0.3702	0.5773	2.30

Table 11: Wasserstein distance (W_1, ↓) between predicted and observed cells in a training-fitted 50-component PCA space, for every cell of Table 5. **Bold** marks the best value in each row. Marks and conventions follow Table 5.

		Copy and regression baselines									Dynamics models		
Task	Target	Persist.	Nearest prior	Test ctx.	Neighbor interp.	Obs. mean	Src. mean	Linear	PCA-ridge	MLP (pop.)	OT-CFM	PRESCIENT	scNODE
LARRY in vitro													
Interp.	day 4	**0.860**	1.129	1.005	1.566	1.585	1.626	1.567	1.555	$1.556_{\pm 0.001}$	$0.919_{\pm 0.002}$	$0.866_{\pm 0.051}$	$0.881_{\pm 0.017}$
Forecast	day 6 (2+4)	0.747		**0.688**		1.763	1.726	1.829	1.746	$1.695_{\pm 0.002}$	$1.195_{\pm 0.012}$	$1.143_{\pm 0.053}$	$1.033_{\pm 0.072}$
Cross-fate	neutrophil	1.303		**0.901**		1.870	1.950	2.031	1.952	$1.908_{\pm 0.004}$	$1.191_{\pm 0.005}{}^{t}$	$1.660_{\pm 0.044}$	$1.601_{\pm 0.036}$
Cross-fate	monocyte	1.241		**0.792**		1.752	1.861	1.992	1.930	$1.828_{\pm 0.005}$	$1.150_{\pm 0.016}{}^{t}$	$1.757_{\pm 0.100}$	$1.614_{\pm 0.037}$
Cross-fate	basophil	1.310		**0.446**		1.689	1.816	2.039	1.975	$1.842_{\pm 0.004}$	$1.216_{\pm 0.019}{}^{t}$	$1.635_{\pm 0.100}$	$1.783_{\pm 0.049}$
Start.-pop.	LK	0.933		**0.747**		1.848	1.957	2.063	1.971	$1.934_{\pm 0.003}$	$1.080_{\pm 0.004}{}^{t}$	$1.560_{\pm 0.006}$	$1.468_{\pm 0.040}$
Start.-pop.	LSK	**0.722**		0.787		1.618	1.593	1.826	1.823	$1.598_{\pm 0.004}$	$1.071_{\pm 0.007}{}^{t}$	$1.360_{\pm 0.018}$	$1.244_{\pm 0.020}$
LARRY cytokine													
Interp.	day 4	1.019	1.063	0.985	1.573	1.577	1.660	1.577	1.575	1.578	0.882	**0.873**	1.068
Forecast	day 6 (2+4)	0.927		**0.774**		1.771	1.709	1.771	1.736	1.676	1.381	1.324	1.140
Interp., full cocktail	day 4	2.306	1.955	1.129	1.268	1.243	1.356	1.274	1.271	1.272	n/a	**0.919**	1.298
Forecast, full cocktail	day 6 (2+4)	1.632		**0.970**		1.557	1.657	2.197	2.084	1.920	n/a	1.555	3.299
Cross-condition	full cocktail	**0.902**		0.978		1.851	1.856	1.856	1.856	1.856	n/a	n/a	n/a
Schiebinger													
Interp.	day 5	1.351	**0.363**		0.912	1.206	1.463	1.159	1.152	1.053	0.415	0.554	0.984
Interp.	day 10	1.086	**0.504**		1.437	1.607	1.743	1.611	1.610	1.548	0.613	0.738	1.295
Interp.	day 15	1.038	**0.590**		1.741	1.823	1.955	1.775	1.781	1.755	0.684	0.728	1.529
Forecast	day 16	**0.946**				1.744	1.637	1.709	1.709	1.792	1.534	0.995	1.500
Forecast	day 17	**0.574**				1.726	1.582	1.744	1.773	1.760	1.857	0.638	1.425
Forecast	day 18	**0.820**				1.744	1.639	1.754	1.771	1.883	2.835	0.888	1.504
Farrell													
Interp.	4.7 hpf	1.151	0.871		1.202	1.446	1.695	1.265	1.253	1.150	**0.731**	0.808	0.981
Interp.	6.0 hpf	1.068	**0.842**		1.377	1.432	1.693	1.360	1.357	1.336	1.074	1.075	1.573
Interp.	8.0 hpf	0.966	**0.376**		1.137	1.209	1.546	1.244	1.245	1.185	0.637	0.880	2.213
Forecast	10 hpf	**0.644**				1.562	1.300	1.381	1.394	1.195	2.065	1.071	3.639
Forecast	11 hpf	**0.900**				1.670	1.401	1.500	1.515	1.252	2.643	1.132	4.167
Forecast	12 hpf	**0.935**				1.706	1.435	1.585	1.607	1.306	3.151	1.220	4.670
Pijuan-Sala													
Interp.	E7.0	0.797	**0.553**		1.076	1.118	1.260	1.059	1.067	1.059	0.574	0.630	0.807
Interp.	E7.5	0.772	**0.514**		1.124	1.110	1.243	1.097	1.100	1.096	0.530	0.698	0.723
Interp.	E8.0	0.574	**0.446**		1.099	1.091	1.142	1.081	1.081	1.083	0.509	0.741	0.778
Forecast	E8.5	**0.728**				1.173	1.150	1.250	1.236	1.156	8.049	0.780	1.047
Cross-lineage	E8.5 mesoderm	0.751		**0.577**		1.208	1.136	1.194	1.188	1.164	n/a	n/a	n/a
Matched control	E8.5 mesoderm	0.689		**0.604**		1.310	1.200	1.214	1.215	1.195	n/a	n/a	n/a
CellTag													
Interp.	day 9	0.963	0.604	0.582	1.281	1.387	1.650	1.354	1.352	1.362	**0.509**	0.918	0.964
Forecast	day 21	**0.848**		0.935		1.521	1.413	1.620	1.596	1.476	1.154	1.508	3.997
Forecast	day 28	**1.220**		1.284		1.764	1.699	2.082	2.015	1.759	3.118	1.842	6.122
Cross-fate fc.	dead end to succ.	**0.976**		0.993		1.379	1.332	1.591	1.582	1.382	n/a	n/a	n/a
Cross-fate fc.	succ. to dead end	1.403		**1.059**		1.375	1.308	1.523	1.509	1.336	n/a	n/a	n/a
Cross-fate int.	dead end to succ.	1.963	0.522	**0.379**	0.988	1.126	1.430	1.108	1.109	1.091	n/a	0.595	1.025
Cross-fate int.	succ. to dead end	2.148	0.317	**0.242**	0.848	0.991	1.203	1.000	1.011	0.977	n/a	0.465	0.846
Moon													
Interp.	day 7.5	0.810	**0.615**		1.122	1.157	1.303	1.156	1.164	1.163	0.674	0.854	0.899
Interp.	day 13.5	0.629	**0.449**		1.244	1.255	1.345	1.282	1.283	1.270	0.589	0.910	0.966
Interp.	day 19.5	0.541	**0.304**		1.261	1.288	1.310	1.279	1.280	1.314	0.537	0.724	1.012

Table 12: Pearson correlation (↑) between predicted and observed pseudobulk means on the 18,791-gene axis, for every cell of Table 5. **Bold** marks the best value in each row. Marks and conventions follow Table 5.

		Copy and regression baselines									Dynamics models		
Task	Target	Persist.	Nearest prior	Test ctx.	Neighbor interp.	Obs. mean	Src. mean	Linear	PCA-ridge	MLP (pop.)	OT-CFM	PRESCIENT	scNODE
LARRY in vitro													
Interp.	day 4	0.954	0.945	0.939	0.988	0.978	0.957	0.988	0.989	0.988$_{\pm 0.001}$	**0.990**$_{\pm 0.001}$	0.989$_{\pm 0.003}$	0.985$_{\pm 0.003}$
Forecast	day 6 (2+4)	0.963		0.947		0.942	0.965	0.968	0.975	0.985$_{\pm 0.001}$	0.972$_{\pm 0.003}$	**0.986**$_{\pm 0.003}$	0.969$_{\pm 0.007}$
Cross-fate	neutrophil	0.930		0.944		0.908	0.905	0.887	0.896	0.917$_{\pm 0.000}$	**0.952**$_{\pm 0.001}{}^{t}$	0.900$_{\pm 0.009}$	0.893$_{\pm 0.003}$
Cross-fate	monocyte	0.928		**0.958**		0.907	0.887	0.865	0.878	0.897$_{\pm 0.000}$	0.950$_{\pm 0.003}{}^{t}$	0.885$_{\pm 0.005}$	0.869$_{\pm 0.003}$
Cross-fate	basophil	0.903		**0.979**		0.890	0.869	0.829	0.833	0.866$_{\pm 0.001}$	0.889$_{\pm 0.005}{}^{t}$	0.860$_{\pm 0.009}$	0.836$_{\pm 0.009}$
Start.-pop.	LK	0.950		0.965		0.926	0.900	0.922	0.934	0.925$_{\pm 0.001}$	**0.981**$_{\pm 0.000}{}^{t}$	0.920$_{\pm 0.008}$	0.906$_{\pm 0.011}$
Start.-pop.	LSK	0.970		0.943		0.953	**0.975**	0.917	0.914	**0.975**$_{\pm 0.001}$	0.969$_{\pm 0.001}{}^{t}$	0.950$_{\pm 0.010}$	0.922$_{\pm 0.008}$
LARRY cytokine													
Interp.	day 4	0.937	0.938	0.933	0.986	0.976	0.937	0.986	0.986	0.986	**0.987**	**0.987**	0.983
Forecast	day 6 (2+4)	0.944		0.939		0.909	0.945	0.960	0.961	**0.967**	0.963	0.941	0.963
Interp., full cocktail	day 4	0.903	0.874	0.841	0.927	**0.930**	0.900	0.927	0.927	0.927	n/a	0.929	0.916
Forecast, full cocktail	day 6 (2+4)	**0.927**		0.890		0.892	0.921	0.872	0.881	0.903	n/a	0.846	0.870
Cross-condition	full cocktail	**0.970**		0.968		0.968	**0.970**	**0.970**	**0.970**	**0.970**	n/a	n/a	n/a
Schiebinger													
Interp.	day 5	0.957	0.997		**0.999**	0.985	0.957	0.987	0.988	0.994	0.996	0.995	0.984
Interp.	day 10	0.962	0.992		**0.994**	0.975	0.962	0.975	0.975	0.982	0.989	0.992	0.975
Interp.	day 15	0.977	0.992		**0.993**	0.983	0.977	0.990	0.990	0.992	0.992	0.992	0.980
Forecast	day 16	**0.979**				0.962	**0.979**	0.967	0.966	0.942	0.955	**0.979**	0.948
Forecast	day 17	**0.992**				0.975	**0.992**	0.970	0.968	0.955	0.940	**0.992**	0.959
Forecast	day 18	**0.983**				0.969	**0.983**	0.964	0.962	0.932	0.877	0.982	0.947
Farrell													
Interp.	4.7 hpf	0.826	0.963		0.985	0.934	0.826	0.982	0.983	0.991	0.988	**0.992**	0.985
Interp.	6.0 hpf	0.880	0.972		0.980	0.964	0.880	0.978	0.979	0.980	0.981	**0.982**	0.953
Interp.	8.0 hpf	0.952	0.992		**0.998**	0.994	0.952	0.990	0.990	0.995	0.997	0.997	0.899
Forecast	10 hpf	0.981				0.940	0.981	0.978	0.977	**0.988**	0.812	0.987	0.784
Forecast	11 hpf	0.958				0.910	0.958	0.957	0.956	0.972	0.674	**0.973**	0.756
Forecast	12 hpf	0.952				0.903	0.952	0.946	0.944	**0.965**	0.618	0.962	0.753
Pijuan-Sala													
Interp.	E7.0	0.985	0.995		0.997	0.995	0.984	**0.998**	0.997	**0.998**	0.996	0.997	0.995
Interp.	E7.5	0.988	0.993		**0.998**	0.997	0.988	**0.998**	**0.998**	**0.998**	0.995	0.996	0.997
Interp.	E8.0	0.995	0.996		0.997	**0.998**	0.995	**0.998**	**0.998**	**0.998**	0.996	0.995	0.996
Forecast	E8.5	0.988				0.987	0.988	0.982	0.984	0.986	0.620	**0.990**	0.983
Cross-lineage	E8.5 mesoderm	0.989		**0.993**		0.980	0.986	0.983	0.983	0.985	n/a	n/a	n/a
Matched control	E8.5 mesoderm	0.991		**0.993**		0.984	0.991	0.991	0.991	0.991	n/a	n/a	n/a
CellTag													
Interp.	day 9	0.944	0.991	0.992	**0.996**	0.988	0.945	0.992	0.992	0.991	0.995	0.994	0.987
Forecast	day 21	0.976		0.970		0.960	**0.977**	0.949	0.946	0.974	0.960	0.953	0.603
Forecast	day 28	0.954		0.944		0.937	**0.956**	0.915	0.913	**0.956**	0.772	0.933	0.591
Cross-fate fc.	dead end to succ.	**0.956**		0.927		0.950	0.954	0.921	0.921	0.951	n/a	n/a	n/a
Cross-fate fc.	succ. to dead end	**0.952**		0.906		0.929	0.951	0.925	0.929	0.949	n/a	n/a	n/a
Cross-fate int.	dead end to succ.	0.900	0.991	**0.992**	0.986	0.961	0.901	0.978	0.978	0.979	n/a	0.988	0.964
Cross-fate int.	succ. to dead end	0.915	0.994	**0.995**	0.989	0.966	0.915	0.975	0.975	0.979	n/a	0.988	0.966
Moon													
Interp.	day 7.5	0.976	0.985		**0.996**	0.993	0.976	0.994	0.993	0.993	0.994	0.994	0.991
Interp.	day 13.5	0.989	0.992		**0.998**	0.997	0.989	0.995	0.994	0.996	0.997	0.996	0.993
Interp.	day 19.5	0.993	0.997		**0.998**	0.994	0.993	0.996	0.996	0.991	0.997	0.997	0.994

B ADDITIONAL RESULTS

Target-day supervision. A separate clone-level pipeline learns from other training clones at the target day (Table 7). On the in-domain day-6 task, the training mean scores 0.0243 and the clone-paired MLP 0.0312 ± 0.0005. The matched-fate control can draw on same-fate day-6 training cells, whereas the held-out-fate task cannot. The resulting RMSE gap therefore does not isolate learned dynamics. On held-out fates in this track, the zero-call selector is best on monocytes and within 0.0002 of the clone-level OT-CFM on neutrophils. It extrapolates the test clones' own pseudobulk. Per-clone persistence is best on basophils. Table 19 gives the same track on LARRY cytokine, where the training mean is best on seven of eight tasks. Table 20 adds cross-fate interpolation and the cytokine matched-condition controls.

Condition shift and CellTag fates. On the strict held-out cytokine condition, persistence scores 0.0840 and the fitted controls 0.0831. No dynamics model can be run on the task, because none accepts a condition descriptor. The task is a forecast made after the model has seen the early response of the condition. It does not predict an unseen intervention. On the CellTag cross-fate tasks, pooled persistence is best when forecasting a held-out fate (0.1177 and 0.1202). The test-context copy is best when interpolating one (0.0504 and 0.0385), ahead of PRESCIENT (0.0615 and 0.0630).

Replicate transfer. CellTag has two Drop-seq replicates. Trained on one replicate and asked for the whole other one, the observed-history mean reaches an RMSE of 0.0326 in both directions, and every other applicable control lies between 0.0956 and 0.1031 (Table 21). The mean expression profile of one replicate transfers almost unchanged to the other.

C DATASETS AND SPLITS

Gene axis. All datasets are projected onto one axis of 18,791 human gene symbols, in the same order for every dataset. Mouse genes are mapped through MGI homology classes on linear values before normalization. Zebrafish genes are mapped through the same homology resource, and only one-to-one orthologs are kept. This keeps 9,061 Farrell genes, and 9,030 of them carry data. The other columns are zero for every Farrell cell and every model, which scales Farrell RMSE values by about 0.7. Counts are renormalized to 10,000 per cell and log1p-transformed. Every score uses the whole axis.

LARRY *in vitro*. Mouse hematopoietic progenitors in culture carry the LARRY barcode and are sampled at days 2, 4 and 6 (28,249, 48,498 and 54,140 cells). Two sorted starting populations are present, LSK (72,946 cells) and LK (57,941). Terminal fate is annotated at day 6 for three mature populations, neutrophil (14,448 cells), monocyte (10,727) and basophil (6,721). Of the 130,887 cells, 49,302 carry a clone identifier across 5,864 clones. The tasks use the 407 clones that are observed at all three days and have a confident fate. These hold 5,022 cells (578, 1,826 and 2,618 at

Table 13: LARRY *in vitro* day-6 forecast from days 2 and 4, common evaluator, full gene axis, lower is better. All rows above the rule see only cells from before the target day. PRESCIENT and scNODE integrate from day 2, and the other rows start from day 4. The two rows below the rule see training clones' day-6 cells and serve only as access ablations.

Method	RMSE	Notes
Persistence	0.0874	Copies the pooled day-4 population
Source-population mean	0.0850	
Linear time	0.0886	
PCA-ridge	0.0800	
Population-time MLP	0.0600 ± 0.0016	
scNODE	0.0773 ± 0.0141	
PRESCIENT	$\mathbf{0.0519} \pm 0.0116$	Seeds 0.0415, 0.0643, 0.0498
OT-CFM	0.0882 ± 0.0068	Trained without day-6 cells
OT-CFM, original	0.0492 ± 0.0005	Trained on other clones' day-6 cells
Clone-track training mean	0.0243	Mean of other clones' day-6 cells

Table 14: Pijuan-Sala held-out mesoderm at E8.5, with a matched control on identical target cells, common evaluator, single seed. Both tasks score the same 1,162 E8.5 mesoderm cells. Gap is the relative increase in RMSE when mesoderm is absent from training. The validation split excludes mesoderm. OT-CFM, PRESCIENT and scNODE do not support the task.

Method	Held out	Matched control	Gap
Test-context copy	0.0505	0.0505	0%
Persistence	0.0637	0.0569	+11.9%
Observed-history mean	0.0832	0.0749	+11.0%
Source-population mean	0.0714	0.0565	+26.4%
Population-time MLP	0.0725	0.0563	+28.9%
Linear time	0.0768	0.0564	+36.1%
PCA-ridge	0.0766	0.0557	+37.7%

Table 15: Matched control for the LARRY held-out fates. Each fate's clones are split in half (seed 1729, stratified by starting population); both tasks score the same test half, and the control returns the other half to training (80% train, 20% validation). Gap is the relative increase in RMSE when the fate is held out; fitted rows average three seeds. The test-context copies read only the test clones' cells, so their gap is zero by construction. The test-context copy scores 0.1309, 0.1047 and 0.0639 on the three test halves.

Method	Neutrophil	Monocyte	Basophil
Test-context copy (day 4 or day 2)	0%	0%	0%
Persistence	+3.9%	+7.6%	+6.3%
Observed-history mean	+2.0%	+5.1%	+4.5%
Source-population mean	+7.3%	+13.9%	+6.0%
Linear time	+20.2%	+25.8%	+8.4%
PCA-ridge	+23.7%	+23.5%	+7.8%
Population-time MLP	+15.5%	+20.1%	+7.0%
PRESCIENT, pooled	+23.9%	+25.7%	+9.3%
PRESCIENT, cond. d2	+27.2%	+26.0%	+23.9%
PRESCIENT, cond. d4	−1.0%	+16.8%	+60.5%
scNODE, pooled	+20.3%	+28.3%	+14.2%
scNODE, cond. d2	+19.3%	+27.4%	+12.1%
scNODE, cond. d4	+8.4%	+24.6%	+10.1%
OT-CFM, cond. (no day 6)	+9.9%	+25.1%	+87.9%

Table 16: Group-conditioned, pre-target fitted models on the LARRY held-out groups, pseudobulk RMSE (↓), mean ± standard deviation over three seeds. PRESCIENT and scNODE start from the held-out clones' own day-2 (d2) or day-4 (d4) cells instead of the pooled population; OT-CFM starts from their day-4 cells and trains without day-6 cells. Training is unchanged and holds no target-day cell. The day-2 and day-4 copies copy the held-out clones' own cells; the best copy baseline is the source-population mean on LSK and the day-4 copy elsewhere. Pooled PRESCIENT is the row of Table 5. With three seeds, one seed can move a mean: PRESCIENT d4 on LSK has seeds 0.0887, 0.0523 and 0.0545 (median 0.0545; seed 1729 stopped at an essentially untrained checkpoint), and PRESCIENT d2 on neutrophils 0.1452, 0.1220 and 0.1904 (median 0.1452; seed 91 fit the seen fates of a validation set without neutrophils).

	Copy baselines			PRESCIENT	PRESCIENT, cond.		scNODE, cond.		OT-CFM, cond.
Held out	Best	Day 4	Day 2	pooled	d2	d4	d2	d4	no day 6
Neutrophil	0.1119	0.1119	0.2083	0.1281	$0.1525_{\pm 0.0348}$	$0.1089_{\pm 0.0017}$	$0.1392_{\pm 0.0027}$	$0.1251_{\pm 0.0042}$	$\mathbf{0.1085}_{\pm 0.0031}$
Monocyte	**0.0918**	**0.0918**	0.2059	0.1479	$0.1577_{\pm 0.0099}$	$0.1149_{\pm 0.0005}$	$0.1620_{\pm 0.0076}$	$0.1479_{\pm 0.0056}$	$0.1009_{\pm 0.0043}$
Basophil	**0.0657**	**0.0657**	0.1909	0.1571	$0.1642_{\pm 0.0064}$	$0.1216_{\pm 0.0003}$	$0.1726_{\pm 0.0047}$	$0.1508_{\pm 0.0061}$	$0.1595_{\pm 0.0080}$
LK	**0.0941**	**0.0941**	0.2349	0.1323	$0.1725_{\pm 0.0060}$	$0.1119_{\pm 0.0036}$	$0.1542_{\pm 0.0210}$	$0.1366_{\pm 0.0222}$	$0.1194_{\pm 0.0083}$
LSK	0.0751	0.1020	0.1651	0.0933	$\mathbf{0.0615}_{\pm 0.0039}$	$0.0652_{\pm 0.0204}$	$0.1234_{\pm 0.0127}$	$0.0784_{\pm 0.0078}$	$0.0794_{\pm 0.0082}$

Table 17: Shift magnitude on LARRY and Pijuan-Sala: pseudobulk RMSE and marginal W_1 between the test target cells and the training cells at the target time, in the evaluator's PCA space. Each LARRY shift is compared with the in-domain day-6 forecast; the Pijuan-Sala held-out lineage with its matched control on the same target cells.

Task	Shift RMSE	Shift W_1	Target cells	Training cells
LARRY in-domain day 6	0.0243	0.511	381	1,847
Held-out LK	0.0789	1.153	1,216	1,213
Held-out LSK	0.0794	1.169	1,402	1,033
Held-out monocyte	0.1352	1.729	1,017	1,410
Held-out neutrophil	0.1374	1.791	946	1,472
Held-out basophil	0.1495	1.620	655	1,620
Pijuan-Sala mesoderm, matched control	0.0504	0.612	1,162	6,797
Pijuan-Sala mesoderm, held out	0.0701	0.771	1,162	4,579

Table 18: Test and training sizes of every temporal task. Test gives test-split prediction-target cells (clones); train context gives train-split observed-context cells (clones). Context by time splits the observed-context cells of that split by time. Clones are clone barcodes on LARRY, lineage ids on the CellTag time-holdout tasks, and clone-tag strings on the CellTag cross-outcome tasks (see Appendix C); n/a marks datasets without clone structure. †For the matched-condition tasks only the test cells are recorded.

Task	Target	Test (clones)	Train ctx. (clones)	Train ctx. by time	Test ctx. by time
LARRY in vitro					
Interp.	day 4	251 (61)	2,249 (285)	2:402; 6:1,847	2:90; 6:381
Forecast	day 6 (2+4)	381 (61)	1,706 (285)	2:402; 4:1,304	2:90; 4:251
Cross-fate	neutrophil	946 (160)	1,303 (210)	2:303; 4:1,000	2:222; 4:675
Cross-fate	monocyte	1,017 (159)	1,132 (211)	2:295; 4:837	2:232; 4:779
Cross-fate	basophil	655 (88)	1,618 (271)	2:384; 4:1,234	2:124; 4:372
Start.-pop.	LK	1,216 (163)	861 (208)	2:291; 4:570	2:233; 4:1,154
Start.-pop.	LSK	1,402 (244)	1,132 (138)	2:200; 4:932	2:345; 4:672
LARRY cytokine					
Interp.	day 4	172 (58)	1,797 (267)	2:441; 6:1,356	2:94; 6:276
Forecast	day 6 (2+4)	276 (58)	1,354 (267)	2:441; 4:913	2:94; 4:172
Interp., full cocktail	day 4	12 (10)	215 (40)	2:60; 6:155	2:17; 6:33
Forecast, full cocktail	day 6 (2+4)	33 (10)	105 (40)	2:60; 4:45	2:17; 4:12
Cross-condition	full cocktail	3,827 (1,074)	6,639 (2,454)	4:6,639	2:1,878; 4:165
Schiebinger					
Interp.	day 5	7,227 (–)	224,918 (–)	37 times (1,956–12,530 cells)	–
Interp.	day 10	4,124 (–)	231,262 (–)	37 times (1,956–12,530 cells)	–
Interp.	day 15	8,368 (–)	224,055 (–)	37 times (1,956–12,530 cells)	–
Forecast	day 16	9,827 (–)	166,365 (–)	31 times (1,956–8,962 cells)	–
Forecast	day 17	8,110 (–)	166,365 (–)	31 times (1,956–8,962 cells)	–
Forecast	day 18	7,437 (–)	166,365 (–)	31 times (1,956–8,962 cells)	–
Farrell					
Interp.	4.7 hpf	1,484 (–)	36,657 (–)	10 times (200–7,094 cells)	–
Interp.	6.0 hpf	1,042 (–)	32,410 (–)	10 times (200–7,094 cells)	–
Interp.	8.0 hpf	6,323 (–)	28,792 (–)	10 times (200–7,094 cells)	–
Forecast	10 hpf	7,094 (–)	20,597 (–)	3:3:311; 3:7:200; 4:3:1,174; 4:7:1,484; 5:3:5,863; 6:1,042; 6:5:4,200; 8:6,323	–
Forecast	11 hpf	1,589 (–)	20,597 (–)	3:3:311; 3:7:200; 4:3:1,174; 4:7:1,484; 5:3:5,863; 6:1,042; 6:5:4,200; 8:6,323	–
Forecast	12 hpf	4,413 (–)	20,597 (–)	3:3:311; 3:7:200; 4:3:1,174; 4:7:1,484; 5:3:5,863; 6:1,042; 6:5:4,200; 8:6,323	–
Pijuan-Sala					
Interp.	E7.0	14,749 (–)	71,552 (–)	6.5:3,484; 6.75:2,075; 7.25:13,537; 7.5:5,074; 7.75:11,766; 8:12,202; 8.25:9,691; 8.5:13,723	–
Interp.	E7.5	10,994 (–)	82,546 (–)	6.5:2,041; 6.75:2,075; 7:11,397; 7.25:7,641; 7.75:9,867; 8:16,681; 8.25:15,935; 8.5:16,909	–
Interp.	E8.0	16,681 (–)	77,778 (–)	6.5:2,378; 6.75:2,075; 7:8,106; 7.25:7,209; 7.5:10,673; 7.75:14,493; 8.25:15,935; 8.5:16,909	–
Forecast	E8.5	16,909 (–)	44,839 (–)	6.5:3,484; 6.75:2,075; 7:14,749; 7.25:13,537; 7.5:10,994	–
Cross-lineage	E8.5 mesoderm	1,162 (–)	43,876 (–)	6.5:3,480; 6.75:1,917; 7:10,036; 7.25:5,054; 7.5:1,907; 7.75:7,386; 8:10,406; 8.25:3,690	7:493; 7.25:211; 7.75:1,443; 8.25:1,575
Matched control	E8.5 mesoderm	1,162 (–)	58,680 (–)	6.5:3,484; 6.75:2,075; 7:11,474; 7.25:6,328; 7.5:2,527; 7.75:9,867; 8:16,681; 8.25:6,244	7:493; 7.25:211; 7.75:1,443; 8.25:1,575
CellTag					
Interp.	day 9	5,809 (649)	55,241 (364)	0:11,786; 3:7,582; 6:6,657; 12:6,485; 21:8,305; 28:14,426	0:20; 3:51; 6:372; 12:3,113; 21:4,309; 28:5,943
Forecast	day 21	13,169 (576)	37,458 (261)	0:11,791; 3:7,601; 6:6,768; 9:4,670; 12:6,628	0:7; 3:25; 6:212; 9:731; 12:2,790
Forecast	day 28	20,532 (356)	37,601 (309)	0:11,790; 3:7,603; 6:6,786; 9:4,714; 12:6,708	0:4; 3:13; 6:93; 9:355; 12:1,867
Cross-fate fc.	dead end to succ.	269 (9)	833 (7)	6:1; 9:15; 12:145; 15:263; 21:409	6:3; 9:12; 12:83; 15:112
Cross-fate fc.	succ. to dead end	257 (9)	342 (8)	6:2; 9:10; 12:69; 15:86; 21:175	6:2; 9:21; 12:193; 15:294
Cross-fate int.	dead end to succ.	112 (9)	775 (7)	6:1; 9:15; 12:145; 21:409; 28:205	12:83; 21:204
Cross-fate int.	succ. to dead end	294 (9)	517 (8)	6:2; 9:10; 12:69; 21:175; 28:261	12:193; 21:511
Moon					
Interp.	day 7.5	2,951 (–)	10,522 (–)	1.5:2,789; 19.5:3,937; 25.5:3,796	–
Interp.	day 13.5	2,498 (–)	10,522 (–)	1.5:2,789; 19.5:3,937; 25.5:3,796	–
Interp.	day 19.5	2,626 (–)	11,011 (–)	1.5:2,789; 7.5:4,426; 25.5:3,796	–
LARRY in vitro, clone track					
Forecast	day 4 (2)	251 (61)	402 (285)	2:402	2:90
Cross-fate int.	neutrophil, day 4	675 (160)	1,775 (210)	2:303; 6:1,472	–
Cross-fate int.	monocyte, day 4	779 (159)	1,705 (211)	2:295; 6:1,410	–
Cross-fate int.	basophil, day 4	372 (88)	2,004 (271)	2:384; 6:1,620	–
LARRY cytokine, clone track					
Forecast	day 4 (2)	172 (58)	441 (267)	2:441	2:94
Forecast	day 6 (2)	276 (58)	441 (267)	2:441	2:94
Forecast, full cocktail	day 4 (2)	12 (10)	60 (40)	2:60	2:17
Forecast, full cocktail	day 6 (2)	33 (10)	60 (40)	2:60	2:17
Matched cond.	EPO†	177 (–)	–(–)	–	–
Matched cond.	G-CSF†	165 (–)	–(–)	–	–
Matched cond.	M-CSF†	401 (–)	–(–)	–	–

Table 19: LARRY cytokine forecasts that start from each test clone's own early cells, pseudobulk RMSE (↓). Conventions follow Table 7. A dash marks a model that was not run on the task. The full-cocktail targets hold 12 or 33 cells.

	In-domain				Full cocktail			
	Forecast			Interp.	Forecast			Interp.
Method	day 6 (2+4)	day 4 (2)	day 6 (2)	day 4	day 6 (2+4)	day 4 (2)	day 6 (2)	day 4
Persistence	0.1091	0.1897	0.2126	0.1897	0.1664	0.2035	0.2242	0.2035
Mean shift	0.0665	0.1410	0.0848	0.1410	0.1325	0.1685	0.1372	0.1685
Training mean	**0.0279**	**0.0329**	**0.0279**	**0.0329**	**0.0804**	**0.1259**	**0.0804**	0.1259
PCA-ridge	0.0482	0.0856	0.0846	0.0891	0.0930	0.1298	0.0973	0.1311
PCA-kNN	0.0526	0.1063	0.0880	0.1063	0.0939	0.1314	0.0975	0.1314
scGPT + kNN	0.0519	0.1021	0.0878	0.1021	0.1054	0.1345	0.0992	0.1345
scNODE	$0.0429_{\pm 0.0031}$	$0.0620_{\pm 0.0038}$	$0.0760_{\pm 0.0148}$	$0.0620_{\pm 0.0038}$	$0.0921_{\pm 0.0003}$	$0.1349_{\pm 0.0037}$	$0.1464_{\pm 0.0195}$	$0.1349_{\pm 0.0037}$
kNN-copy	–	–	–	–	$0.1405_{\pm 0.0037}$	$0.1590_{\pm 0.0035}$	$0.1431_{\pm 0.0012}$	$0.2002_{\pm 0.0178}$
MLP (clone)	–	–	–	–	$0.0868_{\pm 0.0004}$	$0.1260_{\pm 0.0012}$	$0.0900_{\pm 0.0016}$	**0.1227**$_{\pm 0.0006}$
Linear	–	–	–	–	0.2531	–	–	0.1596
OT-CFM (clone)	–	–	–	–	$0.0917_{\pm 0.0021}$	$0.1612_{\pm 0.0018}$	$0.1997_{\pm 0.0090}$	$0.1693_{\pm 0.0037}$

Table 20: Two further clone-level tasks, pseudobulk RMSE (↓), mean and standard deviation over three seeds. **Bold** marks the best value in each row. Top, LARRY *in vitro* interpolation of day 4 for clones of a held-out fate. OT-CFM integrates forward from the clones' day-2 cells, backward from their day-6 cells, or pools both predicted populations. Bottom, LARRY cytokine day-6 forecasts for one condition with that condition in training, the in-domain counterparts of the held-out condition task. Linear interpolates between the clones' own day-2 and day-6 cells, and every other model here learns from training clones' target-day cells.

LARRY in vitro*, cross-fate interpolation, day 4*

Held-out fate	Linear	kNN-copy	MLP (clone)	OT-CFM forward	OT-CFM backward	OT-CFM pooled
Neutrophil	0.0670	$0.1039_{\pm 0.0028}$	**0.0515**$_{\pm 0.0012}$	$0.0761_{\pm 0.0019}$	$0.0673_{\pm 0.0028}$	$0.0597_{\pm 0.0017}$
Monocyte	0.0682	$0.0926_{\pm 0.0019}$	$0.0641_{\pm 0.0047}$	$0.0826_{\pm 0.0023}$	$0.0636_{\pm 0.0025}$	**0.0543**$_{\pm 0.0025}$
Basophil	**0.0878**	$0.1253_{\pm 0.0006}$	$0.1004_{\pm 0.0016}$	$0.1175_{\pm 0.0024}$	$0.1076_{\pm 0.0019}$	$0.1000_{\pm 0.0009}$

LARRY cytokine, matched condition, day 4 → 6

Condition	kNN-copy	MLP (clone)	OT-CFM (clone)
EPO	$0.0765_{\pm 0.0060}$	**0.0572**$_{\pm 0.0005}$	$0.0825_{\pm 0.0029}$
G-CSF	$0.0690_{\pm 0.0028}$	**0.0589**$_{\pm 0.0019}$	$0.0653_{\pm 0.0010}$
M-CSF	$0.0365_{\pm 0.0015}$	$0.0265_{\pm 0.0002}$	**0.0227**$_{\pm 0.0003}$

days 2, 4 and 6), with 160 neutrophil, 159 monocyte and 88 basophil clones, and 244 LSK and 163 LK clones.

LARRY cytokine. The same system under five cytokine conditions (full cocktail, M-CSF, G-CSF, EPO and SCF) holds 65,075 cells at days 2, 4 and 6. Day 2 precedes the exposure, and all 15,728 day-2 cells are full cocktail. SCF has no day-6 cells, so it cannot serve as a forecast target. Of 5,257 clones, 2,098 (39.9%) appear under more than one condition, so a clone-level split does not by itself separate conditions. The strict held-out-condition task therefore scores only test clones that never appear in training or validation. Cytokine doses and exposure times are not recorded in the released metadata.

CellTag. Mouse embryonic fibroblasts are reprogrammed into induced endoderm progenitors and profiled by Drop-seq, 85,010 cells at days 0, 3, 6, 9, 12, 15, 21 and 28 in two replicates of 30,733 and 54,277 cells. Three sequentially delivered CellTag libraries define clones, and 22,210 cells carry a usable tag across 2,408 clones. The reprogramming fate, dead end or successful, is labeled only for clones with a unanimous annotation. The time-holdout splits keep whole lineages on one side, where a lineage groups the clones that share any tag within a replicate. The fate tasks cannot be split this way: in replicate 2, every clone of both outcomes falls in one tag-connected lineage, so a lineage-safe split leaves no training cells. They are therefore split by clone-tag string, and relatives of a test clone may appear in training; two tag-connected lineages link nearly every test clone to

Table 21: CellTag replicate transfer, single run. Linear time reproduces persistence exactly, because the task offers a single training time, where it falls back to persistence, and PCA-ridge and the MLP stay within 1% of it. Each direction trains on one of the two Drop-seq replicates and predicts the whole other replicate. **Bold** marks the best value in each column. kNN-copy, the clone-level MLP, OT-CFM, PRESCIENT, scNODE and CellTempo need observed cells of the held-out replicate to start from, which this task does not provide.

	Replicate 1 held out			Replicate 2 held out		
Method	RMSE	Pearson	W_1	RMSE	Pearson	W_1
Persistence	0.0966	0.9659	0.7072	0.1025	0.9611	**0.7459**
Observed-history mean	**0.0326**	**0.9962**	1.4246	**0.0326**	**0.9962**	1.4137
Source-population mean	0.0958	0.9664	1.5586	0.1019	0.9616	1.5475
Linear time	0.0966	0.9659	0.7072	0.1025	0.9611	**0.7459**
PCA-ridge	0.0956	0.9666	**0.6965**	0.1024	0.9611	0.7529
Population-time MLP	0.0965	0.9660	0.7070	0.1031	0.9606	0.7617

training. Dropping the test clones with a relative in training leaves 19 to 43 target cells per task and raises every copy baseline's error by 0.006 to 0.044 (for example, persistence 0.1177 to 0.1620 on the dead-end-to-successful forecast), so the relatives flatter the copies; the population-time MLP rises the same way (0.1255 to 0.1664) and still does not beat them. The reported CellTag fate numbers are therefore upper bounds on copy accuracy. The day-9 interpolation scores 5,809 target cells. The cross-fate forecasts score 269 and 257 cells. The cross-fate interpolations predict day 15 from observed days 12 and 21 and score 112 and 294 cells.

Schiebinger. The reprogramming course holds 254,197 mouse cells sampled at 39 time points, every half day from day 0 to day 18 plus days 8.25 and 8.75, without barcodes. Interpolation holds out days 5, 10 and 15 and observes the half days on both sides. Forecasting observes days up to 14 and predicts days 16, 17 and 18.

Farrell. The zebrafish course holds 39,315 cells at 12 times from 3.3 to 12 hours post fertilization (hpf). Interpolation holds out 4.7, 6.0 and 8.0 hpf. Forecasting predicts 10, 11 and 12 hpf as three separate folds, one target time each.

Pijuan-Sala. The gastrulation atlas holds 108,857 mouse cells from E6.5 to E8.5. Interpolation holds out E7.0, E7.5 and E8.0, and forecasting predicts E8.5 from stages up to E7.5. The held-out lineage task removes mesoderm from training and validation and draws validation cells from the lineages seen in training. Its control returns mesoderm cells to the training pools. Both score the same 1,162 E8.5 mesoderm cells.

Moon. The embryoid-body course holds 31,161 human cells in five windows (days 0 to 3, 6 to 9, 12 to 15, 18 to 21 and 24 to 27). Each task holds out one interior window, trains on three of the other four and validates on the fourth, with window-disjoint and cell-disjoint splits.

D MODELS AND SETTINGS

Copy baselines. Persistence predicts the empirical distribution of the cells at the latest observed time. Nearest-prior persistence uses the latest observed time before the target. Test-context persistence copies the held-out group's own latest observed cells. Neighbor interpolation weights the two observed populations around an interpolation target linearly in time. The observed-history mean predicts the cell-weighted mean over all observed times, and the source-population mean predicts the mean of the training cells at the latest observed time.

Regression baselines. Linear time fits a weighted least-squares line per gene through the per-time pseudobulk means of the training cells. PCA-ridge fits PCA on the training cells, with centering only, and regresses the per-time latent means on time with an unpenalized intercept. The number of components (5, 10, 25, 50 or 100) and the ridge penalty (0, 0.1, 1, 10 or 100) are chosen by

Table 22: Static perturbation track (Appendix C), common evaluator, single run. **Bold** marks the best value in each row. Norman has no time axis, so only the source-population mean applies there, and n/a marks the controls that condition on a target time.

Pseudobulk RMSE (↓) Task	Persist.	Obs. mean	Src. mean	Linear	PCA-ridge	MLP (pop.)
Dixit, dendritic cells, 0 → 3 h LPS	**0.1544**	**0.1544**	0.1546	0.1546	0.1546	0.1546
Dixit, held-out perturbation	0.0850	0.0850	**0.0849**	**0.0849**	**0.0849**	**0.0849**
Dixit, K562, day 7 → 13	**0.0849**	**0.0849**	**0.0849**	**0.0849**	**0.0849**	**0.0849**
Norman, no unseen gene	n/a	n/a	**0.0146**	n/a	n/a	n/a
Norman, one unseen gene	n/a	n/a	**0.0126**	n/a	n/a	n/a
W_1 *(↓)* Task	Persist.	Obs. mean	Src. mean	Linear	PCA-ridge	MLP (pop.)
Dixit, dendritic cells, 0 → 3 h LPS	**1.193**	1.510	1.511	1.511	1.511	1.511
Dixit, held-out perturbation	**0.446**	1.131	1.131	1.131	1.131	1.131
Dixit, K562, day 7 → 13	**0.729**	1.102	1.101	1.101	1.101	1.101
Norman, no unseen gene	n/a	n/a	**0.842**	n/a	n/a	n/a
Norman, one unseen gene	n/a	n/a	**0.796**	n/a	n/a	n/a

leave-one-time-out error over the training times. The population-time MLP maps standardized time to the latent mean with two hidden layers of width 256, dropout 0.1 and the AdamW optimizer at a learning rate of 10^{-3}. A task offers between 2 and 37 training times, so the network is small and its capacity is set on training data. The number of PCA components (5, 10, 25, 50 or 100) and the weight decay (10^{-4}, 10^{-3} or 10^{-2}) are chosen by the same leave-one-time-out error, and the held-out time of each fold serves for early stopping. With fewer than three training times, fixed low-capacity defaults replace this selection. All three regression baselines predict one expression profile, their fitted mean at the target time, clipped at zero.

Dynamics models. scNODE integrates from the forecast origin of each task. PRESCIENT selects its checkpoint on a validation window over the interval of the test. Its potential's initialization varies between runs with the same seed, so LARRY *in vitro* results report the mean and standard deviation over three seeds. OT-CFM uses the benchmark entry point, which supports tasks that hold out a time. On the LARRY day-6 forecast it trains on days 2 and 4 only, as the task protocol specifies. On the LARRY fate and starting-population tasks it trains on the full trajectories of the other populations.

Pretrained models. The released CellTempo checkpoint generates one cell state per step, and each of its forecasts takes a single step. A mapping from steps to days fitted on Schiebinger training and validation data was not identifiable, so CellTempo has no score there. A checkpoint was also pretrained on the scBaseTraj collection and fine-tuned on each LARRY task. MaxToki predicts rank-encoded transcriptomes, so it is scored on order-only quantities within the genes each prediction covers. A MaxToki checkpoint was also pretrained from scratch on scBaseTraj. scGPT provides frozen cell embeddings for a nearest-neighbor predictor.

Clone-level pipeline. On LARRY, a separate pipeline predicts each test clone's day-6 cells from its own early cells. Its simple controls are per-clone persistence, a per-gene mean shift, a linear trend through the test clones' own pseudobulk and the training-population mean. Its other predictors are PCA-ridge, nearest-neighbor predictors in PCA and scGPT space, scNODE, OT-CFM, a nearest-neighbor copy of training clones and a clone-paired MLP.

Seeds and applicability. The learned models run with three seeds on LARRY *in vitro* and with seed 1729 elsewhere. The evaluator fits its PCA on at most 5,000 training cells with seed 1729. A model is not applicable when its implementation does not support a task. The OT-CFM entry point supports only held-out times, scNODE and PRESCIENT do not support the held-out lineage task, and no model reads a condition label.

E AGENTIC SYSTEMS

Harness. Every agent uses DeepSeek V4.1 Flash through OpenRouter. The single agent runs in the OpenCode harness inside a sandbox that exposes only the authorized training material. It receives population-mean trajectories of the training cells, a fixed library of numerical forecasting tools with declared parameter ranges and sealed practice episodes built from training data. It submits a plan that names tools, parameters, weights and optional gene groups, and the host executes the plan with deterministic numerical code. The variant with validation feedback may score up to 12 sealed validation episodes and receives aggregate metrics only. The zero-call selector explores the same tool library and parameter grid and selects on the sealed validation episodes, with no language model. No condition ever receives test targets, test scores or residuals.

Recursive system. The recursive system follows recursive language models (Zhang et al., 2025). Its root inspects the authorized context through small programs, sends focused questions to analyst children and asks engineer children to write candidate forecasting programs, which the host scores on the practice episodes. The recursion depth is one. The control runs the same root with delegation disabled under the same budget of calls, tokens, wall time and practice submissions. Recursive runs are reported on the four LARRY tasks only.

Specialists. Parallel specialists each predict one of 73 gene modules for the same cells, and a coordinator merges the blocks. Each module is clustered from gene-gene correlations in the training data. The control assigns size-matched random gene groups.

Direct values and recall control. Asked for expression values directly, in one reply without tools, the model returned only 3,620 and 159 of the 18,791 genes in two attempts, so direct values are not scored. The recall control runs under neutral run-folder names, so a condition cannot be read from a path. Its anonymized condition renames gene symbols and fate labels and removes the dataset name. Its noise condition replaces each gene's values with noise matched to that gene's training mean and variance. The agent's messages in these runs name no dataset, fate or gene.

Prompts. The system prompts below are reproduced verbatim, with em dashes typed as double hyphens, and the working name of the benchmark is replaced by CellShift. RLM denotes the recursive language-model system, and a frozen budget is one fixed before the run. Task data, gene lists and time grids are filled in from the authorized task payload.

Single agent, no validation feedback. (`algorithm_selection.md`)

```
You are a single-cell temporal modelling agent. You receive authorized
population-mean trajectories, a fixed tool library (/workspace/toolkit.json),
and sealed practice episodes built from training data only. Choose the executed
numerical algorithm: name tools, set declared parameters within range, weight
components, and optionally partition genes into disjoint groups. Submit a final
plan as {"episode_id": "final", "plan": {...}} to
/workspace/submissions/final.json; the host executes it with deterministic
numerical code. You may not edit tool implementations. This run provides no
validation feedback of any kind: no scores, residuals, or aggregates from
withheld data will be returned. Justify your choice from the authorized
evidence alone and freeze your strategy before final delivery. Biological
annotations are not part of this ablation's authorized data.
```

Single agent with validation feedback. (`algorithm_selection_validation.md`)

```
You are a single-cell temporal modelling agent. You receive authorized
population-mean trajectories, a fixed tool library (/workspace/toolkit.json),
and sealed practice episodes built from training data only. Choose the executed
numerical algorithm: name tools, set declared parameters within range, weight
components, and optionally partition genes into disjoint groups. Submit a final
plan as {"episode_id": "final", "plan": {...}} to
/workspace/submissions/final.json; the host executes it with deterministic
numerical code. You may not edit tool implementations. You may submit
predictions for sealed validation episodes listed in
/workspace/validation/manifest.json; each episode accepts exactly one committed
```

```
prediction and returns aggregate numerical metrics only (MAE/RMSE family). No
residuals, per-gene values, or withheld futures are ever returned. Freeze your
strategy before final delivery. Biological annotations are not part of this
ablation's authorized data.
```

Flat prompt. (`flat_prompt.md`)

```
You are a single-cell temporal modelling assistant. Below you receive
authorized population-mean trajectories (rows = observed timepoints, columns =
genes, log1p CP10K) under an information-access policy. Reason about the
biology of the observed window, then state expression values directly for the
held-out horizon timepoints. This is a single reply with no tools, no files, no
code execution, and no further rounds: nothing will be returned to you. Values
must be finite and nonnegative. Biological annotations are not part of this
condition's authorized data.
```

Recursive system, root. (`root_core.md`)

```
# RLM root -- role and constraints

You are the root of a bounded recursive forecasting agent. You have no direct
access to the raw expression data: you never read a data file, and you never
see the expression of the cells you are asked to predict. Hidden-test
prediction-target expression is evaluator-only and is not in your context.

What you do have is an **external context interface**: a JSON-able context
digest of authorized training material (task summary, gene-space head,
per-time training pseudobulk, allowed metadata counts, validation aggregates
when the run's access policy permits them) and a bounded code sandbox. You
examine the context **programmatically**: you write small programs that
compute aggregates over the authorized context and return compact JSON. You
do not guess numbers; every number you reason about comes from a program you
wrote and from the reports your children return.

## Your tools

- `inspect(context)` -- you write a program defining `run(context)` that
  returns a compact JSON summary. Output size is capped. Set `"continue":
  false` in the reply when you have seen enough. This is the defining
  behavior of this agent: programmatic external-context inspection, not
  reading raw context.
- **Delegation** -- you dispatch focused analyst/engineer briefs. Children
  cannot delegate further; you are the only node that can recurse, and the
  recursion is capped at depth one.

## Your two kinds of child

- `analyst` -- receives one focused question plus the named context slices,
  measures what the question needs, and returns JSON findings with measured
  numbers. It does not write forecasting code and submits nothing.
- `engineer` -- receives the analyst findings it needs and authors a bounded
  numerical candidate: a program defining `forecast(context)` returning
  either a gene-pseudobulk vector (practice episodes) or a cell-by-gene
  matrix (the final population). The host executes it against sealed
  practice episodes and reports aggregate scores back.

## The loop you are here to run

1. **Inspect** -- one or a few bounded programs to measure the context
   (times observed, pseudobulk shapes, variability, metadata counts).
2. **Decompose** -- split the problem into focused residual/gene-program
   questions. Each brief names its grounding context slices.
3. **Dispatch** -- send analyst briefs and then engineer briefs grounded in
   the analyst findings. Keep each brief short and specific: name the
   slices, name the quantities, and stop.
4. **Integrate** -- choose one declared, host-executed rule over the scored
   candidates: `select_best_practice`, `first`, or `weighted_average` with
   explicit weights keyed by candidate_id.
```

```
5. **Freeze** -- the strategy freezes once. Validation-driven revision, when
   your ablation permits it, works on aggregate feedback only and within the
   same frozen budget. Hidden-test results are never yours to see.

Report failures and empty findings; do not silently retry past your budget.
A later delivery replaces an earlier one, so bank an acceptable candidate
early before improving on it.
```

Recursive system, delegation disabled. (`nonrecursive.md`)

```
# Nonrecursive control -- same material, no delegation

You run the identical task, context digest, sandbox, sealed-practice
protocol, and frozen budget as the RLM root, but **delegation is disabled**.
You are the paired control that isolates what recursion adds.

What stays the same: the root prompt material above, the programmatic
context inspection (you write bounded `run(context)` programs), the
engineer's candidate contract, the host-executed scoring, the integration
rules, and the freeze-once semantics with the same total budget.

What is removed: analyst/engineer child calls. There is no decomposition
tree beyond the root, and any attempt to dispatch a child brief is refused
by the runtime and recorded. You author your candidate inline, in the
same format the engineer child would have used:

'''json
{"program": "<python source defining forecast(context)>"}
'''

The comparison against the recursive rung is budget-matched: identical caps
on LLM calls, tokens, wall time, sandbox calls, and practice submissions.
```

Specialist. (`specialist_system.md`)

```
# Specialist agent system prompt (parallel gene-group/pathway agents)

You are one parallel gene-group specialist in the CellShift benchmark. Your
scope is a single frozen gene group; other specialists handle the remaining
genes in parallel against the **same shared population of latent particles**.

## Inputs you may read

- The shared population object: particle ids and their latent coordinates
  (training-fitted PCA). These particle identities are fixed for the whole
  panel; every specialist writes predictions onto exactly these ids, in order.
- Your group's gene list (frozen before any target access; see the group-set
  fingerprint in your task card).
- Training-data statistics for your genes (means, variances, correlations
  within the training set).
- In the `shared_context` ablation only: one aggregate global-context vector
  shared identically by every specialist (training pseudobulk module means).
  It contains no target or future information.

## Rules

1. Never request or inspect hidden-test prediction-target cells. Any tool or
   read that would expose them is forbidden and will be recorded as a failure.
2. Predict **group-specific gene quantities** for every particle of the shared
   population. Do not sample your own cells; do not propose a different
   population. Independent cell sampling destroys cross-gene covariance and
   will be rejected by the coordinator.
3. Keep all model fitting on training data. If a validation aggregate is
   provided for weight selection, use it only through the declared interface.
4. Report your per-gene predictions as a matrix aligned to the shared particle
   ids plus a short trace: what you read, what you fit, and your confidence.
5. If your group's genes cannot be predicted from the shared population
   (insufficient training signal, zero variance), return the training mean
```

```
   profile for your genes and record 'fallback=train_mean' in the trace rather
   than fabricating structure.

## Output contract

Per specialist:

- 'group_id', 'genes' (order must match the frozen group),
- 'values' with one row per shared particle id, same order,
- 'trace' with accesses, backend, and fit statistics,
- resource usage (calls, tokens, wall time) for the shared budget report.

The host-side coordinator performs the fixed merge (reconciliation policy with
validation-selected weights). You do not merge; you do not see other
specialists' outputs except in the 'shared_context' ablation's aggregate form.
```

F SUPPLEMENTARY ANALYSES

The analyses in this section use saved predictions and held-out targets after scoring. They help explain failures but create no benchmark rows, and a flow step rescaled using the true target serves only as a post-hoc diagnostic.

F.1 SIMPLE CONTROLS AND TIME DIRECTION

Interpolation controls. Plain persistence chooses the latest context time even when that time falls after an interpolation target, so on interpolation tasks it copies a later population. Two copy baselines with fixed context rules address this. Nearest-prior persistence copies the closest earlier time, and neighbor interpolation weights the two observed populations around the target linearly in time. Neither has fitted parameters, and neither reads the target. Neighbor interpolation is the best method on 9 of 16 in-domain interpolation tasks (Table 6), with Pijuan-Sala the clearest exception.

Forecast horizons. In Schiebinger and Pijuan-Sala, the most recent observed change in expression is only a weak guide to the later change, and linear extrapolation can perform worse than copying the latest population. Learned dynamics err in both directions. The LARRY day-6 OT-CFM run trained with day-6 cells moves cells too little. On Schiebinger, integrating beyond the training interval carries cells past the target, and the OT-CFM error grows from 0.119 to 0.204 between days 16 and 18. On Farrell, OT-CFM with a cap selected on a held-out training time is the weakest fitted forecaster (0.21 to 0.38 against 0.06 to 0.11 for the MLP), because it integrates from the latest permitted time, which lies in the segment held out for selection; its interpolations improve (0.0285 at 8.0 hpf). The Pijuan-Sala E8.5 OT-CFM forecast diverges (0.7901). In the LARRY OT-CFM runs trained without day-6 cells, 36 to 66 of 251 predicted cells leave the observed data range.

F.2 FATE AND POPULATION ERRORS

Held-out fate. The population-level cross-fate persistence rows all share the same pooled day-4 source, and using each test clone's own day-4 cells instead changes the apparent cross-fate gap. Figure 4 shows that PRESCIENT, scNODE and the MLP predict the marker levels of the training fates. Much of the error is concentrated in fate-marker genes, and the models show similar patterns of cross-fate gene error. Table 23 shows why kNN and fitted methods fail most visibly on held-out basophils. A reference classifier labels each predicted cell. Copying the held-out clones' early cells, with or without a mean shift, keeps 77 to 83% of predicted cells in the basophil class, because these clones already look basophil at day 4. Methods that learn from the training fates keep 0 to 28% and place the rest in the training fates. They route monocytes and neutrophils the same way, but copying is weaker there. Their RMSE is 2.0 to 2.3 times that of copying on basophils, against 0.8 to 1.6 times on the other fates (Table 7). The basophil result therefore reflects how far each model moves cells from their early state. It does not show that basophil biology is intrinsically harder to predict.

Clone-track comparisons. The LARRY clone-level methods use training clones' cells at the target day (Table 7). Their matched control has access to same-fate target-day training cells, whereas

Table 23: Where held-out clones go on LARRY *in vitro* (setting of Table 7). The 15-nearest-neighbor reference classifier of Section 3.3 labels each predicted day-6 cell; the test-context copy is per-clone persistence. Own gives the share of cells labeled as the held-out fate. TV gives the fate-proportion error (↓).

	Neutrophil held out		Monocyte held out		Basophil held out	
Model	Own (%)	TV	Own (%)	TV	Own (%)	TV
Test-context copy	55	0.41	50	0.43	77	0.19
Mean shift	**59**	**0.37**	**51**	**0.42**	**83**	**0.13**
PCA-ridge	1	0.95	4	0.89	28	0.66
PCA-kNN	0	0.98	4	0.89	5	0.91
scGPT + kNN	0	0.98	2	0.91	26	0.69
scNODE	0	0.96	0	0.93	0	0.96
Training-population mean	0	0.99	0	0.94	0	0.99
Observed day-6 cells	95		93		95	

the cross-fate arm sees only target-day cells from other fates. This setting is thus a distinct scientific task from a forecast trained only on days before the target. Within the clone track, matched expression errors can be compared. Wasserstein distances cannot be differenced, because the control refits PCA on a different training pool. Moreover, several clone-level rows labeled as interpolation use only pre-target cells. They repeat the corresponding forward forecasts and do not test whether a model exploits observations on both sides of a missing day.

Diversity and composition. Mean predictors output point masses, and some models match the average while missing held-out cell types (Table 23), which is why cell-type proportions are reported alongside pseudobulk error.

F.3 Metrics and rankings

Expression RMSE and PCA Wasserstein distance need not rank models in the same order. Copying observed cells can preserve a realistic point cloud while missing the change in mean, and a low mean error can coexist with wrong fate proportions. On LARRY *in vitro*, pooled persistence has a lower PCA Wasserstein distance than every fitted model in the population-level table, except OT-CFM on the three held-out fates, which trains on the other fates' target-day cells. MaxToki is scored with one rank convention and permutation nulls. Under this metric, none of the 41 change-based cells of the released checkpoint's seven valid predictions is above chance. A checkpoint pretrained from scratch on scBaseTraj passes 5 of 106 such cells, about the rate chance gives at this threshold, and no level cell of either checkpoint beats copying the reference population (Table 24). Lastly, interpolation rankings are more similar across datasets than forecast rankings, but there are too few comparable forecast tasks to support a general claim about ranking transfer.

F.4 Agents, representations and uncertainty

The zero-call selector and single agents often choose the same trend, and several runs with validation feedback do not use it. The recursive agents build their context from training clones' target-day cells, and their best day-6 runs reproduce the mean of those cells. On held-out fates they are worse than the selector. Specialist agents assigned biological gene groups have not shown a clear advantage over matched random groups. The recall control shows that anonymizing the inputs leaves accuracy unchanged, while noise inputs degrade it. The agents therefore use the data and do not recall a published answer. Outside the four LARRY *in vitro* tasks, the two single-agent conditions ran once per task. Both beat their selector on the CellTag day-21 forecast (0.0957 and 0.1018 against 0.1077) and on the strict cytokine condition task (0.1070 and 0.1044 against 0.1071). These single runs are not interpreted as gains. Within the target-day-supervised clone track, frozen scGPT neighbors show no consistent advantage over PCA neighbors and tend to fare worse on the shifted LARRY tasks. This comparison does not test scGPT as a temporal generator.

Table 24: MaxToki under rank metrics with permutation nulls, one prediction per row, seed 1729, greedy decoding from three context cells. Change cells compare the genes that change most from the reference population, top 50, 100 and 300 in two gene pools. Level cells compare the most strongly ranked genes the same way, and pathway cells the top enriched pathways. Each count gives the cells above a 1,000-permutation null (one-sided $p < 0.05$, no correction for multiple comparisons) out of the cells defined for that prediction. Level $>$ copy counts the level cells that beat copying the reference population. Schiebinger forecasts of the released checkpoint without a length cap and CellTag predictions on earlier splits are not listed.

Dataset	Task	Ranked genes	Change	Level	Level $>$ copy	Pathway
Released checkpoint						
LARRY *in vitro*	Forecast day 6 (2+4)	123	0/5	2/4	0/4	0/1
Schiebinger	Interp. day 5	482	0/6	1/6	0/6	0/1
Schiebinger	Interp. day 10	471	0/6	1/6	0/6	0/1
Schiebinger	Interp. day 15	520	0/6	1/6	0/6	0/1
Schiebinger	Forecast day 16	673	0/6	0/6	0/6	0/1
Schiebinger	Forecast day 17	674	0/6	0/6	0/6	0/1
Schiebinger	Forecast day 18	644	0/6	3/6	0/6	0/1
Pretrained from scratch on scBaseTraj						
LARRY *in vitro*	Forecast day 6 (2+4)	935	0/6	1/6	0/6	0/1
LARRY *in vitro*	Forecast day 4 (2)	817	0/6	1/6	0/6	0/1
LARRY *in vitro*	Interp. day 4	433	0/6	1/6	0/6	0/1
LARRY *in vitro*	Cross-fate int. neutrophil	138	0/5	1/4	0/4	0/1
LARRY *in vitro*	Cross-fate int. monocyte	110	0/5	2/4	0/4	0/1
LARRY *in vitro*	Cross-fate int. basophil	163	0/5	2/4	0/4	0/1
LARRY *in vitro*	Matched control neutrophil	91	1/4	1/2	0/2	0/1
LARRY *in vitro*	Matched control monocyte	940	0/6	0/6	0/6	0/1
LARRY *in vitro*	Matched control basophil	922	0/6	0/6	0/6	0/1
LARRY cytokine	Interp. day 4, full cocktail	304	0/6	3/6	0/6	1/1
LARRY cytokine	Forecast day 4 (2), full cocktail	706	1/6	6/6	0/6	0/1
LARRY cytokine	Forecast day 6 (2+4), full cocktail	81	0/4	2/2	0/2	0/1
LARRY cytokine	Forecast day 6 (2), full cocktail	329	0/6	2/6	0/6	0/1
Schiebinger	Interp. day 5	876	0/6	2/6	0/6	0/1
Schiebinger	Interp. day 10	906	0/6	1/6	0/6	1/1
Schiebinger	Interp. day 15	890	0/6	0/6	0/6	0/1
Schiebinger	Forecast day 16	256	0/5	1/4	0/4	0/1
Schiebinger	Forecast day 17	704	0/6	1/6	0/6	0/1
Schiebinger	Forecast day 18	533	3/6	5/6	0/6	0/1

Clone resampling makes several small gaps between models uncertain (Appendix F.5), PRESCIENT seeds vary by up to 70% (Table 16), and the full-cocktail cytokine targets hold 12 or 33 cells. Table 18 gives the target and clone counts of every task.

F.5 Bootstrap intervals

Intervals are computed from the saved predictions without refitting any model: bias-corrected and accelerated (BCa) 95% intervals from 1,000 resamples (seed 1729) of test clones on LARRY and CellTag and of test cells elsewhere, with the evaluator's frozen PCA. A ratio between a fitted model and a copy uses the same resample for both, and a win requires the whole interval below 1. Of the 74 pairs computed, 20 are wins, 5 ties and 49 losses. The counts in the main text use the 5% margin rule, which applies uniformly to every task. On held-out neutrophils, PRESCIENT and scNODE started from day-4 cells and OT-CFM trained without day-6 cells tie the day-4 copy (e.g. PRESCIENT 0.973, interval 0.918 to 1.206); on monocytes, basophils and LK every such model loses. On LSK, PRESCIENT (from day-2 or day-4 cells), scNODE from day-4 cells and OT-CFM beat the day-4 copy (PRESCIENT 0.639, 0.594 to 0.669), and on the in-domain day-6 forecast every conditioned variant beats it.

F.6 PROTOCOL REVISIONS

Before the reported results were scored, the task construction and model selection were revised as follows, and every reported number comes from the revised version. CellTag time-holdout splits separate whole lineages instead of tag strings. The validation cells of the Pijuan-Sala held-out-mesoderm task come from the lineages seen in training. The Farrell forecast predicts each late stage as a separate target. The cytokine held-out-condition task scores only clones absent from training and validation. PRESCIENT validates over the interval of the test, and OT-CFM selects its step cap on validation loss, or on a held-out training time where no validation split exists (CellTag, Schiebinger, Farrell, Moon and Pijuan-Sala E8.5). The released CellTempo checkpoint is loaded with its released tokenizer.

G EXCLUDED MODELS

Waddington-OT, the canonical method for the Schiebinger data, has an adapter and a synthetic self-test but was not run on the benchmark tasks. Neither it nor LineageOT receives future populations in a forecasting experiment. The other exclusions fall into four groups. Several share a formulation with a model already in the benchmark and were deprioritized: MIOFlow and TrajectoryNet learn continuous flows like OT-CFM, PLS regression and gradient boosting are regression baselines like PCA-ridge, and Geneformer is a second frozen-embedding model like scGPT. GRU and temporal-transformer sequence models, iTransformer and displacement networks are architecture variants, and JKOnet*, scIMF and CellPace were released after the model set of the benchmark was fixed. Several fall outside the definitions of the benchmark. D-SPIN models equilibrium response distributions, LineageOT is a lineage-informed diagnostic, RegVelo-style GRN-ODEs need spliced and unspliced counts that are not wired in, and CellFlow targets perturbations. GEARS, CPA, PerturbGen, State, Tahoe-100M-SCVI and the scGPT perturbation head were not adapted to the temporal tasks; CPA in particular could read the cytokine condition and is the most direct extension of the condition task. scDiffEq, DeepRUOT and TIGON or WFR-FM were not implemented.