FORECASTING CELLULAR AND POPULATION DYNAMICS VIA MULTI-HORIZON FLOW MATCHING

Anonymous authors
Paper under double-blind review

ABSTRACT

Forecasting cellular populations from time-resolved single-cell RNA-seq requires predicting how cell states evolve and how these state-level changes reshape the population over time. We propose scHOMURA, a generative method for forecasting cellular and population dynamics from unpaired temporal snapshots. scHOMURA combines deterministic temporal displacement to capture shared state progression, conditional residual dynamics to model heterogeneous cellular outcomes, and a state-dependent growth field to model changes in relative population contribution. By directly learning transitions over multiple forecast horizons, scHOMURA reduces reliance on recursive one-step propagation and supports forecasting over both short and extended temporal intervals. Across five time-resolved single-cell datasets and seven forecasting splits, scHOMURA achieves strong distribution-, gene-expression-, and composition-level performance and remains stable as predictions extend beyond the observed temporal window. Beyond global population reconstruction, scHOMURA captures rare-state expansion, reaches future cellular states absent from the training period, and recovers developmental transitions consistent with known biological relationships. Together, scHOMURA provides a multi-horizon generative framework for forecasting cell-state dynamics and population remodeling from unpaired temporal snapshots.

1 INTRODUCTION

Time-series single-cell RNA sequencing (scRNA-seq) captures cellular populations at multiple stages of dynamic biological processes, providing snapshots of how cellular systems evolve over time (Ding et al., 2022; Liu et al., 2024). Forecasting future cellular and population dynamics from these data requires modeling how individual cell states change and how these state-level changes collectively give rise to population-level remodeling. This is fundamentally challenging because scRNA-seq measurements are destructive: cells observed at different time points are unpaired, and future population dynamics must therefore be inferred from a sequence of independent snapshots.

Existing approaches have advanced temporal single-cell modeling, from trajectory reconstruction and distribution transport to generative prediction and population-growth modeling (Schiebinger et al., 2019; Tong et al., 2020; Huguet et al., 2022; Yeo et al., 2021; Zhang et al., 2024; Sha et al., 2024; Wang et al., 2026). However, most methods are optimized for matching the overall cellular distribution at unobserved time points, leaving two challenges less explored: maintaining accuracy over extended forecast horizons and anticipating rare or emerging populations. Beyond **future-population forecasting**, these settings require **extended-horizon stability** and reliable prediction of **rare and emerging populations** that are weakly represented or absent in observed data.

These settings expose two limitations of conventional one-step recursive forecasting. First, recursive propagation compounds transition errors over time, motivating direct supervision across multiple forecast horizons. Second, state transition alone is insufficient to model rare and emerging populations, because their future prevalence depends not only on how cells evolve, but also on how strongly different source states contribute to the future population. These considerations motivate a forecasting model that couples multi-horizon transition learning with explicit population remodeling.

To this end, we propose **scHOMURA** (Single-Cell Heterogeneous Outcome Modeling from Unpaired Multi-horizon Residual-Abundance Dynamics), a generative model for forecasting cellular and population dynamics from unpaired temporal snapshots (Fig. 1). scHOMURA jointly models

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

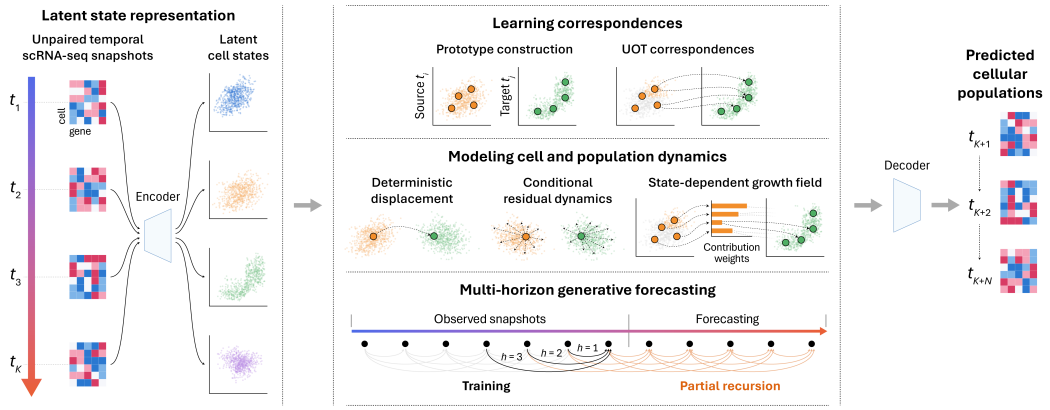


Figure 1: Overview of scHOMURA. Unpaired temporal scRNA-seq snapshots are encoded into a shared latent space, where prototype-level unbalanced optimal transport (UOT) establishes correspondences between source and target populations. scHOMURA then models cell and population dynamics through deterministic displacement, conditional residual dynamics, and a state-dependent growth field. Direct multi-horizon training learns transitions over multiple snapshot intervals, while partial recursion extends forecasting beyond the observed range. Predicted latent populations are decoded into future single-cell expression profiles.

cell-state progression and population remodeling: deterministic displacement captures shared temporal change, conditional residual dynamics capture heterogeneous outcomes, and a state-dependent growth field models changes in relative population contribution. By learning transitions directly over multiple forecast horizons, scHOMURA reduces reliance on recursive one-step propagation and supports forecasting over both short and extended temporal intervals.

To evaluate these capabilities, we consider three forecasting settings: future-population forecasting, extended-horizon forecasting, and rare and emerging population forecasting. Across five time-resolved single-cell datasets and seven splits, scHOMURA performs strongly at the distribution, gene-expression, and composition levels, remains stable beyond the observed temporal window, and captures both rare-state expansion and future states absent from training. Targeted analyses further separate forecasting into distinct capabilities: identifying populations that will expand, calibrating their future abundance, and recovering cellular states that emerge later. scHOMURA also recovers developmental transitions consistent with known biological relationships, indicating that its learned dynamics preserve meaningful temporal structure. Our main contributions are:

- We extend future single-cell population forecasting beyond standard future-population prediction to two more challenging settings: extended-horizon forecasting and rare and emerging population forecasting.
- We propose scHOMURA, a generative model that jointly models shared temporal progression, heterogeneous cellular outcomes, and state-dependent population contributions while directly learning transitions across multiple horizons from unpaired temporal snapshots.
- We demonstrate that extended-horizon stability, population expansion, and future-state recovery are distinct forecasting capabilities, and show that scHOMURA can capture them within a unified framework.

2 Related Work

Learning cellular dynamics from single-cell snapshots. Time-resolved scRNA-seq provides population snapshots rather than longitudinal measurements of individual cells, motivating methods that infer dynamics from unpaired observations. Early work focused on trajectory reconstruction using pseudotime, optimal transport, and continuous-flow models (Lönnberg et al., 2017; Schiebinger et al., 2019; Tong et al., 2020), while more recent approaches extend toward interpolation and prediction at unobserved or future time points (Yeo et al., 2021; Huguet et al., 2022; Zhang et al., 2024;

Jiang & Wan, 2024; Balik & Lähdesmäki, 2026). These developments have shifted temporal single-cell modeling from retrospective reconstruction toward prospective prediction, but still primarily emphasize state-space evolution rather than joint state and population dynamics.

Population mass variation and unbalanced transport. Cellular populations change in relative abundance through proliferation, death, and transitions between states, violating the mass-conservation assumption of standard OT. Unbalanced transport methods address this by jointly modeling state transitions and population mass variation (Sha et al., 2024; Wang et al., 2026). These approaches extend temporal single-cell modeling beyond state transport alone, but typically treat population change as an auxiliary mass term rather than explicitly forecasting how state-level dynamics reshape future population composition.

Flow matching for temporal cellular dynamics. Flow matching provides a flexible framework for learning transport between probability distributions (Lipman et al., 2022). Recent single-cell methods apply flow-based dynamics to temporal state transitions and population mass variation (Wang et al., 2026; Pu et al., 2026). scHOMURA further extends this framework to multi-horizon forecasting of heterogeneous cellular and population dynamics.

3 METHOD

3.1 OVERVIEW

Let $X_k = \{x_i^{(k)}\}_{i=1}^{N_k}$ denote the single-cell population observed at snapshot k, where $x_i^{(k)}$ is the gene-expression profile of cell i. Given an ordered sequence of snapshots, our goal is to forecast the cellular-state distribution at a future snapshot $k' = k + h$. Because scRNA-seq is destructive, cells are unpaired across time, and the dynamics must be inferred from population snapshots. Forecasting therefore requires modeling how cells move through state space, how their outcomes diversify, and how source states differentially contribute to the future population.

scHOMURA models future population evolution through three complementary processes: shared temporal progression, heterogeneous outcomes among cells with similar starting states, and state-dependent changes in population contribution, modeled by a deterministic displacement network μ_θ, a conditional residual flow Φ_ϕ, and a state-dependent growth field r_ω, respectively (Fig. 1).

Given a source latent population $\{z_i\}$ at snapshot index k, scHOMURA generates a future cell at index $k' = k + h$ by first sampling a source state according to a growth-weighted distribution and then applying a stochastic state transition:

$$i_n \sim q_\omega^{(h)}, \qquad \hat{z}_n = z_{i_n} + \mu_\theta\left(z_{i_n}, \bar{k}, \bar{h}\right) + \sigma\Phi_\phi\left(\eta_n \mid z_{i_n}, \bar{k}, \bar{h}\right), \qquad \eta_n \sim \mathcal{N}(0, I).$$

Let k and k' denote the ordinal indices of the source and target snapshots. We define

$$h = k' - k, \qquad \bar{k} = k/K_{\text{last}}, \qquad \bar{h} = h/K_{\text{last}},$$

where K_{last} is the ordinal index of the final training snapshot. $\bar{k}$ and $\bar{h}$ can preserve relative sampling time and elapsed time, and for stage-based data, these coordinates represent developmental order and the number of snapshot steps rather than calibrated physical time.

The distribution $q_\omega^{(h)}$ is the source-sampling distribution induced by the state- and snapshot-dependent growth field. This formulation separates three aspects of future evolution: where cells move in state space, how heterogeneous their future outcomes are, and how strongly different source states contribute to the future population.

3.2 SCHOMURA ARCHITECTURE

Latent state representation. We first map gene-expression profiles into a low-dimensional latent space. Specifically, a deterministic autoencoder encodes each expression profile x as

$$z = e(x), \qquad \hat{x} = d(z),$$

where e and d denote the encoder and decoder, respectively. The autoencoder is trained using only the observed training snapshots and is then fixed, so that all subsequent temporal modeling is performed in a shared latent space.

State transition model. Given a source latent state z, scHOMURA models its future state as

$$z' = z + \mu_\theta(z, \bar{k}, \bar{h}) + \sigma\Phi_\phi(\eta \mid z, \bar{k}, \bar{h}).$$

where μ_θ represents the shared temporal displacement and Φ_ϕ denotes a conditional residual flow learned by flow matching (Lipman et al., 2022). The scalar σ learned jointly with the transition model controls the global magnitude of the residual variation.

Both components are conditioned on the source state, normalized source-snapshot index, and normalized ordinal horizon. A single set of transition-network parameters is shared across all directly supervised horizons. The deterministic displacement captures the common direction of state progression, whereas the residual flow models heterogeneous deviations around this progression, allowing similar source states to evolve toward distinct future states.

State-dependent growth field. State transition alone does not determine future population composition, because different source states may contribute unequally to the future population. scHOMURA therefore introduces a state- and snapshot-dependent growth field $r_\omega(z, \bar{k})$ to model the relative contribution of each source state:

$$q_i^{(h)} = \frac{\exp\big(h\, r_\omega(z_i, \bar{k})\big)}{\sum_j \exp\big(h\, r_\omega(z_j, \bar{k})\big)}.$$

Here, $q_i^{(h)}$ defines the probability of selecting source state z_i for an h-snapshot forecast. Each selected source state is then propagated through the stochastic transition model to produce a future cell. Thus, r_ω captures relative changes in population contribution rather than absolute cell numbers or mechanistic proliferation and death rates.

3.3 Learning from Unpaired Temporal Snapshots

Because cells are unpaired across time, scHOMURA learns temporal dynamics from distributional correspondences between population snapshots. We use unbalanced optimal transport (UOT) (Chizat et al., 2018) to construct latent pseudo-transitions, while state transition, population contribution, and cross-horizon consistency are learned with separate objectives.

Transition supervision. For each observed source–target pair, we first summarize each latent population by a set of local prototypes, each representing a cluster of nearby cells. UOT is then used to establish correspondences between source and target prototypes. Cell-level pseudo-transitions (z, z') are then sampled from cells associated with coupled source and target prototypes, defining an observed latent displacement

$$v^\star = z' - z.$$

We decompose this displacement into a shared component and a residual component. The deterministic displacement network is trained to capture the shared temporal progression:

$$\mathcal{L}_{\text{mean}} = \mathbb{E}\left[w^{(\text{under})} \left\|\mu_\theta(z, \bar{k}, \bar{h}) - v^\star\right\|_2^2\right].$$

The remaining variation is represented by the normalized residual

$$\varepsilon^\star = \big[v^\star - \text{sg}\big(\mu_\theta(z, \bar{k}, \bar{h})\big)\big]/\sigma,$$

where sg denotes stop-gradient and $w^{(\text{under})}$ is the undercoverage-aware transition weight defined in Appendix A.1. A conditional flow-matching objective is then used to learn the distribution of $\varepsilon^\star$ conditioned on the source state, normalized ordinal source coordinate, and normalized snapshot horizon, enabling scHOMURA to capture heterogeneous deviations around the shared temporal progression.

Growth supervision. Population remodeling is supervised by separating state transition from changes in source-state contribution. Let a_p and c_p denote the empirical mass and center of source prototype p, respectively. The growth field reweights the contribution of each source prototype as

$$q_p^{(h)} = \frac{a_p \exp\big(h\, r_\omega(c_p, \bar{k})\big)}{\sum_k a_k \exp\big(h\, r_\omega(c_k, \bar{k})\big)}.$$

Let P denote the prototype transition matrix induced by the UOT coupling. While P specifies where source states transition, $q^{(h)}$ determines how strongly they contribute to the future population. The resulting target composition $\hat{b} = q^{(h)}P$ is matched to the observed target composition b using

$$\mathcal{L}_{\text{growth}} = \text{JS}(\hat{b}, b) + \mathcal{R}_{\text{growth}},$$

where $\mathcal{R}_{\text{growth}}$ regularizes large deviations from the empirical source composition and encourages temporal smoothness of the growth field.

Cross-horizon consistency regularization. In addition to transition-level supervision, we align the generated and observed target populations in latent space using an RBF-MMD loss $\mathcal{L}_{\text{MMD}}$ (Gretton et al., 2012). We further encourage consistency across forecast horizons. Specifically, the same future population can be reached either through a direct two-step prediction or by composing two one-step transitions; these two forecasts should therefore agree distributionally. We define

$$\mathcal{L}_{\text{cons}} = \text{MMD}^2_{\text{RBF}}\left(\hat{Z}^{\text{direct}}_{t\to t+2}, \hat{Z}^{\text{composed}}_{t\to t+1\to t+2}\right).$$

The complete training objective combines residual flow matching, deterministic transition supervision, population-level distribution matching, cross-horizon consistency, and growth supervision:

$$\mathcal{L} = \mathcal{L}_{\text{flow}} + \lambda_{\text{mean}}\mathcal{L}_{\text{mean}} + \lambda_{\text{MMD}}\mathcal{L}_{\text{MMD}} + \lambda_{\text{cons}}\mathcal{L}_{\text{cons}} + \mathcal{L}_{\text{growth}}.$$

3.4 Generative Forecasting

At inference, scHOMURA generates a future population by first sampling source states according to the growth-weighted distribution $q^{(h)}$ and then propagating them through the stochastic transition model. Each generated cell is therefore determined jointly by its source-state contribution and its conditional future-state distribution.

scHOMURA is trained directly on every available snapshot horizon $h \in \mathcal{H}$, with a maximum direct horizon of three snapshots. All horizons share the same transition networks and are distinguished through the normalized ordinal-horizon condition $\bar{h}$. To forecast a target snapshot $k^\star$, the model uses all available source populations within the trained horizon range: an h-step forecast is generated from the population at $k^\star - h$. Candidate forecasts for $h = 1, 2, 3$ are combined using fixed base weights 0.6, 0.3, and 0.1, after unavailable horizons are removed and the remaining weights are renormalized. When $k^\star$ extends beyond the observed range, previously generated populations are added to the available history and can serve as sources for subsequent predictions. Extended-horizon forecasting therefore combines direct multi-horizon transitions with partial recursion, reducing dependence on repeated one-step propagation.

Details of the scHOMURA model, including prototype construction, UOT settings, regularization terms, loss weights, and other training settings, are provided in Appendix A.1.

4 Experiments

4.1 Experimental Setup

We evaluate scHOMURA on five time-resolved single-cell RNA-seq datasets spanning mouse cardiac development (CD) (Feng et al., 2022), cellular reprogramming (CR) (Schiebinger et al., 2019), zebrafish embryogenesis (Farrell et al., 2018), *in vitro* beta-cell differentiation (BetaDiff) (Veres et al., 2019), and retinal organoid development (Wahle et al., 2023). These datasets yield seven forecasting splits covering diverse developmental, reprogramming, and differentiation processes. For both CD and CR, we include a shorter-range and a longer-range forecasting setting to evaluate performance across different temporal horizons (Table S1).

We compare scHOMURA with five representative methods for temporal single-cell population modeling: LGP-OT (Balik & Lähdesmäki, 2026), scNODE (Zhang et al., 2024), MIOFlow (Huguet et al., 2022), PRESCIENT (Yeo et al., 2021), and VGFM (Wang et al., 2026). All methods use the same train–test splits, expression representation, generated-cell budget, and evaluation pipeline to ensure a consistent comparison.

We assess forecasting performance at the distribution, gene-expression, and population-composition levels. Distributional agreement is measured using empirical 2-Wasserstein distance (W_2) and maximum mean discrepancy (MMD). Gene-level dynamics are evaluated by differential-expression average precision (DE-AP), direction agreement (DE-dir), Pearson correlation of mean-expression changes (Δ Pearson), and variogram error. For datasets with cell-type annotations, we additionally measure Jensen–Shannon divergence between observed and predicted cell-type proportions (Type JS). Lower values indicate better performance for W_2, MMD, variogram error, and Type JS, whereas higher values are better for DE-AP, DE-dir, and Δ Pearson. Detailed datasets, metric definitions and implementation settings are provided in Appendix A.2.

4.2 Future-Population Forecasting

We first evaluate scHOMURA under the standard future-population forecasting setting, where models are trained on observed snapshots and evaluated on held-out future populations. Performance is aggregated across held-out time points to assess how accurately each method predicts future cellular distributions, transcriptional changes, and population composition.

Across all seven splits, scHOMURA achieves the lowest W_2, indicating consistently accurate future-population forecasts (representative splits in Table 1; remaining splits in Table S2; aggregate rankings in Table S3). On the four annotated splits, it also obtains the lowest Type JS on three, demonstrating accurate recovery of future population composition. scHOMURA further performs strongly across gene-level metrics, capturing both differential-expression patterns and the direction of transcriptional change. Together, these results establish strong baseline forecasting performance before moving to the more challenging extended-horizon and rare/emerging population settings.

Table 1: Benchmark performance on three representative forecasting splits, reported as mean $\pm$ standard deviation (SD) over five seeds. For each seed, metrics are first averaged across held-out time points; the reported statistics are then computed across the five seed-level averages. **Bold** and underlined values indicate the best and second-best results, respectively.

Method	$W_2 \downarrow$	DE-AP$\uparrow$	DE-dir$\uparrow$	$\Delta r \uparrow$	MMD$\downarrow$	Variogram$\downarrow$	Type-JS$\downarrow$
CD-long							
scHOMURA	**16.35** $\pm$ **0.17**	**0.298** $\pm$ **0.039**	**0.768** $\pm$ **0.035**	**0.369** $\pm$ **0.041**	**0.132** $\pm$ **0.015**	**0.0226** $\pm$ **0.0011**	**0.079** $\pm$ **0.006**
LGP-OT	17.94 $\pm$ 0.33	0.238 $\pm$ 0.008	0.672 $\pm$ 0.053	0.218 $\pm$ 0.021	0.136 $\pm$ 0.009	0.0254 $\pm$ 0.0019	0.111 $\pm$ 0.014
scNODE	19.83 $\pm$ 0.68	0.241 $\pm$ 0.015	0.603 $\pm$ 0.029	0.169 $\pm$ 0.032	0.150 $\pm$ 0.026	0.0451 $\pm$ 0.0076	0.115 $\pm$ 0.023
MIOFlow	24.15 $\pm$ 1.28	0.286 $\pm$ 0.043	0.729 $\pm$ 0.027	0.258 $\pm$ 0.030	0.336 $\pm$ 0.026	0.1047 $\pm$ 0.0181	0.206 $\pm$ 0.052
PRESCIENT	19.20 $\pm$ 0.07	0.269 $\pm$ 0.005	0.401 $\pm$ 0.024	$-$0.293 $\pm$ 0.006	0.306 $\pm$ 0.006	0.0607 $\pm$ 0.0005	0.129 $\pm$ 0.000
VGFM	1050.81 $\pm$ 943.81	0.159 $\pm$ 0.025	0.598 $\pm$ 0.074	0.096 $\pm$ 0.045	0.779 $\pm$ 0.222	15.2742 $\pm$ 14.3838	0.359 $\pm$ 0.103
BetaDiff							
scHOMURA	**34.48** $\pm$ **0.71**	**0.706** $\pm$ **0.039**	**0.915** $\pm$ **0.032**	**0.678** $\pm$ **0.056**	**0.188** $\pm$ **0.039**	0.0955 $\pm$ 0.0063	**0.124** $\pm$ **0.048**
LGP-OT	36.28 $\pm$ 0.27	0.656 $\pm$ 0.010	0.507 $\pm$ 0.021	0.368 $\pm$ 0.032	0.352 $\pm$ 0.012	**0.0798** $\pm$ **0.0063**	0.251 $\pm$ 0.017
scNODE	39.98 $\pm$ 1.12	0.632 $\pm$ 0.013	0.552 $\pm$ 0.077	0.256 $\pm$ 0.125	0.237 $\pm$ 0.031	0.1549 $\pm$ 0.0131	0.167 $\pm$ 0.018
MIOFlow	44.03 $\pm$ 2.05	0.679 $\pm$ 0.025	0.373 $\pm$ 0.177	$-$0.276 $\pm$ 0.195	0.473 $\pm$ 0.126	0.2155 $\pm$ 0.0303	0.319 $\pm$ 0.081
PRESCIENT	55.55 $\pm$ 0.19	0.558 $\pm$ 0.007	0.364 $\pm$ 0.007	$-$0.290 $\pm$ 0.005	0.676 $\pm$ 0.003	0.4315 $\pm$ 0.0023	0.383 $\pm$ 0.003
VGFM	64.96 $\pm$ 4.80	0.651 $\pm$ 0.016	0.648 $\pm$ 0.015	0.234 $\pm$ 0.046	0.306 $\pm$ 0.021	0.2233 $\pm$ 0.0190	0.147 $\pm$ 0.025
CR-short							
scHOMURA	**15.62** $\pm$ **0.25**	**0.152** $\pm$ **0.010**	**0.669** $\pm$ **0.044**	0.142 $\pm$ 0.057	0.094 $\pm$ 0.005	0.0208 $\pm$ 0.0012	—
LGP-OT	15.78 $\pm$ 0.29	0.102 $\pm$ 0.032	0.570 $\pm$ 0.074	**0.192** $\pm$ **0.050**	0.103 $\pm$ 0.014	**0.0114** $\pm$ **0.0012**	—
scNODE	16.15 $\pm$ 0.30	0.059 $\pm$ 0.011	0.539 $\pm$ 0.053	0.035 $\pm$ 0.036	**0.074** $\pm$ **0.010**	0.0192 $\pm$ 0.0017	—
MIOFlow	17.68 $\pm$ 0.73	0.105 $\pm$ 0.027	0.450 $\pm$ 0.078	$-$0.147 $\pm$ 0.078	0.093 $\pm$ 0.018	0.0433 $\pm$ 0.0111	—
PRESCIENT	18.42 $\pm$ 0.41	0.084 $\pm$ 0.021	0.557 $\pm$ 0.033	$-$0.088 $\pm$ 0.048	0.275 $\pm$ 0.024	0.0663 $\pm$ 0.0042	—
VGFM	20.28 $\pm$ 0.58	0.075 $\pm$ 0.013	0.451 $\pm$ 0.033	$-$0.072 $\pm$ 0.043	0.255 $\pm$ 0.050	0.0633 $\pm$ 0.0082	—

4.3 Extended-Horizon Forecasting

We next evaluate forecasting stability as predictions extend further beyond the observed temporal range. CD-long and CR-long include seven and thirteen consecutive held-out snapshots, respectively, enabling evaluation over substantially longer forecasting horizons.

As shown in Figure 2, scHOMURA accumulates error more slowly with increasing forecast horizon. On CD-long, its W_2 remains relatively stable across seven held-out stages, while LGP-OT, scNODE, and MIOFlow increase by approximately 43%, 63%, and 139%, respectively, and VGFM becomes unstable. A similar pattern is observed on CR-long, where scHOMURA maintains low error across

13 future snapshots and achieves the lowest W_2 at 8 of 13 time points. Together, these results show that multi-horizon learning improves stability over extended forecasts and slows error accumulation beyond the observed window.

4.4 Forecasting Rare and Emerging Populations

We next examine rare and emerging populations in the CD-short and BetaDiff settings. We consider populations that are initially rare and later expand, as well as populations absent from the training snapshots but present at future stages. To probe these cases, we define three settings with progressively less target information available during training: **rare-state expansion**, where the target is observed at low abundance; **natural emergence**, where it is absent from all training snapshots; and **controlled masked emergence**, where an otherwise observed target is removed from every training snapshot. Together, these settings test whether early observations contain sufficient information to anticipate future changes in population identity and abundance (details provided in Appendix A.2.3).

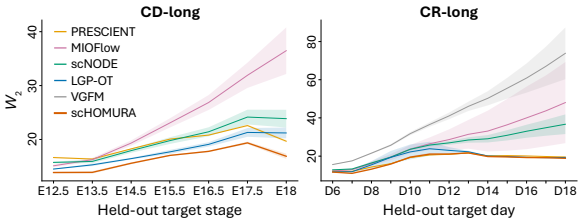


Figure 2: Extended-horizon forecasting performance. W_2 is reported across successive held-out future time points. Lines and shaded regions show mean ± SD over five seeds. VGFM is omitted from CD-long as its error grows far beyond the scale of the other methods.

Rare-state expansion. A central challenge in developmental forecasting is to identify populations that are rare at observed stages but later expand substantially. On CD-short, scHOMURA consistently prioritizes these future-expanding populations and predicts their abundance changes more accurately than methods that largely preserve the observed composition (Table 2); Figure S1 and S2 provide complementary views of the predicted future-state distributions and rare-population abundances. scHOMURA is the only learned method to achieve positive mean GrowthSkill (+0.172), outperforming persistence of the last observed composition. BetaDiff provides a more data-limited test (Table S4): from the single observed S3c→S4c transition, scHOMURA correctly anticipates expansion of the rare `sst_hhex` population but underestimates its magnitude. Together, these results separate expansion identification from abundance calibration, with Section 4.6 further showing that multi-horizon learning primarily improves the latter.

Natural emergence. A more challenging setting arises when a population is absent from all training snapshots and appears only at future stages. This analysis is restricted to BetaDiff, where `exo`, `foxj1`, `sc_beta`, and `sc_ec` are absent from both training snapshots and appear at test time; CD contains no qualifying future-only populations. scHOMURA generates cells that reach their future expression-space regions and achieves the highest emergence precision (Table 2). This recovery is selective rather than indiscriminate: compared with persistence and local-noise baselines, a larger fraction of extrapolated cells fall near states subsequently observed. These results support selective recovery of future-emergent expression regions rather than absolute proof of de novo bi-

Table 2: Forecasting rare and emerging populations on CD-short and BetaDiff.

	Rare-state expansion (CD-short)				**Natural emergence (BetaDiff)**		
Method	Expn. AP↑	Dir.↑	ΔMAE↓	GrowthSkill↑	Accepted↑	Unsupported↓	Precision↑
scHOMURA	**0.800**	**0.650**	**0.015**	**+0.172**	0.064	0.043	**0.594**
LGP-OT	0.500	0.250	0.038	−1.138	0.000	**0.000**	—
scNODE	0.650	0.250	0.042	−1.386	0.009	0.046	0.168
MIOFlow	0.617	0.350	0.039	−1.168	0.023	0.049	0.269
PRESCIENT	0.683	0.250	0.054	−2.039	0.028	0.705	0.038
VGFM	0.650	0.300	0.050	−1.787	0.050	0.495	0.093
Persistence	0.633	0.450	0.018	0.000	0.041	0.055	0.427
Noise-halo	0.683	0.450	0.018	−0.001	**0.101**	0.107	0.484

Table 3: Forecasting controlled masked emergence on BetaDiff.

	sst_hhex (true fraction: 1.2–2.7%)				sc_alpha (true fraction: 33–45%)			
Method	Detected	Fraction	Cond. $W_2 \downarrow$	OOS-prec.↑	Detected	Fraction	Cond. $W_2 \downarrow$	OOS-prec.↑
scHOMURA	**7**/10	**2.2%**	**16.3**	**0.036**	**10**/10	12.1%	**23.3**	0.477
LGP-OT	0/10	0.0%	—	0.000	**10**/10	9.9%	26.5	0.182
scNODE	0/10	0.0%	—	0.000	**10**/10	**23.4%**	30.9	**0.870**
MIOFlow	0/10	0.0%	—	0.000	**10**/10	4.4%	29.1	0.174
PRESCIENT	0/10	0.0%	—	0.000	**10**/10	2.3%	42.3	0.024
VGFM	0/10	0.0%	—	0.000	**10**/10	8.4%	56.3	0.103
Persistence	0/10	0.1%	—	—	1/10	0.4%	23.7	—
Noise-halo	0/10	0.1%	—	0.000	**10**/10	2.9%	27.3	0.071

ological state generation, given the non-zero empirical false-positive floor of the support estimator (Appendix A.2.3).

Controlled masked emergence. We next ask whether a future cellular state can still be recovered when all examples of the target population are removed from the training snapshots in BetaDiff. For sst_hhex, scHOMURA recovers the masked population in 7 of 10 seed–time combinations, whereas none of the competing methods or simple controls recover it (Table 3); the predicted population also reaches an abundance comparable to that observed at the corresponding future stages. Because no sst_hhex cells are available during training, this result shows that the dynamics of the remaining populations can provide sufficient information to recover a future state absent from the training observations. The sc_alpha experiment represents a complementary regime: unlike the rare sst_hhex population, it is abundant and lies within a densely observed endocrine trajectory. Accordingly, all generative methods recover the masked state, while scHOMURA places it close to the observed future population but underestimates its abundance. Together, the two cases suggest that masked-state recovery depends on available dynamical support, and that state recovery and abundance calibration are distinct forecasting capabilities.

All methods, including scHOMURA, fail to recover the masked targets on CD-short (Table S5), suggesting limited dynamical support for these states and highlighting the identifiability limits of masked-state emergence forecasting. scHOMURA nevertheless maintains strong global forecasting accuracy on the remaining population.

4.5 Interpreting Model-Inferred Developmental Transitions

We next asked whether scHOMURA-inferred transitions are consistent with known developmental relationships. We focused on fibroblast-like cells in CD, which expand from 2.21% at E11.5 to 7.21% at E12.5 and are known to receive developmental contributions from both epicardial and endocardial lineages. Epicardial progenitors constitute a major source of ventricular fibroblasts, while lineage-tracing studies have also identified endocardial contributions, including through intermediate mesenchymal states (Acharya et al., 2012; Moore-Morris et al., 2014; Ali et al., 2014; Huang et al., 2019). This provides a clear test of whether scHOMURA both anticipates the population increase and links the predicted fibroblast-like states to plausible earlier cellular sources.

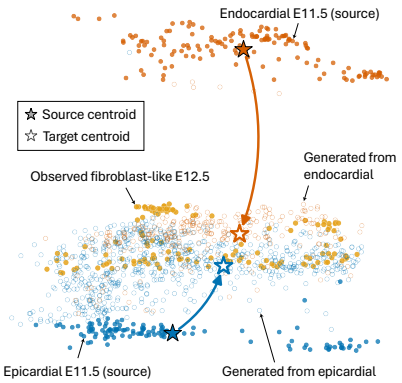


Figure 3: UMAP visualization of model-inferred transitions from E11.5 epicardial and endocardial states toward E12.5 fibroblast-like cells.

Consistent with these developmental relationships, scHOMURA links E12.5 fibroblast-like states to two transcriptionally distinct E11.5 source regions associated with epicardial and endocardial populations (Figure 3). One route follows the canonical epicardial-to-fibroblast direction, whereas the other is linked to endocardial-associated source states, consistent with the additional endocardial contribution reported by lineage-tracing studies.

Table 4: Ablation study of scHOMURA on the three annotated forecasting splits. For each seed, metrics are first averaged across held-out time points; values are then reported as mean ± SD over five seeds. **Bold** indicates metrics degraded by at least 40% relative to the full model.

Dataset	Variant	$W_2 \downarrow$	DE-AP↑	DE-dir↑	$\Delta r \uparrow$	MMD↓	Variogram↓	Type JS↓
CD-long	**Full (scHOMURA)**	16.35±0.17	0.298±0.039	0.768±0.035	0.369±0.041	0.132±0.015	0.0226±0.0011	0.079±0.006
	w/o residual flow	17.33±0.20	0.211±0.028	0.638±0.049	**0.195±0.049**	**0.205±0.023**	0.0247±0.0007	**0.161±0.032**
	w/o multi-horizon	16.36±0.16	0.306±0.015	0.768±0.073	0.365±0.069	0.141±0.011	0.0237±0.0027	0.083±0.013
	w/o growth modeling	16.21±0.16	0.326±0.045	0.789±0.038	0.399±0.058	0.127±0.011	0.0223±0.0009	0.061±0.013
CD-short	**Full (scHOMURA)**	15.09±0.16	0.263±0.085	0.575±0.118	0.261±0.059	0.044±0.008	0.0142±0.0014	0.013±0.005
	w/o residual flow	15.12±0.22	0.318±0.062	0.595±0.123	0.337±0.024	0.041±0.009	0.0130±0.0013	0.014±0.002
	w/o multi-horizon	15.48±0.34	0.235±0.063	0.478±0.084	0.181±0.103	0.048±0.009	0.0157±0.0021	**0.022±0.012**
	w/o growth modeling	16.01±0.21	0.267±0.096	0.529±0.134	**−0.016±0.102**	0.053±0.010	0.0135±0.0014	**0.028±0.008**
BetaDiff	**Full (scHOMURA)**	34.48±0.71	0.706±0.039	0.915±0.032	0.678±0.056	0.188±0.039	0.0955±0.0063	0.124±0.048
	w/o residual flow	34.31±0.72	0.709±0.052	0.844±0.086	0.643±0.082	**0.324±0.063**	0.0902±0.0061	**0.279±0.114**
	w/o multi-horizon	34.48±0.71	0.706±0.039	0.915±0.032	0.678±0.056	0.188±0.039	0.0955±0.0063	0.124±0.048
	w/o growth modeling	35.47±0.78	0.667±0.055	0.781±0.207	0.457±0.347	0.258±0.046	0.1057±0.0131	**0.190±0.046**

Importantly, cell-type identities are not provided to scHOMURA during training and are used only after forecasting to interpret the inferred transitions. The recovery of both epicardial- and endocardial-associated source states therefore suggests that the learned temporal dynamics preserve biologically meaningful developmental structure. In this example, scHOMURA not only anticipates the expansion of a rare fibroblast-like population, but also links its predicted future states to earlier cellular populations that are consistent with known developmental origins.

4.6 Ablation Study

We ablate the three main components of scHOMURA, residual dynamics, multi-horizon learning, and the state-dependent growth field, to assess their roles in future-population forecasting (Tables 4 and S6). Removing the residual flow mainly degrades internal population structure and cell-type composition while leaving coarse state transport relatively intact, indicating that deterministic displacement captures shared temporal progression whereas residual dynamics model heterogeneous outcomes.

Table 5: Effect of multi-horizon learning on rare-state expansion (CD-short).

Variant	Expn. AP↑	GrowthSkill↑
Full (scHOMURA)	0.800	$+0.172$
w/o multi-horizon	1.000	-1.329

Removing multi-horizon learning preserves rare-state ranking but substantially impairs abundance calibration over longer intervals: GrowthSkill drops from $+0.172$ to -1.329 despite strong expansion ranking (Table 5). Removing the growth field has little effect on state transport but weakens recovery of future population composition when relative abundances shift. Together, these ablations show complementary roles for the three components: residual dynamics capture heterogeneous outcomes, multi-horizon learning supports extended-range forecasting and abundance calibration, and the growth field models population remodeling.

5 Discussion

scHOMURA provides a multi-horizon generative framework for forecasting both cell-state transitions and population remodeling from unpaired temporal snapshots. Beyond improving aggregate forecasting accuracy, our results show that future-population prediction comprises distinct capabilities: reaching future state support, recovering emerging populations, and calibrating their abundance need not coincide. This separation exposes failure modes that are hidden by global distributional metrics and motivates a more target-aware view of single-cell forecasting.

Limitations. scHOMURA predicts relative population composition rather than absolute cell numbers or mechanistic birth–death rates. Recovery of unseen states also depends on sufficient temporal support; the failure on controlled CD masking illustrates this identifiability boundary. In addition, transcriptional proximity alone cannot establish causal lineage relationships. Future work should incorporate absolute population dynamics, lineage and spatial information, predictive uncertainty, and broader evaluation of when emergence forecasting is biologically identifiable.

AI USE STATEMENT

In this work, we used generative AI tools to assist with the implementation of methods, including the development of code and model-training scripts; data cleaning and preprocessing; translation and language assistance; and brainstorming and providing feedback on aspects of the model architecture and research methodology. We did not use generative AI tools to generate synthetic datasets, formulate mathematical claims, provide ingredients for or assist with mathematical proofs, support qualitative or thematic data analysis, or interpret experimental results. Other required disclosure tasks were not applicable to this work.

Additionally, we used generative AI tools for literature search, identification and summarization of relevant literature, and drafting and editing parts of the manuscript. All AI-assisted outputs were reviewed by the authors. References identified with AI assistance were verified against the original publications, AI-assisted code and training scripts were inspected and tested for correctness, and data-processing outputs were checked using the same quality-control procedures applied throughout the study. The authors independently evaluated all methodological decisions, experimental results, and scientific claims. We take responsibility for the final content of this work, including text, claims, code, and other artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

Our code is available at https://anonymous.4open.science/r/scHomura/. All datasets used in this study are publicly available from the sources reported in the original studies. The preprocessed datasets and corresponding trained models used in our experiments are available at the anonymous Figshare repository: https://figshare.com/s/9a5efce688db0f8a7574. Full methodological details, implementation settings, and experimental protocols are provided in the Appendix.

REFERENCES

Asha Acharya, Seung Tae Baek, Guo Huang, Banu Eskiocak, Sean Goetsch, Caroline Y Sung, Serena Banfi, Marion F Sauer, Gregory S Olsen, Jeremy S Duffield, et al. The bhlh transcription factor tcf21 is required for lineage-specific emt of cardiac fibroblast progenitors. *Development*, 139(12):2139–2149, 2012.

Shah R Ali, Sara Ranjbarvaziri, Mahmood Talkhabi, Peng Zhao, Ali Subat, Armin Hojjat, Paniz Kamran, Antonia MS Müller, Katharina S Volz, Zhaoyi Tang, et al. Developmental heterogeneity of cardiac fibroblasts does not predict pathological proliferation and activation. *Circulation research*, 115(7):625–635, 2014.

Mehmet Yigit Balik and Harri Lähdesmäki. Modeling temporal scrna-seq data with latent gaussian process and optimal transport. *arXiv preprint arXiv:2605.20989*, 2026.

Lenaic Chizat, Gabriel Peyré, Bernhard Schmitzer, and François-Xavier Vialard. Scaling algorithms for unbalanced optimal transport problems. *Mathematics of computation*, 87(314):2563–2609, 2018.

Jun Ding, Nadav Sharon, and Ziv Bar-Joseph. Temporal modelling using single-cell transcriptomics. *Nature Reviews Genetics*, 23(6):355–368, 2022.

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.

Wei Feng, Abha Bais, Haoting He, Cassandra Rios, Shan Jiang, Juan Xu, Cindy Chang, Dennis Kostka, and Guang Li. Single-cell transcriptomic analysis identifies murine heart molecular features at embryonic and neonatal stages. *Nature communications*, 13(1):7960, 2022.

Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The journal of machine learning research*, 13:723–773, 2012.

Xinyan Huang, Teng Feng, Zhen Jiang, Jufeng Meng, Shan Kou, Zhengkai Lu, Weize Chen, Chao-Po Lin, Bin Zhou, and Hui Zhang. Dual lineage tracing identifies intermediate mesenchymal stage for endocardial contribution to fibroblasts, coronary mural cells, and adipocytes. *Journal of Biological Chemistry*, 294(22):8894–8906, 2019.

Guillaume Huguet, Daniel Sumner Magruder, Alexander Tong, Oluwadamilola Fasina, Manik Kuchroo, Guy Wolf, and Smita Krishnaswamy. Manifold interpolating optimal-transport flows for trajectory inference. *Advances in neural information processing systems*, 35:29705–29718, 2022.

Qi Jiang and Lin Wan. A physics-informed neural sde network for learning cellular dynamics from time-series scrna-seq data. *Bioinformatics*, 40(Supplement_2):ii120–ii127, 2024.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

Yifei Liu, Kai Huang, and Wanze Chen. Resolving cellular dynamics using single-cell temporal transcriptomics. *Current opinion in biotechnology*, 85:103060, 2024.

Tapio Lönnberg, Valentine Svensson, Kylie R James, Daniel Fernandez-Ruiz, Ismail Sebina, Ruddy Montandon, Megan SF Soon, Lily G Fogg, Arya Sheela Nair, Urijah N Liligeto, et al. Single-cell rna-seq and computational analysis using temporal mixture modeling resolves th1/tfh fate bifurcation in malaria. *Science immunology*, 2(9):eaal2192, 2017.

Thomas Moore-Morris, Nuno Guimarães-Camboa, Indroneal Banerjee, Alexander C Zambon, Tatiana Kisseleva, Aurélie Velayoudon, William B Stallcup, Yusu Gu, Nancy D Dalton, Marta Cedenilla, et al. Resident fibroblast lineages mediate pressure overload–induced cardiac fibrosis. *The Journal of clinical investigation*, 124(7):2921–2934, 2014.

Siyu Pu, Qingqing Long, Xiaohan Huang, Haotian Chen, Jiajia Wang, Meng Xiao, Xiao Luo, Hengshu Zhu, Yuanchun Zhou, and Xuezhi Wang. From snapshots to trajectories: Learning single-cell gene expression dynamics via conditional flow matching. *arXiv preprint arXiv:2605.22340*, 2026.

Geoffrey Schiebinger, Jian Shu, Marcin Tabaka, Brian Cleary, Vidya Subramanian, Aryeh Solomon, Joshua Gould, Siyan Liu, Stacie Lin, Peter Berube, et al. Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming. *Cell*, 176(4):928–943, 2019.

Yutong Sha, Yuchi Qiu, Peijie Zhou, and Qing Nie. Reconstructing growth and dynamic trajectories from single-cell transcriptomics data. *Nature Machine Intelligence*, 6(1):25–39, 2024.

Alexander Tong, Jessie Huang, Guy Wolf, David Van Dijk, and Smita Krishnaswamy. Trajectorynet: A dynamic optimal transport network for modeling cellular dynamics. In *International conference on machine learning*, pp. 9526–9536. PMLR, 2020.

Adrian Veres, Aubrey L Faust, Henry L Bushnell, Elise N Engquist, Jennifer Hyoje-Ryu Kenty, George Harb, Yeh-Chuin Poh, Elad Sintov, Mads Gürtler, Felicia W Pagliuca, et al. Charting cellular identity during human in vitro β-cell differentiation. *Nature*, 569(7756):368–373, 2019.

Philipp Wahle, Giovanna Brancati, Christoph Harmel, Zhisong He, Gabriele Gut, Jacobo Sarabia Del Castillo, Aline Xavier da Silveira dos Santos, Qianhui Yu, Pascal Noser, Jonas Simon Fleck, et al. Multimodal spatiotemporal phenotyping of human retinal organoid development. *Nature Biotechnology*, 41(12):1765–1775, 2023.

Dongyi Wang, Yuanwei Jiang, Zhenyi Zhang, Xiang Gu, Peijie Zhou, and Jian Sun. Joint velocity-growth flow matching for single-cell dynamics modeling. *Advances in Neural Information Processing Systems*, 38:160552–160588, 2026.

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.

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.

A Appendix

A.1 Details of scHOMURA

A.1.1 Latent Representation and Prototype-level UOT Construction

For each dataset, we use 2,000 highly variable genes selected from the training snapshots only. No held-out snapshot is used for feature selection. We encode expression profiles with a deterministic autoencoder whose encoder has widths $2000 \to 512 \to 128 \to 32$, with SiLU activations, LayerNorm after the first hidden layer, and a final $\tanh$ activation. The decoder has widths $32 \to 128 \to 512 \to 2000$, with SiLU hidden activations and a Softplus output. The autoencoder is trained using mean-squared reconstruction error on training snapshots and is subsequently frozen. All temporal modeling is performed in the resulting 32-dimensional latent space.

Let K_{train} denote the number of training snapshots. We construct source–target training pairs for every available snapshot horizon in

$$\mathcal{H} = \{1, \ldots, \min(3, K_{\mathrm{train}} - 1)\}.$$

For each source–target pair with $h \in \mathcal{H}$, we independently cluster the source and target latent populations into at most 64 prototypes using MiniBatchKMeans with batch size $1{,}024$, three initializations, and pair-specific random seeds, where each prototype represents a local cluster of nearby cells. Let $a_i = n_i^0/N_0$ and $b_j = n_j^1/N_1$ denote the empirical masses of source prototype i and target prototype j, and let c_i^0 and c_j^1 denote their corresponding latent-space centers. We define the transport cost between source and target prototypes as

$$C_{ij} = \left\| c_i^0 - c_j^1 \right\|_2^2.$$

To reduce sensitivity to the overall scale of the latent space, we normalize the cost matrix as

$$\widetilde{C}_{ij} = \frac{C_{ij}}{\mathrm{median}\,\{C_{kl} : C_{kl} > 0\}}.$$

An unbalanced optimal transport coupling γ is then computed from $\widetilde{C}$ using entropically regularized unbalanced Sinkhorn transport with entropy regularization 0.08, marginal-relaxation parameter 1.0, and at most $3{,}000$ iterations. We row-normalize γ to obtain a transition matrix P, where each row defines the conditional distribution over target prototypes given a source prototype.

Training pseudo-transitions are sampled hierarchically. We first sample a source–target prototype pair according to the transport mass in γ, and then independently sample one source cell and one target cell from the corresponding prototype clusters. This yields a latent pseudo-transition pair (z, z') used to supervise the state-transition model.

A.1.2 Ordinal Snapshot Conditioning and Deterministic Displacement

Snapshots are assigned ordinal indices $k = 0, \ldots, K-1$. Let K_{last} be the index of the final training snapshot. For a transition from source index k to target index $k' = k + h$, we use the normalized source coordinate and normalized horizon

$$\bar{k} = \frac{k}{K_{\mathrm{last}}}, \qquad \bar{h} = \frac{h}{K_{\mathrm{last}}}.$$

The normalized snapshot coordinates $\bar{k}$ and $\bar{h}$ can preserve relative sampling time and elapsed time. For stage-based data, they represent developmental order and the number of snapshot steps rather than calibrated physical time. Each scalar condition v is encoded as

$$\psi(v) = [v, \sin(\pi v), \cos(\pi v)].$$

The displacement network receives $\psi(\bar{k})$ and $\psi(\bar{h})$, whereas the residual-flow network additionally receives the flow-path encoding $\psi(s)$.

For each sampled pseudo-transition (z, z'), we define the observed latent displacement as

$$v^\star = z' - z.$$

The deterministic displacement network is trained to predict the shared transition:

$$\mathcal{L}_{\text{mean}} = \mathbb{E}\left[w^{(\text{under})} \left\|\mu_\theta(z, \bar{k}, \bar{h}) - v^\star\right\|_2^2\right].$$

Here, $w^{(\text{under})}$ denotes the transition reweighting term defined below. This objective encourages μ_θ to capture the shared displacement associated with the source state, ordinal source position, and snapshot horizon.

A.1.3 Conditional Residual Flow Matching

The residual flow models transition variability not explained by the deterministic displacement. For each pseudo-transition, we define the normalized residual target as

$$\varepsilon^\star = \frac{v^\star - \text{sg}\big(\mu_\theta(z, \bar{k}, \bar{h})\big)}{\sigma},$$

where sg denotes stop-gradient and σ controls the global residual scale. The global residual scale is learned jointly with the transition model. Specifically, we parameterize

$$\sigma = \text{clip}\left(\exp(\rho),\, 0.05,\, 5.0\right),$$

where ρ is a single scalar parameter initialized to zero. Consequently, the overall magnitude of residual variation is inferred from the observed transitions rather than fixed as a hyperparameter. To learn the conditional residual distribution, we use flow matching between Gaussian noise and $\varepsilon^\star$. We sample

$$s \sim U(0,1), \qquad \eta \sim \mathcal{N}(0, I),$$

and define the linear interpolation

$$\xi_s = (1-s)\eta + s\varepsilon^\star,$$

whose target velocity is

$$u^\star = \varepsilon^\star - \eta.$$

The conditional velocity field f_ϕ is trained with

$$\mathcal{L}_{\text{flow}} = \mathbb{E}\left[w^{(\text{under})} \left\|f_\phi\left(\xi_s, s \mid z, \bar{k}, \bar{h}\right) - u^\star\right\|_2^2\right].$$

At inference, integrating f_ϕ from Gaussian noise yields a residual sample. We use the explicit Euler update

$$\xi^{(m+1)} = \xi^{(m)} + \frac{1}{S} f_\phi\left(\xi^{(m)}, \frac{m+1/2}{S} \mid z, \bar{k}, \bar{h}\right), \qquad m = 0, \ldots, S-1,$$

initialized with $\xi^{(0)} \sim \mathcal{N}(0, I)$. We use $S = 8$ steps for differentiable samples in the training-time MMD and consistency objectives and $S = 24$ steps for final population generation. We denote the resulting conditional flow map by $\Phi_\phi(\eta \mid z, \bar{k}, \bar{h})$.

A.1.4 Undercoverage-aware Transition Reweighting

Because UOT does not enforce the full target marginal, some target prototypes may receive less transported mass than their observed abundance. To reduce this bias, we upweight pseudo-transitions terminating in undercovered target prototypes.

For target prototype j, we first normalize its incoming transported mass into a relative incoming share:

$$\tilde{b}_j = \frac{\sum_i \gamma_{ij}}{\sum_{k,\ell} \gamma_{k\ell}}.$$

We then define its relative undercoverage as

$$u_j = \text{clip}\left(\frac{b_j - \tilde{b}_j}{\max(b_j, \epsilon)}, 0, 1\right),$$

and assign

$$w_j^{(\text{under})} = 1 + 2u_j.$$

For a pseudo-transition terminating in target prototype j, we set $w^{(\text{under})} = w_j^{(\text{under})}$. The weight is applied only to the deterministic-displacement and residual-flow objectives.

A.1.5 State-dependent Growth Parameterization

We parameterize the growth field as

$$r_\omega(z, \bar{k}) = r_{\omega,0}(z) + \tanh(\bar{k})\, r_{\omega,1}(z),$$

where $r_{\omega,0}$ captures the baseline state-dependent contribution and $r_{\omega,1}$ allows this contribution to vary with ordinal source-snapshot position.

For source prototype p with empirical mass a_p and center c_p, the growth field reweights its contribution over forecast horizon h as

$$q_p = \frac{a_p \exp\big(h\, r_\omega(c_p, \bar{k})\big)}{\sum_k a_k \exp\big(h\, r_\omega(c_k, \bar{k})\big)}.$$

Together with the UOT-derived transition matrix P, this gives the predicted target composition

$$\hat{b} = qP.$$

The growth objective is

$$\mathcal{L}_{\text{growth}} = \mathrm{JS}(\hat{b}, b) + 0.1\, \frac{\mathrm{KL}(q \,\|\, a)}{h} + 0.01 \sum_p a_p r_{\omega,1}(c_p)^2.$$

The Jensen–Shannon term matches the predicted and observed target compositions. The KL term discourages unnecessary reweighting of the source population, with the division by h allowing larger compositional changes over longer ordinal forecast horizons. The final term regularizes the magnitude of the snapshot-dependent component of the growth field. The growth score represents relative future contribution per snapshot step and is not interpreted as a proliferation rate per unit of physical time.

A.1.6 Distributional and Multi-horizon Regularization

In addition to transition-level supervision, scHOMURA aligns generated and observed target populations in latent space using an RBF maximum mean discrepancy objective, $\mathcal{L}_{\text{MMD}}$.

For training, we use the biased empirical estimator, including diagonal kernel terms. For equally sized generated and target latent minibatches $\hat{Z} = \{\hat{z}_i\}_{i=1}^n$ and $Z' = \{z'_i\}_{i=1}^n$,

$$\mathcal{L}_{\text{MMD}} = \frac{1}{n^2} \sum_{i,j} k(\hat{z}_i, \hat{z}_j) + \frac{1}{n^2} \sum_{i,j} k(z'_i, z'_j) - \frac{2}{n^2} \sum_{i,j} k(\hat{z}_i, z'_j),$$

where $k(u, v) = \exp(-\gamma \|u - v\|_2^2)$. For each minibatch, γ is the reciprocal of the median positive squared distance over the union of generated and target latent states, with 10^{-8} added for numerical stability.

We further encourage consistency across forecast horizons. For $h = 2$, the same future population can be predicted either directly or by composing two one-step transitions. We therefore define

$$\mathcal{L}_{\text{cons}} = \mathrm{MMD}^2_{\text{RBF}} \left(\hat{Z}^{\text{direct}}_{t \to t+2}, \hat{Z}^{\text{composed}}_{t \to t+1 \to t+2} \right).$$

The consistency objective is evaluated only when a direct two-step training pair is available. It compares a direct $h = 2$ transition with two recursively composed $h = 1$ transitions using an RBF-MMD loss. The term is evaluated every two temporal-model optimization steps using a batch of 256 cells. Growth-based source reweighting is excluded from this consistency calculation.

The complete temporal-model training objective is

$$\mathcal{L} = \mathcal{L}_{\text{flow}} + 5\mathcal{L}_{\text{mean}} + 0.05\mathcal{L}_{\text{MMD}} + 0.05\mathcal{L}_{\text{cons}} + \mathcal{L}_{\text{growth}}.$$

A.1.7 MULTI-HORIZON FORECASTING

scHOMURA is trained directly on every available forecast horizon $h \in \mathcal{H}$, where

$$\mathcal{H} = \{1, \ldots, \min(3, K_{\text{train}} - 1)\}.$$

The maximum direct horizon is therefore three snapshots, but the actual set of trained horizons depends on the number of available training snapshots. All horizons share the same displacement and residual-flow networks and are distinguished through $\bar{h} = h/K_{\text{last}}$. When multiple directly trained horizon forecasts are available for the same target time, they are combined using base mixture weights

$$w_1 = 0.6, \qquad w_2 = 0.3, \qquad w_3 = 0.1.$$

Weights associated with untrained horizons or unavailable source populations are removed, and the remaining weights are renormalized to sum to one.

For each generated cell, a forecast horizon is first sampled according to these mixture weights. The model then uses the population observed or generated h snapshot steps before the target as the source population. A source state is sampled according to the growth-dependent distribution and propagated through the shared stochastic transition model conditioned on $\bar{h}$.

For targets beyond the observed temporal range, newly generated populations are added to the available history and may subsequently serve as source populations. Extended-horizon forecasting therefore combines direct multi-horizon prediction with partial recursion, reducing reliance on repeated one-step propagation.

Architectures and optimization. The deterministic displacement and residual velocity networks each use two hidden layers of width 256 with SiLU activations. The growth field uses separate $32 \to 64 \to 1$ base and slope heads; the final slope layer is initialized to zero. The released implementation retains one shared learned eight-dimensional conditioning vector for checkpoint compatibility; because it is identical for every cell, it does not define discrete states or a VQ codebook. A legacy 32-dimensional GP input channel is fixed identically to zero and is omitted from the mathematical notation; no GP component is active. The autoencoder is optimized for 30 epochs with AdamW using learning rate 3×10^{-4}, weight decay 10^{-5}, and batch size 256. The temporal model is optimized for 40 epochs with AdamW using learning rate 5×10^{-4}, weight decay 10^{-5}, batch size 512, and gradient-norm clipping at 5. Unless otherwise specified, 2,000 cells are generated per test snapshot.

A.2 DETAILS OF EXPERIMENTS

A.2.1 DATASETS

We use the five datasets described in the main text. For datasets with multiple forecasting splits, each split is treated as an independent experiment. Predictions are generated for each held-out time point without access to intermediate test populations. Dataset-specific training and test boundaries are summarized in Table S1.

Mouse cardiac development (CD). The CD dataset profiles mouse heart development across successive embryonic stages, capturing the differentiation and maturation of diverse cardiac and non-cardiac cell populations (Feng et al., 2022).

Cellular reprogramming (CR). The CR dataset follows induced pluripotent stem-cell reprogramming over a densely sampled time course. It captures the progressive departure from the initial somatic state and diversification into distinct reprogramming trajectories (Schiebinger et al., 2019).

Zebrafish embryogenesis (Zebrafish). The Zebrafish dataset profiles whole-embryo development from early blastula and gastrulation stages to somitogenesis. During this period, initially undifferentiated embryonic populations progressively diversify into multiple developmental lineages. (Farrell et al., 2018).

In vitro beta-cell differentiation (BetaDiff). The BetaDiff dataset follows human stem-cell differentiation toward pancreatic endocrine cell identities. The time course spans progenitor and endocrine precursor states through later differentiated populations, accompanied by substantial changes in cellular composition (Veres et al., 2019).

Retinal organoid development (Retinal). The Retinal dataset follows human retinal organoid development over an extended time course. It captures the progression from early retinal progenitor states toward increasingly differentiated neuronal and glial populations(Wahle et al., 2023).

A.2.2 METRICS

2-Wasserstein distance (W_2). W_2 measures the overall discrepancy between the generated and observed future populations while accounting for the geometry of the expression space. For two empirical populations $X = x_i{}_{i=1}^{n}$ and $Y = y_j{}_{j=1}^{m}$ with uniform weights, we compute

$$W_2(X,Y) = \left[\min_{\pi \in \Pi(a,b)} \sum_{i=1}^{n} \sum_{j=1}^{m} \pi_{ij} \|x_i - y_j\|_2^2 \right]^{1/2}.$$

where $a_i = 1/n$, $b_j = 1/m$, and $\Pi(a,b)$ denotes the set of transport plans whose row and column marginals are a and b. Thus, W_2 measures the minimum transport cost required to transform one empirical cellular population into the other under squared Euclidean ground cost. Lower W_2 indicates closer agreement between the generated and observed future populations.

RBF MMD. We use maximum mean discrepancy (MMD) with a radial basis function (RBF) kernel to measure distributional agreement between the observed future population X_t and the generated population $\hat{X}_t$. Let x, x' denote independently sampled cells from X_t, and let $\hat{x}, \hat{x}'$ denote independently sampled cells from $\hat{X}_t$. The squared RBF-MMD is defined as

$$\mathrm{MMD}^2_{\mathrm{RBF}}(X_t, \hat{X}_t) = \mathbb{E}_{x,x' \sim X_t}\left[k(x,x')\right] + \mathbb{E}_{\hat{x},\hat{x}' \sim \hat{X}_t}\left[k(\hat{x},\hat{x}')\right] - 2\mathbb{E}_{x \sim X_t, \hat{x} \sim \hat{X}_t}\left[k(x,\hat{x})\right].$$

The first two terms measure the average within-population similarity of the observed and generated cells, whereas the third term measures their cross-population similarity. The discrepancy becomes small when cells drawn from the two populations exhibit similar kernel-based relationships. We compute these similarities using the RBF kernel

$$k(u,v) = \exp\left(-\frac{\|u-v\|_2^2}{2\sigma^2}\right).$$

where σ denotes the kernel bandwidth, which determines the scale over which two cells are considered similar. With a characteristic RBF kernel, MMD distinguishes probability distributions in the population limit. Lower MMD therefore indicates closer agreement between the generated and observed future populations.

Differential-expression average precision (DE-AP). DE-AP measures whether the predicted future population assigns large expression changes to genes that are differentially expressed in the observed transition. For each seed and forecast time point, we subsample at most 1,000 cells from the last observed population X_0, the observed future population X_t, and the generated population $\hat{X}_t$. Let $\mu_{0,g}$, $\mu_{t,g}$, and $\hat{\mu}_{t,g}$ denote the mean expression of gene g in these three populations. We define the observed and predicted changes in mean expression as

$$\Delta_g = \mu_{t,g} - \mu_{0,g}, \qquad \hat{\Delta}_g = \hat{\mu}_{t,g} - \mu_{0,g}.$$

Ground-truth DE genes are identified from the observed transition $X_0 \to X_t$. For each gene, we perform a two-sided Mann–Whitney U test between cells in X_0 and X_t, followed by Benjamini–Hochberg correction across genes. Gene g is labeled as differentially expressed when

$$q_g \le 0.05 \qquad \text{and} \qquad |\Delta_g| \ge 0.25.$$

If no gene satisfies both criteria at a particular forecast time point, the 100 genes with the largest $|\Delta_g|$ are used as positives to ensure that average precision is defined.

For the generated population, we do not perform a separate differential-expression test. Instead, each gene is assigned the prediction score

$$s_g = |\hat{\Delta}_g|,$$

so that genes predicted to undergo larger expression changes receive higher scores. Genes are ranked by s_g, and DE-AP is computed using non-interpolated average precision,

$$\text{DE-AP} = \sum_k (R_k - R_{k-1}) P_k,$$

where P_k and R_k are the precision and recall at successive score thresholds. Higher DE-AP indicates that genes identified as differentially expressed in the observed future transition tend to receive larger predicted expression changes. Because the ranking score depends on the magnitude $|\hat{\Delta}_g|$, DE-AP evaluates recovery of changing genes but does not distinguish up-regulation from down-regulation; directional agreement is evaluated separately by DE-dir.

Differential-expression direction agreement (DE-dir). DE-dir evaluates whether the model predicts the correct direction of transcriptional change for genes that are differentially expressed in the observed transition. We use the same ground-truth DE set $\mathcal{G}_{\text{DE}}$ defined for DE-AP, including the fallback to the 100 genes with the largest $|\Delta_g|$ when no gene satisfies the DE criteria. For each gene g, the observed and predicted changes in mean expression are

$$\Delta_g = \mu_{t,g} - \mu_{0,g}, \qquad \hat{\Delta}_g = \hat{\mu}_{t,g} - \mu_{0,g}.$$

DE-dir weights each gene by the magnitude of its observed expression change:

$$\text{DE-dir} = \frac{\sum_{g \in \mathcal{G}_{\text{DE}}} |\Delta_g| \mathbf{1}\left[\text{sign}(\hat{\Delta}_g) = \text{sign}(\Delta_g)\right]}{\sum_{g \in \mathcal{G}_{\text{DE}}} |\Delta_g|}.$$

Thus, genes with larger observed expression changes contribute more strongly to the score when their predicted direction is correct. The sign function is applied directly without an additional magnitude threshold. A predicted change of zero is therefore treated as incorrect when the corresponding observed change is nonzero. Higher DE-dir indicates better agreement with the direction of the observed transcriptional changes.

Pearson correlation of mean-expression changes (Δr). Δr measures whether the predicted gene-wise expression changes follow the same overall pattern as the observed changes. Using the observed and predicted mean-expression changes Δ_g and $\hat{\Delta}_g$ defined above, we compute the Pearson correlation across genes:

$$\Delta r = \frac{\sum_{g=1}^{G} (\Delta_g - \bar{\Delta})(\hat{\Delta}_g - \bar{\hat{\Delta}})}{\sqrt{\sum_{g=1}^{G} (\Delta_g - \bar{\Delta})^2} \sqrt{\sum_{g=1}^{G} (\hat{\Delta}_g - \bar{\hat{\Delta}})^2}}.$$

Here, $\bar{\Delta}$ and $\bar{\hat{\Delta}}$ denote the average observed and predicted expression changes across genes. Unlike DE-AP and DE-dir, Δr is computed across all 2,000 selected evaluation genes, without restricting the comparison to the ground-truth DE set. It measures agreement in the overall pattern of gene-wise mean-expression changes. Higher Δr indicates stronger agreement between the predicted and observed transcriptional dynamics.

Variogram error. Variogram error measures agreement in within-cell expression contrasts between pairs of genes in the observed and generated future populations. We randomly sample $M = 20{,}000$ pairs of distinct genes (i_m, j_m), allowing the same pair to be sampled more than once, and use the same set of pairs for both populations. For each population X and sampled gene pair m, we compute

$$v_m(X) = \frac{1}{|X|} \sum_{c \in X} \sqrt{|x_{c,i_m} - x_{c,j_m}|}.$$

The observed future population X_t and generated population $\hat{X}_t$ are each subsampled to at most 1,000 cells. We then compute the mean squared difference between their variogram values across the sampled gene pairs:

$$E_{\text{vario}} = \frac{1}{M} \sum_{m=1}^{M} \left[v_m(X_t) - v_m(\hat{X}_t) \right]^2.$$

Variogram error therefore complements metrics based on gene-wise mean changes by evaluating whether the generated population reproduces the average within-cell expression contrasts observed between gene pairs in the true future population. Lower values indicate better agreement.

Cell-type composition Jensen–Shannon divergence (Type JS). For datasets with cell-type annotations, we construct a fixed post-hoc label-transfer evaluator separately for each held-out target snapshot. Real cells without annotations and cell types represented by fewer than 10 annotated cells at that snapshot are excluded. We standardize each gene and fit a randomized PCA representation with $\min(50, G)$ components using at most 20,000 randomly sampled eligible real cells. A distance-weighted 15-nearest-neighbor classifier is then fitted using all eligible real target cells in this PCA space. The same fitted evaluator is applied to every method and seed for that target snapshot. Target cell-type labels and expression profiles are used for this post-hoc evaluation and are not provided to the generative models.

Every generated cell receives the classifier's hard label; no confidence threshold or unassigned category is used. Let $\mathcal{K}_t$ be the annotated cell types with at least 10 real cells at target time t, and let $\mathbf{p}_t$ and $\hat{\mathbf{p}}_t$ be the observed and predicted type proportions over $\mathcal{K}_t$. Their midpoint distribution is

$$\mathbf{m}_t = \frac{\mathbf{p}_t + \hat{\mathbf{p}}_t}{2}.$$

We compute the Jensen–Shannon divergence as

$$\text{Type-JS}_t = \frac{1}{2} D_{\text{KL}}(\mathbf{p}_t \| \mathbf{m}_t) + \frac{1}{2} D_{\text{KL}}(\hat{\mathbf{p}}_t \| \mathbf{m}_t),$$

where

$$D_{\text{KL}}(\mathbf{p}_t \| \mathbf{m}_t) = \sum_{k \in \mathcal{K}_t} p_{t,k} \log \frac{p_{t,k}}{m_{t,k}}.$$

We use natural logarithms and add 10^{-12} to each proportion before renormalization for numerical stability. Type JS is zero when the observed and predicted cell-type proportions agree and lies in $[0, \log 2]$; lower values indicate better agreement.

Abundance-change error (ΔMAE). For rare-state expansion, let $\mathcal{E}$ denote the set of rare cell types that satisfy the ground-truth expansion criterion, and let $\mathcal{T}_{\text{test}}$ denote the set of held-out future time points. To evaluate the magnitude of rare-state expansion, we measure changes in cell-type abundance relative to the final observed training composition. Let p_c^0 denote the proportion of cell type c at the final training time point, and let $p_{t,c}$ and $\hat{p}_{t,c}$ denote its observed and predicted proportions at future time t. The corresponding observed and predicted abundance changes are

$$\Delta p_{t,c} = p_{t,c} - p_c^0, \qquad \widehat{\Delta p}_{t,c} = \hat{p}_{t,c} - p_c^0.$$

We quantify the error in predicted abundance changes by averaging the absolute deviation across expanding rare types and held-out time points:

$$\Delta\mathrm{MAE}_{\mathrm{model}} = \frac{1}{|\mathcal{T}_{\mathrm{test}}||\mathcal{E}|} \sum_{t\in\mathcal{T}_{\mathrm{test}}} \sum_{c\in\mathcal{E}} \left| \widehat{\Delta p}_{t,c} - \Delta p_{t,c} \right|.$$

Lower ΔMAE indicates more accurate prediction of the magnitude of future compositional changes.

GrowthSkill. To interpret abundance-change error relative to a simple no-change prediction, we compare each model with a Persistence baseline that retains the final training composition at all future time points:

$$\hat{p}_{t,c}^{\mathrm{persist}} = p_c^0.$$

Because Persistence predicts no abundance change, its error is

$$\Delta\mathrm{MAE}_{\mathrm{persist}} = \frac{1}{|\mathcal{T}_{\mathrm{test}}||\mathcal{E}|} \sum_{t\in\mathcal{T}_{\mathrm{test}}} \sum_{c\in\mathcal{E}} |\Delta p_{t,c}|.$$

We then define GrowthSkill as

$$\mathrm{GrowthSkill} = 1 - \frac{\Delta\mathrm{MAE}_{\mathrm{model}}}{\Delta\mathrm{MAE}_{\mathrm{persist}}}.$$

Persistence has GrowthSkill $= 0$ by construction. Positive values indicate that the model predicts future abundance changes more accurately than assuming no compositional change, while negative values indicate larger abundance-change errors than Persistence. A value closer to 1 indicates increasingly accurate recovery of the observed expansion magnitude.

Expansion average precision (ExpnAP). Expansion average precision evaluates whether the model can identify which initially rare populations will subsequently expand. Let p_c^0 denote the observed proportion of cell type c at the final training snapshot, and let $p_{t,c}$ and $\hat{p}_{t,c}$ denote its observed and predicted proportions at held-out time t. We define the set of rare cell types at the forecasting origin as

$$\mathcal{R} = \left\{ c : 0 < p_c^0 \le 0.05 \right\}.$$

A rare type is considered truly expanding if its abundance increases by at least 0.5 percentage points and reaches at least a 1.5-fold increase at one or more held-out time points:

$$c \in \mathcal{E} \iff c \in \mathcal{R}, \qquad \max_{t\in\mathcal{T}_{\mathrm{test}}} \left(p_{t,c} - p_c^0\right) \ge 0.005, \qquad \max_{t\in\mathcal{T}_{\mathrm{test}}} \frac{p_{t,c}}{p_c^0} \ge 1.5.$$

For each rare type, we summarize the predicted expansion by its largest relative abundance increase over the forecasting period:

$$s_c^{\mathrm{pred}} = \max_{t\in\mathcal{T}_{\mathrm{test}}} \frac{\hat{p}_{t,c} - p_c^0}{\max(p_c^0, 0.002)}.$$

The lower bound in the denominator prevents extremely small initial proportions from producing disproportionately large expansion scores. We then assign each rare type the binary label

$$y_c = \mathbf{1}[c \in \mathcal{E}],$$

and compute average precision over the rare-type set:

$$\mathrm{ExpnAP} = \mathrm{AP}\left(\{y_c\}_{c\in\mathcal{R}}, \{s_c^{\mathrm{pred}}\}_{c\in\mathcal{R}}\right).$$

Higher ExpnAP indicates that rare populations that truly expand in the future are ranked above rare populations that remain stable or decline. ExpnAP is undefined when all evaluated rare types belong to the same class.

Abundance-direction accuracy (Dir.). Dir. evaluates whether the model predicts the correct direction of abundance change for initially rare cell types. Let $\mathcal{R} = \{c : 0 < p_c^0 \le 0.05\}$ denote the set of cell types that are rare at the final training snapshot, and let $\mathcal{T}_{\mathrm{test}}$ denote the held-out future time points. For cell type c and future time point t, we define the observed and predicted abundance changes relative to the final training composition as

$$\Delta p_{t,c} = p_{t,c} - p_c^0, \qquad \widehat{\Delta p}_{t,c} = \hat{p}_{t,c} - p_c^0.$$

The abundance-direction accuracy is

$$\mathrm{DirAcc} = \frac{1}{|\mathcal{R}|\,|\mathcal{T}_{\mathrm{test}}|} \sum_{c\in\mathcal{R}} \sum_{t\in\mathcal{T}_{\mathrm{test}}} \mathbf{1}\left[\mathrm{sign}\left(\widehat{\Delta p}_{t,c}\right) = \mathrm{sign}(\Delta p_{t,c})\right],$$

where $\mathrm{sign}(x) \in \{-1, 0, +1\}$ denotes a decrease, no change, or an increase, respectively. Unlike the noise-aware Macro-F1 defined below, Dir. applies no dead zone and therefore evaluates the literal sign of the estimated abundance change. Higher values indicate more accurate prediction of whether rare populations increase or decrease.

Noise-aware abundance-direction Macro-F1. Because small fluctuations in estimated cell-type proportions may arise from finite-sample generation and post-hoc label transfer, we additionally evaluate abundance direction using a three-class dead zone. We set $\delta = 0.005$, corresponding to 0.5 percentage points, and assign the observed direction label

$$y_{t,c} = \begin{cases} +1, & \Delta p_{t,c} > \delta, \\ 0, & |\Delta p_{t,c}| \le \delta, \\ -1, & \Delta p_{t,c} < -\delta, \end{cases}$$

with the predicted label defined analogously:

$$\hat{y}_{t,c} = \begin{cases} +1, & \widehat{\Delta p}_{t,c} > \delta, \\ 0, & |\widehat{\Delta p}_{t,c}| \le \delta, \\ -1, & \widehat{\Delta p}_{t,c} < -\delta. \end{cases}$$

The three classes therefore represent increasing, stable, and decreasing cell-type abundance. For each class $a \in \{-1, 0, +1\}$, we compute

$$\mathrm{Precision}_a = \frac{\mathrm{TP}_a}{\mathrm{TP}_a + \mathrm{FP}_a}, \qquad \mathrm{Recall}_a = \frac{\mathrm{TP}_a}{\mathrm{TP}_a + \mathrm{FN}_a},$$

and

$$F_{1,a} = \frac{2\,\mathrm{Precision}_a\,\mathrm{Recall}_a}{\mathrm{Precision}_a + \mathrm{Recall}_a}.$$

The reported score is the unweighted average over the three direction classes:

$$\text{Macro-F1} = \frac{1}{3} \sum_{a\in\{-1,0,+1\}} F_{1,a}.$$

A class with no predicted or ground-truth positive examples is assigned an F_1 score of zero. Macro-F1 gives equal weight to increasing, stable, and decreasing populations and is therefore less dominated by the most frequent direction class than ordinary accuracy.

A.2.3 DETAILS OF RARE AND EMERGING POPULATION FORECASTING

Motivation. Existing evaluations of temporal single-cell forecasting often equate success with matching the overall future distribution. Yet the biologically consequential changes in development and disease are frequently driven by the selective expansion of rare states and the emergence of previously unobserved populations. Population-level metrics such as Wasserstein distance and MMD are weighted toward abundant states and may barely penalize the omission of small populations; a model can therefore obtain competitive scores simply by preserving the dominant structure of the last observed snapshot. Current evaluations consequently struggle to distinguish genuine learning of future dynamics from conservative maintenance of existing distributions. To directly assess this capability, we construct a three-tier evaluation ladder that progressively reduces the evidence available about the target population during training.

Tasks. Tier 1 evaluates **rare-state expansion**. A target type is present in the last training snapshot at low abundance ($\leq 5\%$), and the model must predict whether it will expand, the direction of its abundance change, and the magnitude of that change. A rare type is considered expanding if its abundance increases by at least 0.5 percentage points with a fold change of at least 1.5 at a held-out future time point.

Tier 2 evaluates **natural emergence**. The target population is absent from every training snapshot but appears naturally during the held-out future. Because a future-only annotation may still overlap the historical expression manifold, we additionally require generated cells to leave the support of the training population and enter the support occupied by a real future-only population.

Tier 3 evaluates **controlled masked emergence**. A target type that was originally observed during training is removed from every training snapshot, while its biologically related precursor, transitional, or lineage-associated populations are retained. The model never observes the target cells or their label. This creates a counterfactual test of whether the dynamics of the retained populations are sufficient to re-compose the hidden state.

Datasets and protocol. We evaluate the three tiers on CD embryonic heart and BetaDiff development. CD-short, trained through E15.5 and evaluated on E16.5 and E17.5, contains repeated observed composition transitions and is the primary test of calibrated rare-state expansion. BetaDiff is evaluated across all three tiers. In Tier 1, `sst_hhex` is observed at low abundance before expanding at S5c/S6c. In Tier 2, `foxj1`, `sc_beta`, `sc_ec`, and `exo` are absent from S3c/S4c and appear naturally during the held-out stages. In Tier 3, `sc_alpha` and `sst_hhex` are removed from all training snapshots. CD `vascular_ec`/`immune` masking provides additional controlled-emergence boundary cases. The retained populations in Tier 3 were verified using transcriptional proximity and established developmental relationships to the masked targets. Masked targets were selected to satisfy three criteria: the target was originally present in the training snapshots, remained sufficiently represented at the held-out stages for stable evaluation, and retained transcriptionally or developmentally related populations after masking. We additionally selected targets spanning distinct support and abundance regimes. In BetaDiff, `sc_alpha` represents an abundant state embedded in a densely sampled endocrine trajectory, whereas `sst_hhex` represents a low-abundance state supported by a much narrower developmental neighborhood. In CD, `vascular_ec` and `immune` provide harder boundary cases in which related populations remain observable but the approach toward the target state is weakly represented after masking. This selection therefore tests masked-state recovery across both favorable and weak-support regimes.

All methods generate 2,000 cells per future time point and are evaluated over five seeds (11, 23, 42, 67, and 89). Persistence resamples the last observed snapshot, whereas Noise-halo adds locally calibrated noise around it. These controls distinguish learned dynamics from doing nothing and from undirected support expansion.

Metrics. Across all tiers, the evaluation asks four related questions: whether the target is identified, whether its change or location is correct, whether its future abundance is calibrated, and whether the generated cells represent directed extrapolation rather than copying or undirected noise.

For Tier 1, identification is measured by expansion average precision (AP), direction by sign accuracy and a noise-aware three-way Macro-F1, and magnitude by abundance-change error ΔMAE

and GrowthSkill. Persistence therefore has zero GrowthSkill by construction, and a positive value indicates improvement over predicting no compositional change.

For Tier 2, a generated cell is *accepted* when it lies outside historical support but inside the support of a real future-only population. An outside-support cell that does not reach any such population is *unsupported*. We report accepted and unsupported fractions, emergence precision $\mathrm{Accepted}/(\mathrm{Accepted} + \mathrm{Unsupported})$, and population-level detection.

For Tier 3, we report target detection, predicted abundance, conditional W_2 between generated and real target cells, and out-of-support target precision. Conditional W_2 is defined only when the target is detected and is therefore always interpreted together with the number of successful seed–time combinations.

False-positive floor of support estimation. Historical support is estimated from finite samples using neighborhood distances in a fixed low-dimensional representation. Because sampling density varies across developmental stages and cellular states, genuine historical cells in sparse or boundary regions may occasionally be classified as outside the estimated historical support. Consistent with this limitation, Persistence, which only resamples the last observed population, obtains a non-zero accepted fraction. Persistence therefore defines an empirical false-positive floor for the support-based evaluator.

We consequently interpret Tier 2 comparatively rather than treating every accepted cell as definitive evidence of a de novo biological state. All methods are evaluated using the same representation, support threshold, future-state reference, and generated-cell budget. Persistence measures finite-sample support-estimation error, whereas Noise-halo tests whether undirected local variation is sufficient to reach future-emergent regions. Under this shared evaluator, scHOMURA achieves higher emergence precision than the learned baselines and both trivial controls, indicating more selective recovery of future-emergent expression regions.

A.3 APPENDIX TABLES

Table S1: Forecasting splits used in the experiments.

Setting	Training snapshots	Test snapshots
CD-short	E9.5, E10.5, E11.5, E12.5, E13.5, E14.5, E15.5	E16.5, E17.5
CD-long	E9.5, E10.5, E11.5	E12.5, E13.5, E14.5, E15.5, E16.5, E17.5, E18
CR-short	D0, D1, D2, D3, D4, D5, D6, D7, D8, D9, D10, D11, D12, D13, D14, D15	D16, D17, D18
CR-long	D0, D1, D2, D3, D4, D5	D6, D7, D8, D9, D10, D11, D12, D13, D14, D15, D16, D17, D18
Zebrafish	High, Oblong, Dome, 30%-epiboly, 50%-epiboly, Shield, 60%-epiboly, 75%-epiboly, 90%-epiboly, Bud	3-somite, 6-somite
BetaDiff	S3c, S4c	S5c, S6c
Retinal	D10, D20, D33, D45, D60, D90	D150, D210

Table S2: Benchmark performance on the remaining four forecasting splits. Protocol and notation follow Table 1. "—" indicates that cell-type annotations are unavailable.

Method	$W_2 \downarrow$	DE-AP↑	DE-dir↑	$\Delta r \uparrow$	MMD↓	Variogram↓	Type-JS↓
CD-short							
scHOMURA	**15.09 ± 0.16**	0.263 ± 0.085	0.575 ± 0.118	0.261 ± 0.059	**0.044 ± 0.008**	0.0142 ± 0.0014	**0.013 ± 0.005**
LGP-OT	16.03 ± 0.21	0.240 ± 0.050	0.674 ± 0.058	0.294 ± 0.080	0.053 ± 0.007	**0.0114 ± 0.0006**	0.038 ± 0.009
scNODE	17.52 ± 0.24	0.139 ± 0.030	0.560 ± 0.106	−0.101 ± 0.095	0.053 ± 0.006	0.0189 ± 0.0025	0.069 ± 0.004
MIOFlow	17.34 ± 0.44	0.189 ± 0.029	0.615 ± 0.088	0.096 ± 0.108	0.045 ± 0.007	0.0249 ± 0.0039	0.047 ± 0.009
PRESCIENT	18.40 ± 0.17	0.100 ± 0.009	0.350 ± 0.042	−0.291 ± 0.036	0.164 ± 0.009	0.0423 ± 0.0015	0.096 ± 0.016
VGFM	18.42 ± 0.23	**0.297 ± 0.100**	**0.714 ± 0.192**	**0.300 ± 0.153**	0.054 ± 0.010	0.0220 ± 0.0016	0.094 ± 0.028
CR-long							
scHOMURA	**17.80 ± 0.24**	0.392 ± 0.027	**0.760 ± 0.074**	0.338 ± 0.133	0.304 ± 0.009	**0.0343 ± 0.0009**	—
LGP-OT	19.02 ± 1.16	**0.451 ± 0.012**	0.673 ± 0.029	**0.396 ± 0.021**	**0.227 ± 0.020**	0.0404 ± 0.0039	—
scNODE	25.48 ± 1.71	0.415 ± 0.014	0.408 ± 0.023	−0.225 ± 0.037	0.410 ± 0.051	0.1202 ± 0.0144	—
MIOFlow	28.63 ± 7.69	0.420 ± 0.015	0.518 ± 0.116	−0.043 ± 0.186	0.497 ± 0.143	0.1382 ± 0.0671	—
PRESCIENT	18.20 ± 0.10	0.402 ± 0.007	0.728 ± 0.012	0.144 ± 0.036	0.318 ± 0.006	0.0442 ± 0.0027	—
VGFM	41.80 ± 3.54	0.363 ± 0.037	0.576 ± 0.056	−0.204 ± 0.046	0.479 ± 0.098	0.2661 ± 0.0624	—
Zebrafish							
scHOMURA	**34.26 ± 0.09**	0.452 ± 0.031	**0.759 ± 0.075**	**0.416 ± 0.078**	0.109 ± 0.011	0.0642 ± 0.0102	—
LGP-OT	35.04 ± 0.33	**0.496 ± 0.046**	0.730 ± 0.057	0.355 ± 0.092	0.118 ± 0.010	**0.0547 ± 0.0105**	—
scNODE	35.70 ± 0.06	0.383 ± 0.038	0.573 ± 0.012	0.137 ± 0.029	0.109 ± 0.006	0.1076 ± 0.0047	—
MIOFlow	35.97 ± 0.30	0.375 ± 0.040	0.607 ± 0.101	0.170 ± 0.155	**0.105 ± 0.018**	0.0834 ± 0.0170	—
PRESCIENT	46.79 ± 1.63	0.491 ± 0.033	0.238 ± 0.007	−0.456 ± 0.020	0.473 ± 0.037	0.3582 ± 0.0330	—
VGFM	39.25 ± 0.14	0.287 ± 0.021	0.737 ± 0.031	0.361 ± 0.021	0.246 ± 0.021	0.1559 ± 0.0036	—
Retinal							
scHOMURA	**43.22 ± 0.62**	0.497 ± 0.039	0.475 ± 0.050	0.022 ± 0.070	**0.201 ± 0.020**	0.0838 ± 0.0241	0.141 ± 0.015
LGP-OT	46.94 ± 0.20	0.517 ± 0.003	0.373 ± 0.015	−0.004 ± 0.013	0.480 ± 0.008	**0.0744 ± 0.0065**	0.312 ± 0.004
scNODE	49.09 ± 2.36	**0.560 ± 0.022**	0.407 ± 0.032	−0.088 ± 0.043	0.294 ± 0.041	0.1978 ± 0.0386	0.215 ± 0.019
MIOFlow	45.94 ± 2.57	0.448 ± 0.053	0.408 ± 0.054	−0.205 ± 0.101	0.266 ± 0.073	0.1765 ± 0.0534	0.200 ± 0.083
PRESCIENT	49.48 ± 0.13	0.402 ± 0.009	**0.670 ± 0.010**	**0.236 ± 0.008**	0.351 ± 0.008	0.3114 ± 0.0031	0.270 ± 0.006
VGFM	60.57 ± 6.18	0.495 ± 0.017	0.567 ± 0.039	0.087 ± 0.036	0.222 ± 0.080	0.1997 ± 0.0556	**0.135 ± 0.013**

Table S3: Aggregate ranking summary across all forecasting split–metric combinations shown in Table 1 and S2. Average rank is computed across all available split–metric cells; wins count the number of rank-1 results.

Method	scHOMURA	LGP-OT	scNODE	MIOFlow	PRESCIENT	VGFM
Avg. rank ↓	**1.67**	2.63	3.76	3.87	4.76	4.30
Wins ↑	**27**	10	2	1	2	4

Table S4: Rare-state expansion on BetaDiff. Models evaluated for the rare `sst_hhex` population. Values are reported as mean ± SD over five seeds. **Bold** and underline indicate the best and second-best results, respectively. Expansion AP is a ranking-based metric and is not defined for BetaDiff, where only one cell type, `sst_hhex`, has a frequency $\leq 5\%$ at the last training time point.

Method	Dir.↑	Macro-F1↑	ΔMAE↓	GrowthSkill↑
scHOMURA	**0.900±0.224**	0.133±0.122	0.017±0.007	−0.585±0.468
LGP-OT	0.300±0.447	0.178±0.099	0.013±0.006	−0.252±0.622
scNODE	0.000±0.000	0.000±0.000	0.019±0.000	−0.827±0.159
MIOFlow	0.000±0.000	0.000±0.000	0.019±0.000	−0.827±0.159
PRESCIENT	0.000±0.000	0.000±0.000	0.019±0.000	−0.827±0.159
VGFM	0.000±0.000	0.000±0.000	0.019±0.000	−0.827±0.159
Persistence	0.700±0.274	0.222±0.000	**0.011±0.001**	**0 (by def.)**
Noise-halo	0.100±0.224	**0.267±0.061**	0.015±0.002	−0.385±0.098

Table S5: Controlled masked emergence on CD-short. Neither vascular endothelial cells (ECs) nor immune cells are detected in any seed–time combination. Global W_2 and Type JS therefore assess the quality of the remaining population forecast rather than successful masked-state recovery. Values are reported as mean ± SD over five seeds. **Bold** and underline indicate the best and second-best results, respectively.

	Vascular ECs (true 5.1–8.0%)		Immune cells (true 4.2–7.4%)	
Method	Global W_2 ↓	Type JS↓	Global W_2 ↓	Type JS↓
scHOMURA	**11.23±0.89**	**0.035±0.013**	**10.91±1.14**	**0.037±0.012**
LGP-OT	11.54±1.07	0.041±0.017	11.39±1.26	0.049±0.026
scNODE	13.66±1.22	0.074±0.019	13.99±1.30	0.079±0.024
MIOFlow	13.85±1.23	0.074±0.018	13.72±1.30	0.082±0.019
PRESCIENT	15.57±0.94	0.174±0.034	14.75±1.07	0.151±0.036
VGFM	12.91±1.55	0.100±0.054	12.85±1.00	0.141±0.060
Persistence	12.41±1.42	0.048±0.020	12.20±1.39	0.047±0.021
Noise-halo	12.38±1.33	0.048±0.020	12.21±1.30	0.047±0.021

Table S6: Additional component ablation results on the two unannotated forecasting splits. Values are reported as mean ± SD over five seeds (11, 23, 42, 67, and 89). For each seed, metrics are first averaged across held-out target time points; the reported statistics are then computed across the five seed-level averages. Type JS is omitted because these datasets do not provide cell-type annotations. **Bold** indicates performance degraded by at least 40% relative to the full model.

Dataset	Variant	W_2 ↓	DE-AP↑	DE-dir↑	Δr ↑	MMD↓	Variogram↓
CR-long	**Full (scHOMURA)**	17.80±0.24	0.392±0.027	0.760±0.074	0.338±0.133	0.304±0.009	0.0343±0.0009
	w/o residual flow	18.46±0.41	0.404±0.009	0.555±0.116	**0.083±0.168**	0.339±0.018	0.0381±0.0019
	w/o multi-horizon	18.07±0.17	0.371±0.017	0.712±0.056	0.257±0.093	0.325±0.013	0.0347±0.0013
	w/o growth modeling	17.71±0.17	0.392±0.024	0.773±0.047	0.372±0.101	0.297±0.006	0.0331±0.0011
Zebrafish	**Full (scHOMURA)**	34.26±0.09	0.452±0.031	0.759±0.075	0.416±0.078	0.109±0.011	0.0642±0.0102
	w/o residual flow	34.08±0.05	0.474±0.022	0.778±0.013	0.475±0.014	0.116±0.008	0.0584±0.0049
	w/o multi-horizon	34.28±0.21	0.458±0.014	0.765±0.049	0.411±0.063	0.115±0.014	0.0592±0.0059
	w/o growth modeling	34.26±0.08	0.447±0.033	0.756±0.076	0.412±0.080	0.109±0.012	0.0652±0.0101

A.4 APPENDIX FIGURES

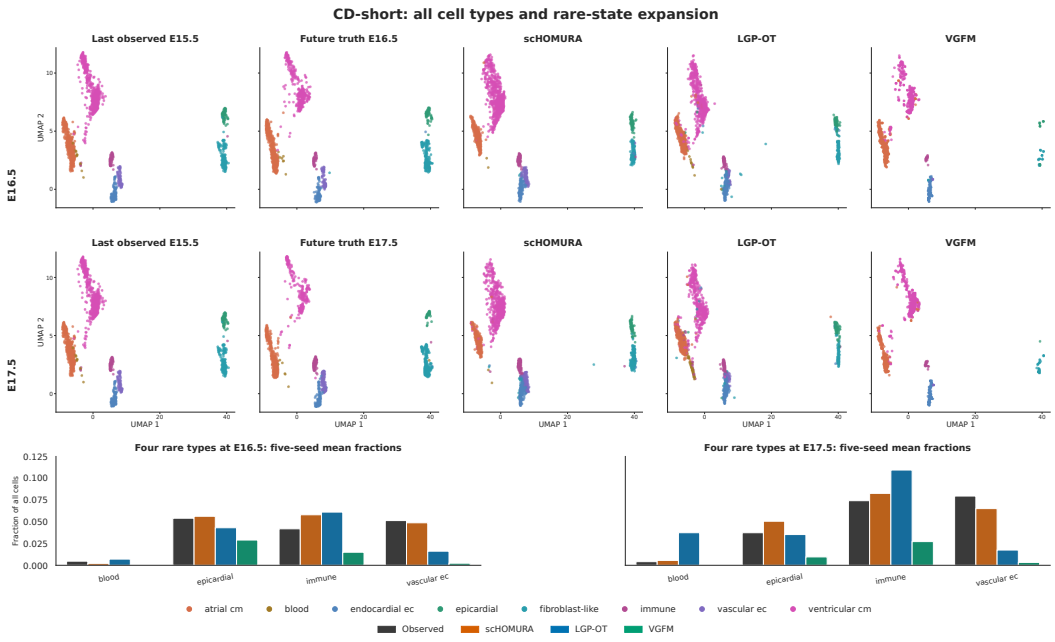


Figure S1: Rare-state expansion on CD-short. UMAPs compare observed and predicted future populations, with rare-cell-type fractions shown at E16.5 and E17.5.

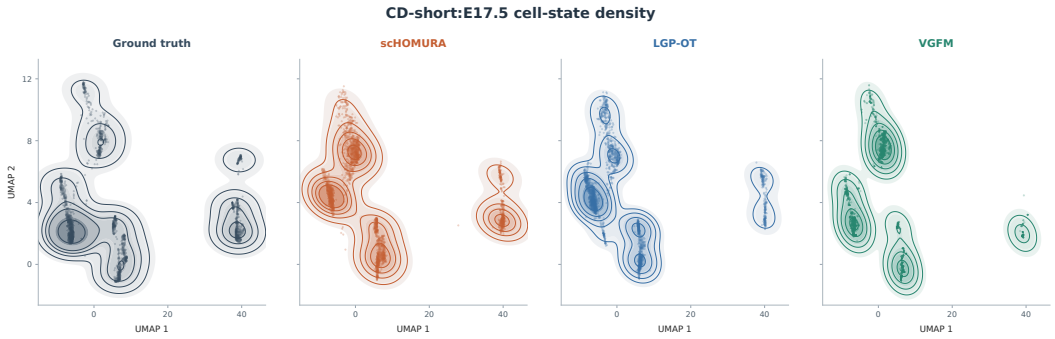


Figure S2: Cell-state density at E17.5 on CD-short.