Supplementary Information (S1)

This document provides additional methods, mathematical formulas, hyperparameters, and supporting results for the main text. It includes Supplementary Figures S1–S15 and Supplementary Tables S1–S3, which are cited within this Supplementary Information.

1 Supporting analysis of missing-value initialization for the *Drosophila* atlas

To assess how the initial placeholder choice influences downstream VAE reconstruction, we compared three initialization schemes for missing gene entries in the *Drosophila* atlas: zero-fill, mean-fill, and random-fill. These schemes were evaluated under two VAE objectives: the full objective, which included reconstruction, KL regularization, smooth sparsity, decoder smoothness, and directional alignment; and a reduced objective omitting the decoder-smoothness and directional-alignment terms.

Fig. S7 summarizes test-set reconstruction performance for the 99-gene decoder as a function of latent dimensionality. Across both training objectives, random-fill initialization consistently yielded the lowest mean absolute error (MAE), the lowest mean squared error (MSE), and the highest mean coefficient of determination (R^2), whereas zero-fill performed worst. Based on this comparison, random-fill was used as the default initialization scheme in subsequent *Drosophila* experiments.

2 Supplementary analyses of teacher-guided mechanistic fitting

This section provides additional qualitative analyses supporting the teacher-guided versus data-only comparisons reported in the main text. The main quantitative comparison is summarized in Table 1 of the main manuscript, where teacher-guided fits achieve equal or lower one-step forecast error after calibration, with the clearest advantage under stronger sparsity.

Here we focus on complementary visualizations that clarify how the black-box teacher shapes downstream mechanistic fitting. Fig. S8 shows two representative black-box gene-space realizations over each interval $[t_k, t_{k+1}]$: an anchored pushforward realization and a decoded latent-path realization. For this figure only, the decoded-path realization was obtained by linear interpolation in latent space between consecutive encoded points before

decoding, in order to isolate the contribution of the decoder separately from full latent-dynamics integration.

Fig. S9 compares mean absolute prediction error across the 27 fully observed genes between the black-box model, the Hill model, and previously reported erf-weighted LAD baselines. Although the LAD quadratic model achieves the lowest raw error, the Hill model offers the complementary advantage of explicit mechanistic nonlinearity and equation-level editability.

Finally, Fig. S10 overlays black-box and Hill-model trajectories directly in the full 99-dimensional gene space. Solid blue dots denote measured values and hollow blue circles denote VAE-imputed values. Blue curves correspond to black-box teacher trajectories, whereas red curves correspond to the fitted Hill-model trajectories. This figure provides a qualitative view of how well the mechanistic student reproduces the learned black-box dynamics in the full partially observed setting.

3 Regional robustness of mechanistic fits across the embryo

This section provides additional analyses of how spatial training coverage affects mechanistic generalization across the embryo. We partitioned the embryo either along the anterior–posterior axis into anterior, middle, and posterior subsets, or along the dorsal–ventral axis into dorsal, medial, and ventral subsets. These regional analyses parallel earlier region-restricted studies, but here are applied to Hill fields fitted from latent-derived teacher velocities.

For consistency across regions, we did not retrain a separate latent neural ODE for each case. Instead, regional Hill models were trained from a common source of black-box-derived gene-space velocities obtained from a fixed VAE pushforward with trivial latent interpolation. This design ensured that differences across regional fits reflected spatial training coverage rather than differences in upstream representation learning.

As a baseline, we also trained random-30% models by uniformly sampling 30% of bins from the training pool, matching the effective training size of each regional partition. To reduce variance, the random-30% sampling was repeated with 10 different seeds and the resulting errors were averaged. Prediction quality was summarized by the relative error

$$\frac{|g - \hat{g}|}{|g + \hat{g}| + \varepsilon},$$

where g is the ground-truth expression and $\hat{g}$ is the Hill-model prediction. This normalization reduced domination by highly expressed genes and made error magnitudes more comparable across genes.

Fig. S12 and Fig. S13 show that balanced spatial coverage improves generalization, with the benefit becoming clearer as sparsity is relaxed. Under stronger regularization, all regional models perform broadly similarly, but the random-30% baseline still provides a small advantage. Under weaker regularization, that advantage becomes larger, indicating that broad spatial sampling is especially helpful when the fitted network retains more couplings.

The spatial overlays in Fig. S14 show that region-restricted supervision produces non-uniform error patterns across the embryo. By contrast, the random-30% baseline gives a more spatially uniform and consistently lower error profile, particularly in posterior and ventral domains where single-region fits tend to overfit. This improvement is not confined to a few genes or locations, but extends across the 27-gene set and the full embryo, supporting the value of balanced spatial coverage during mechanistic distillation.

Supplemental Figures and Tables

List of Figures

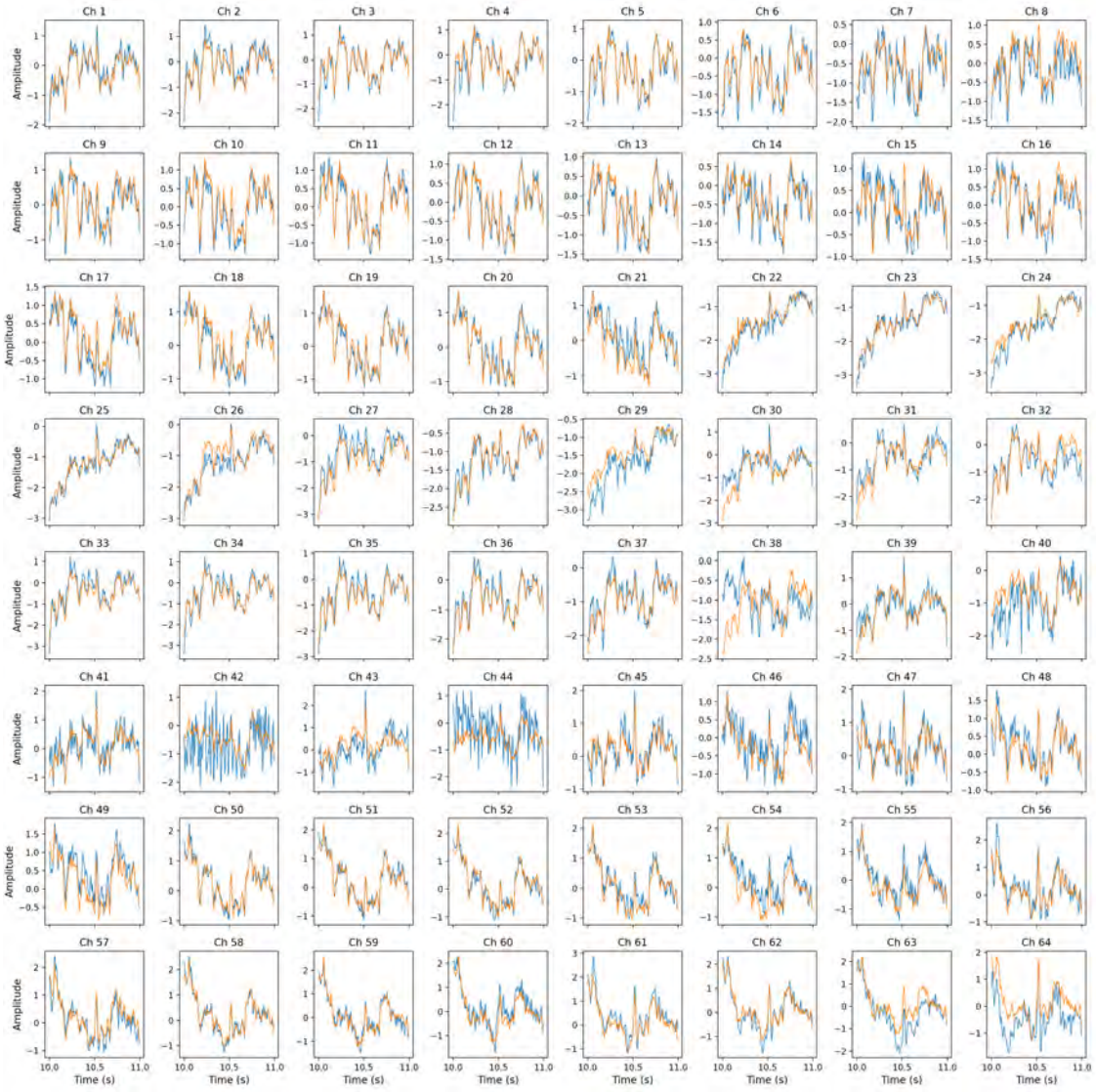

Figure S1: **Universal VAE reconstruction for a 1 s EEG segment (subject S001, run R01).** Each panel shows one of the 64 EEG channels (blue: recorded signal; orange: VAE reconstruction of the center sample in each 9-point window).

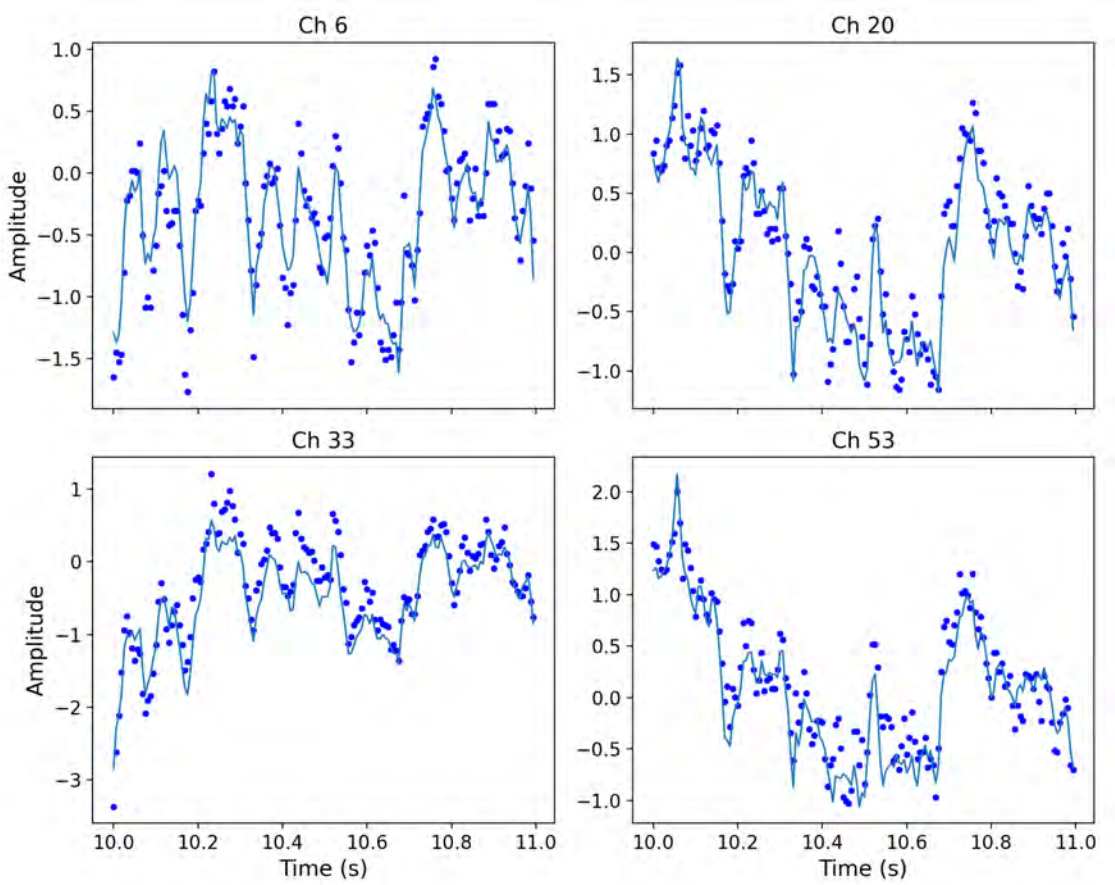

Figure S2: **Zoomed view of four EEG channels from Fig. S1.** Blue dots show measured EEG samples and blue curves show universal VAE center reconstructions, illustrating preservation of both slower trends and fast fluctuations.

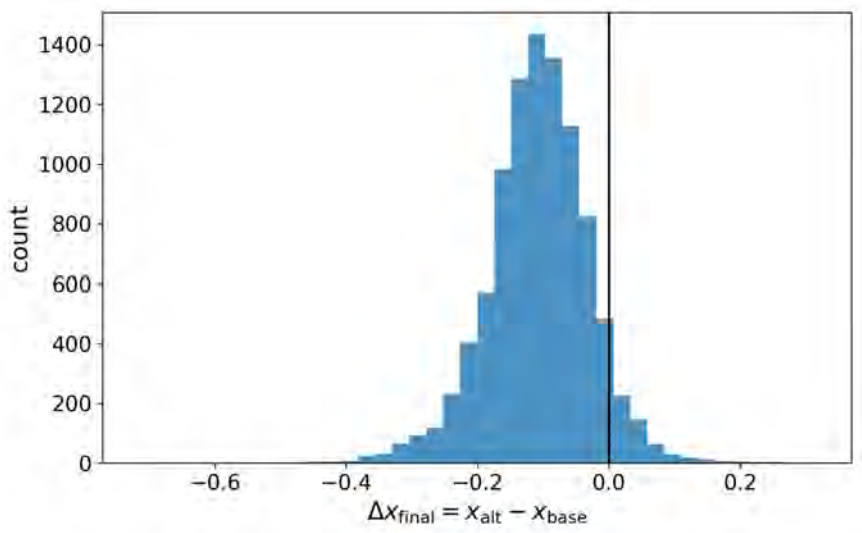

Figure S3: **In silico run-control perturbation in EEG.** For subject S033, we compare baseline run control R01 with a control-direction perturbation toward R03 while keeping the same initial states as in Fig. 4 of the main text. We compute the change in final predicted amplitude $\Delta x_{\text{final}} = x_{\text{alt}} - x_{\text{base}}$ after a 9-step rollout and aggregate across channels and windows. The distribution is shifted below zero (vertical line at 0), indicating a systematic decrease in predicted amplitude under the counterfactual control.

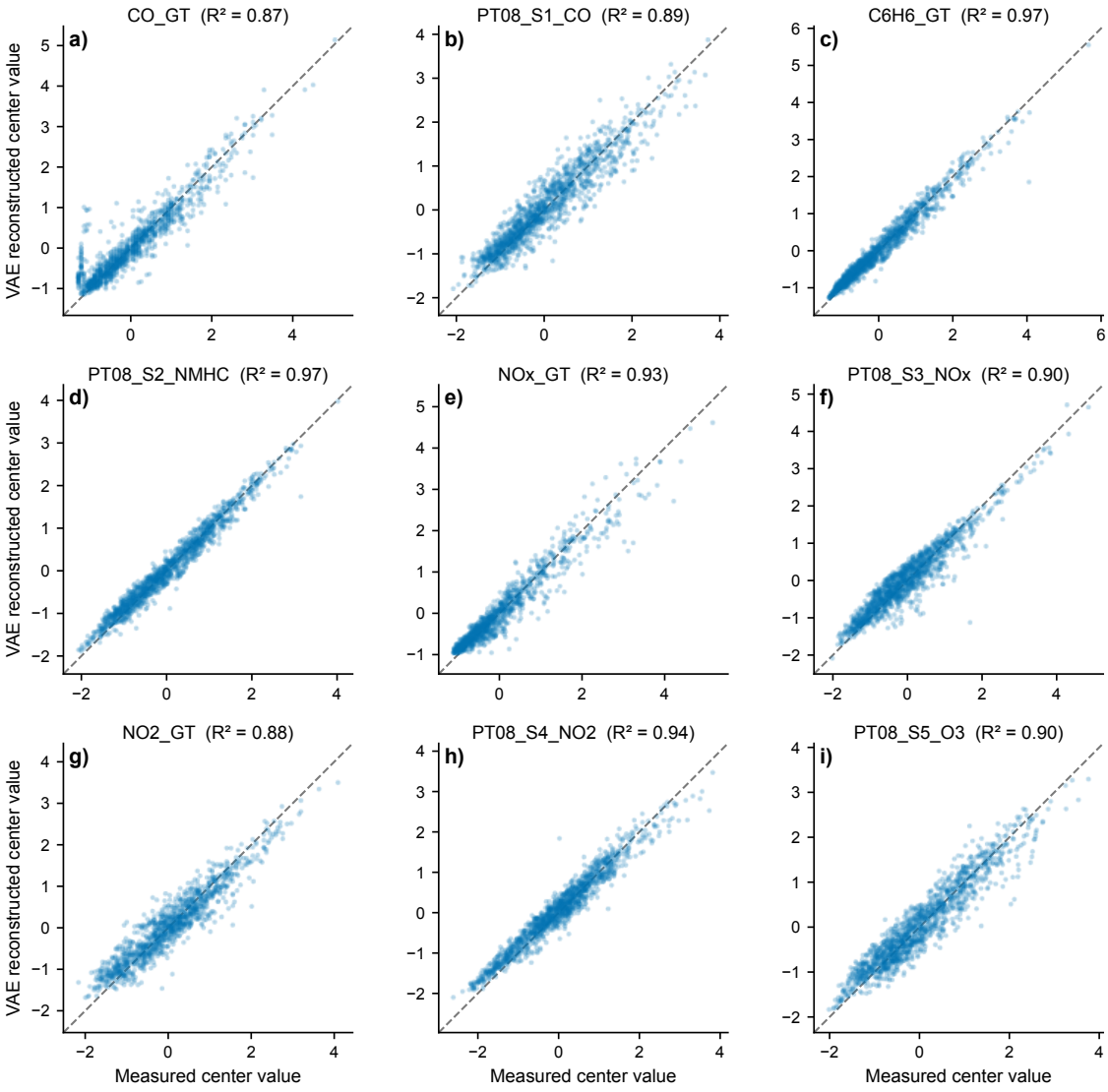

Figure S4: **AirQualityUCI VAE reconstruction scatter on a held-out test subset.** Each panel shows a pollutant variable. Points compare the measured center value (x-axis) to the VAE-reconstructed center value (y-axis) from 3-hour windows. The dashed line indicates $y = x$. Titles report coefficient of determination (R^2) values computed using measured entries only.

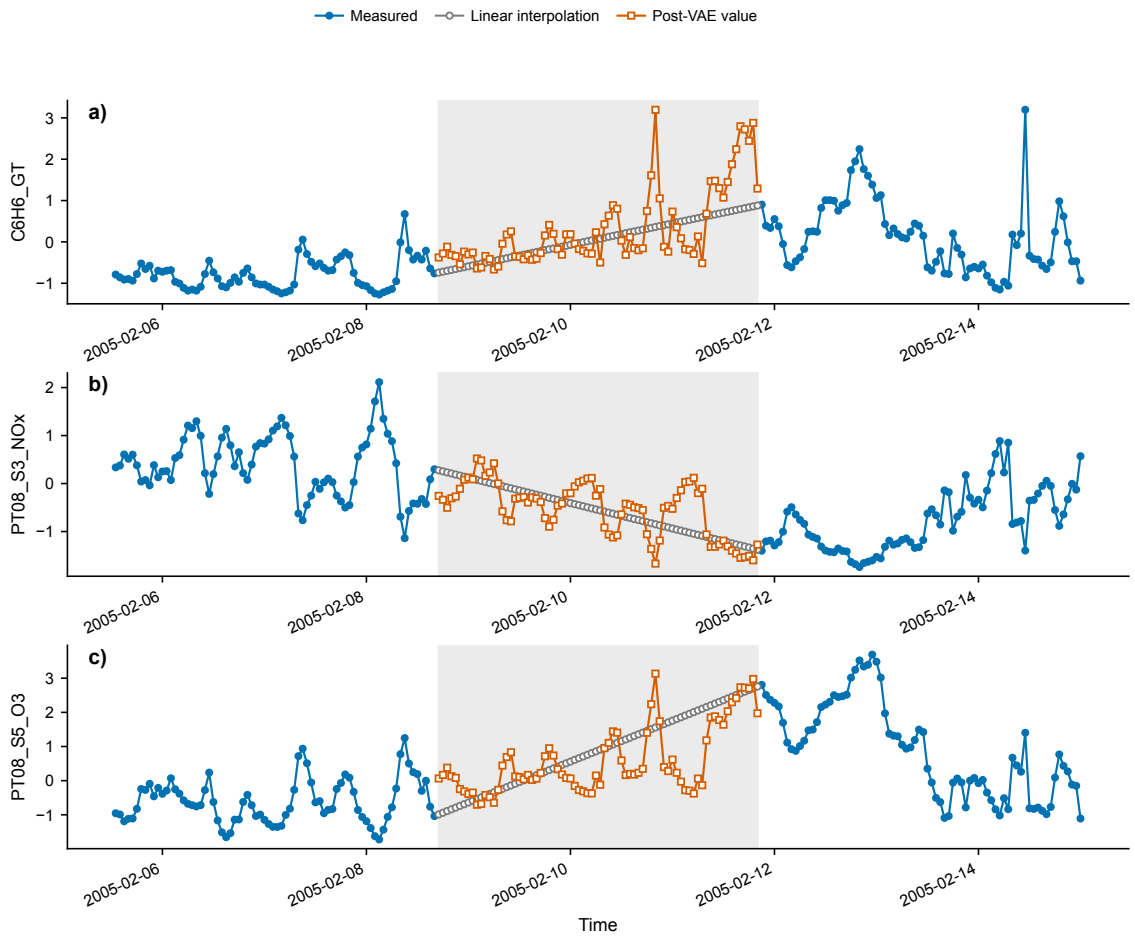

Figure S5: **Local comparison of linear pre-imputation versus post-VAE values within a missing gap.** For each variable (`C6H6_GT_`, `PT08_S3_NOx_`, `PT08_S5_O3_`), the plot shows a local time window centered around the largest missing gap (shaded). Blue: measured values. Black: linear-imputed values in the missing interval. Red: post-VAE values in the missing interval.

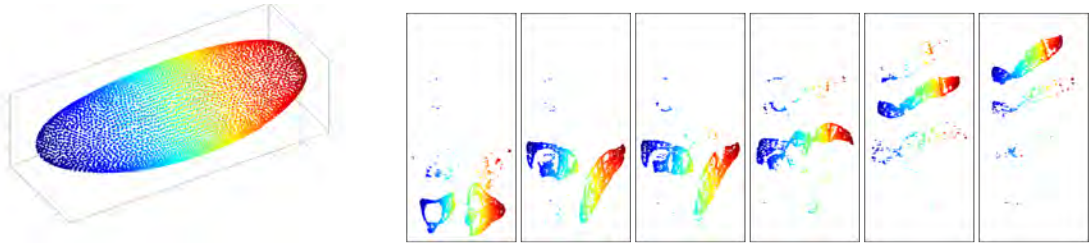

Figure S6: **Embryo layout and UMAP embedding of latent states.** Left: schematic of the blastoderm embryo with 6,078 spatial bins. Right: UMAP visualization of latent encodings across all bins, shown separately for the six developmental time points ($t = 0$ to $t = 5$, left to right). Red corresponds to posterior bins of the virtual embryo and blue corresponds to anterior bins.

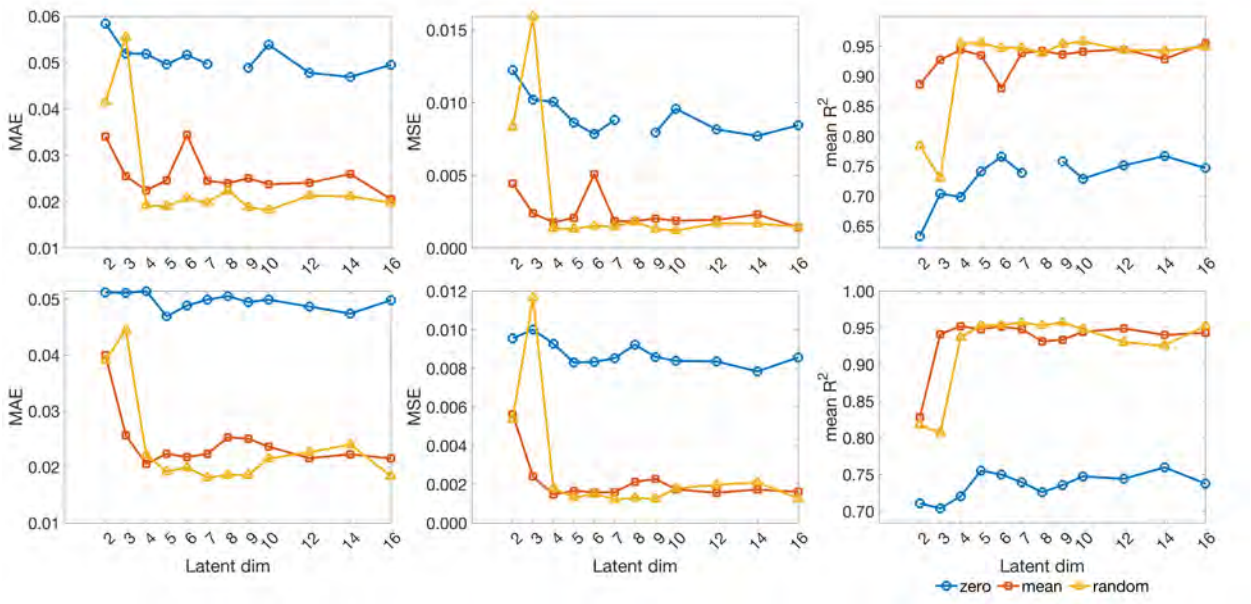

Figure S7: **VAE reconstruction accuracy versus latent dimension on the testing set.** Rows correspond to two training objectives: the full five-term loss (top) and a reduced objective omitting the decoder-smoothness and directional-alignment terms (bottom). Columns show mean absolute error (MAE), mean squared error (MSE), and mean coefficient of determination (R^2). Curves compare zero-fill, mean-fill, and random-fill initialization. Across latent dimensionalities, random-fill consistently gives the lowest error and highest R^2. The zero-initialization point at latent dimension 8 is an outlier and is omitted from the plot.

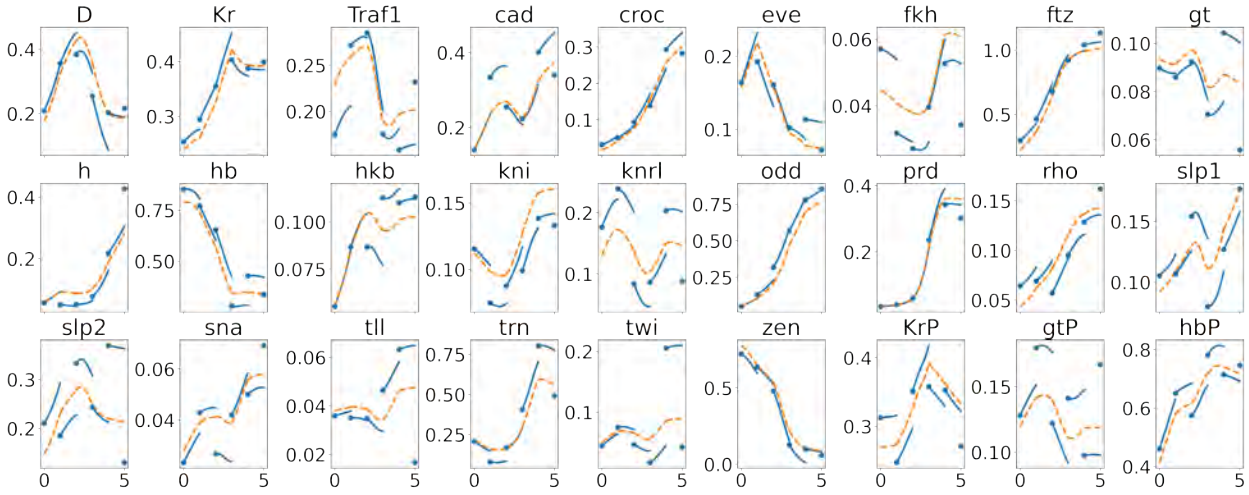

Figure S8: **Two black-box trajectory realizations in the 27-dimensional gene space for one representative cell.** The underlying VAE was trained with $\lambda_{\ell_1} = 10^{-4}$ (see Table 1 of the main text). Blue dots are observed expression at discrete times. Solid blue curves integrate in gene space starting from the observed data, eliminating reconstruction error at the segment start. Dashed red curves decode latent linear interpolations between neighboring encoded states, incurring reconstruction error at both ends of the curves. Gene names are shown above each subplot; segments are not connected across boundaries.

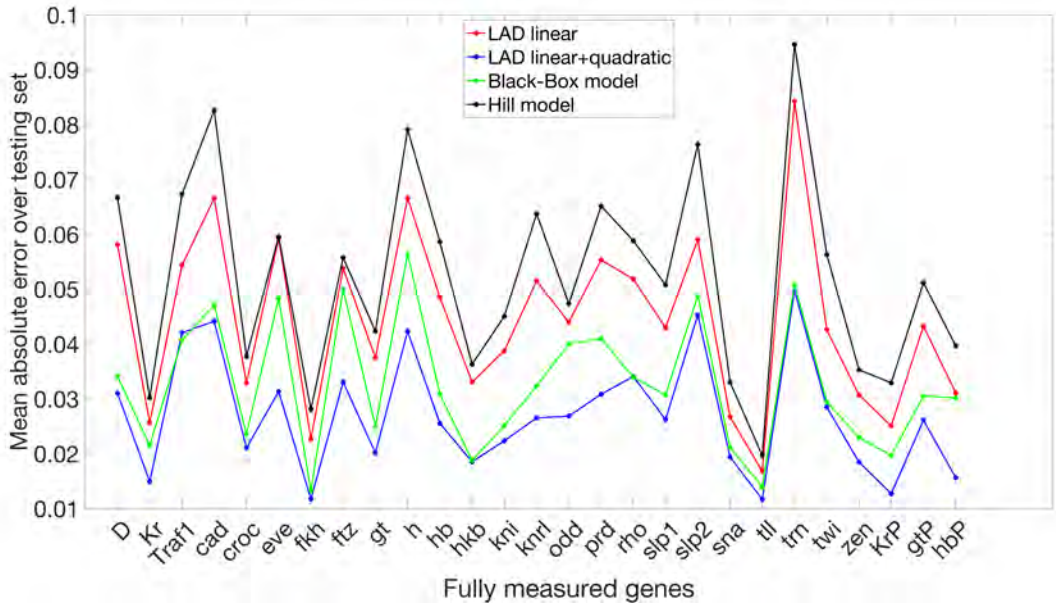

Figure S9: **Mean absolute prediction error on the testing set.** Comparison of mean absolute errors for our black-box model and the Hill model, alongside results reported in[1]. The connecting lines are included solely for visual guidance and do not carry interpretive meaning.

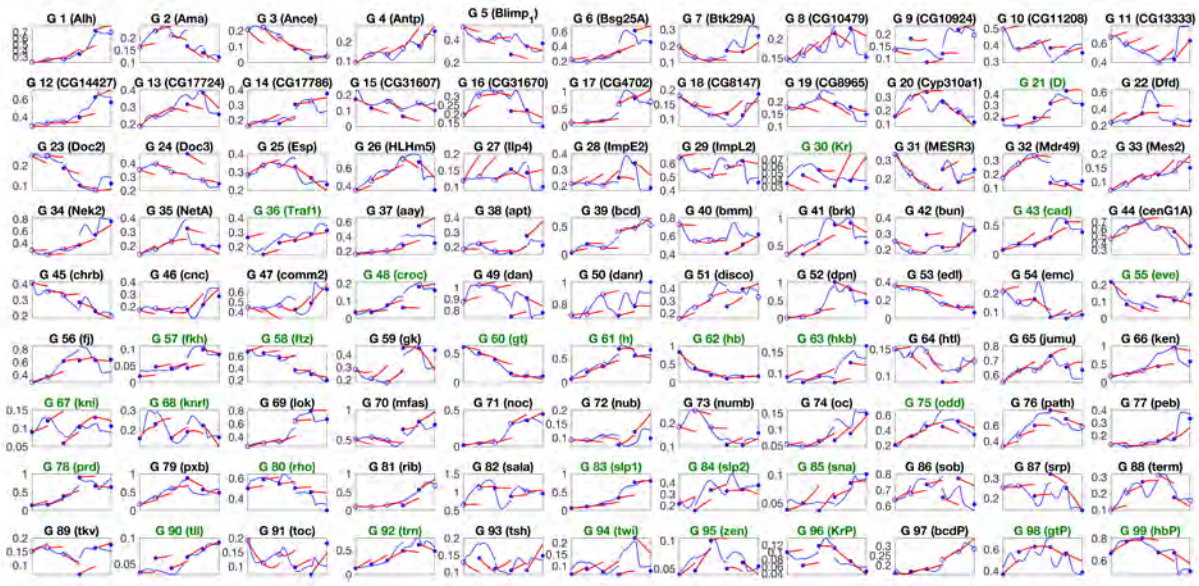

Figure S10: **Black-box versus Hill-field trajectories in 99-dimensional gene space.** Solid blue dots mark observed gene-expression measurements; hollow blue circles indicate values imputed by the VAE. Blue curves are the decoder applied to the latent neural ODE trajectories. Red curves are integrations of the corresponding Hill model trained from this black-box teacher. Subplot titles shown in green denote the 27 genes that are fully observed across all six time points.

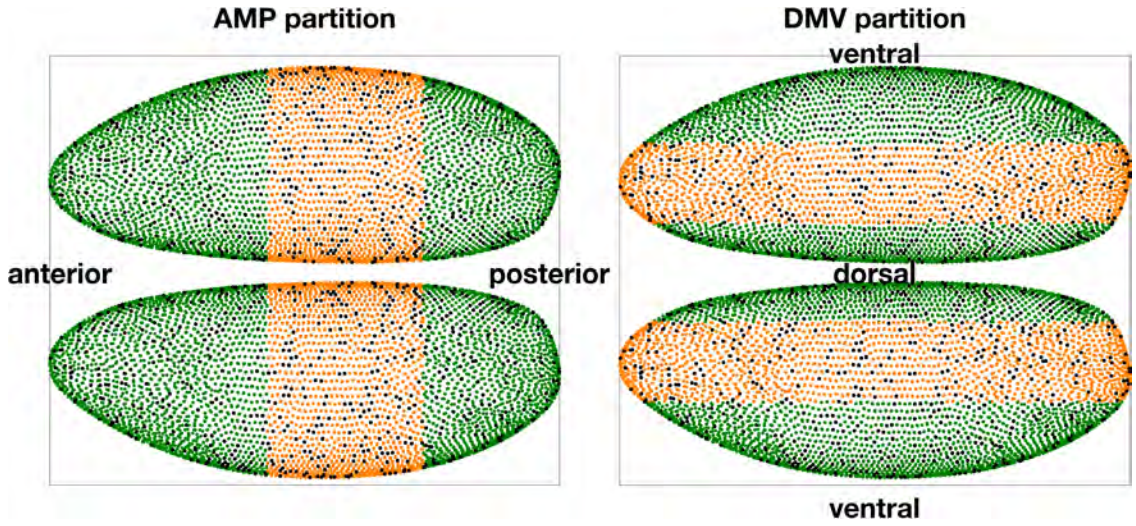

Figure S11: **Spatial partitioning of the embryo into anterior–middle–posterior (AMP) and dorsal–medial–ventral (DMV) subsets.** Cells are colored according to their regional assignment: anterior/posterior or dorsal/ventral in deep green, medial in orange. Black dots indicate the fixed 10% test cells that were excluded from training. These partitions define the subsets of cells used for region-specific Hill-model fitting. As a baseline, we also trained models on a random selection of 30% of bins uniformly distributed across the embryo (random-30%), matching the number of bins used in each regional subset.

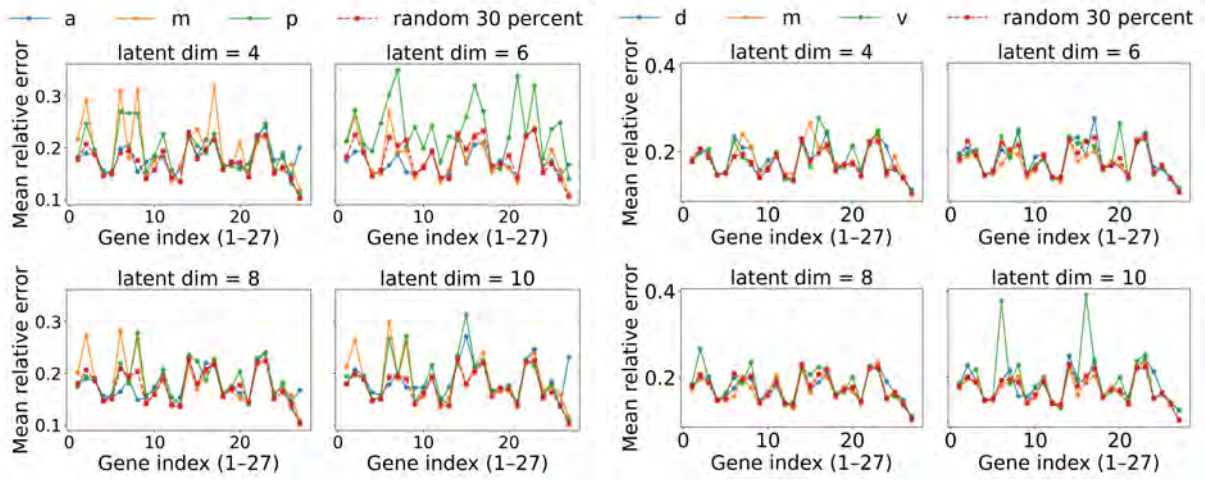

Figure S12: **Universality of Hill-model fitting across regions for $\lambda = 10^{-4}$ (stronger regularization).** Mean relative error per gene for models trained on subsets of cells from either the anterior–middle–posterior (AMP, left) or dorsal–medial–ventral (DMV, right) partition, compared to models trained on randomly selected 30% of bins across the embryo (random-30%). Under stronger sparsity, the random-30% baseline already provides a modest advantage, suggesting that balanced spatial coverage improves generalization even when couplings are heavily pruned.

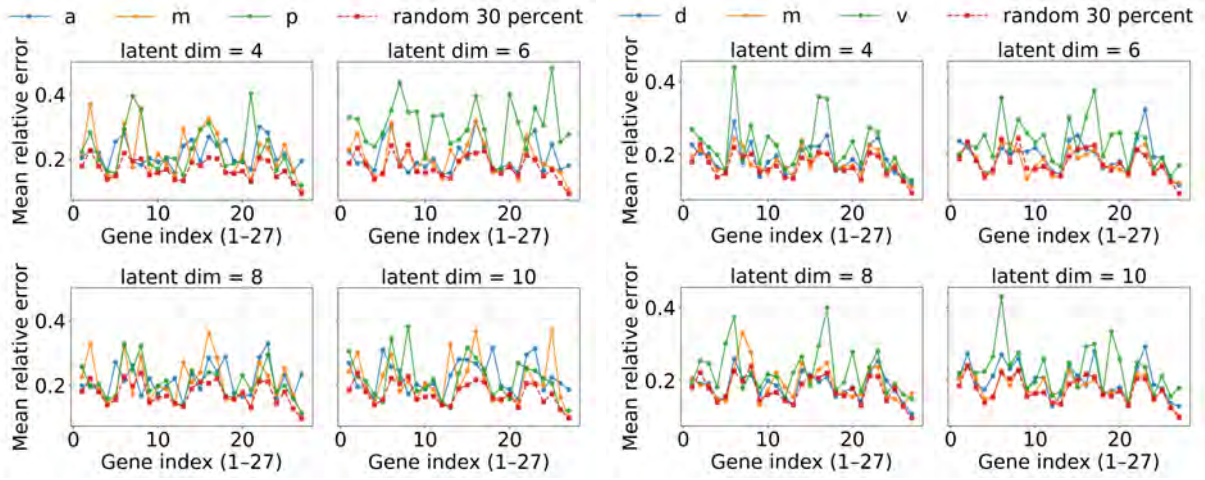

Figure S13: **Universality of Hill-model fitting across regions for $\lambda = 10^{-5}$ (weaker regularization).** Mean relative error per gene for models trained on subsets of cells from either the AMP (left) or DMV (right) partition, compared to random-30%. With weaker sparsity, the advantage of random-30% becomes more pronounced: balanced coverage consistently outperforms any single-region training, indicating that broader sampling is especially valuable when the network retains more couplings.

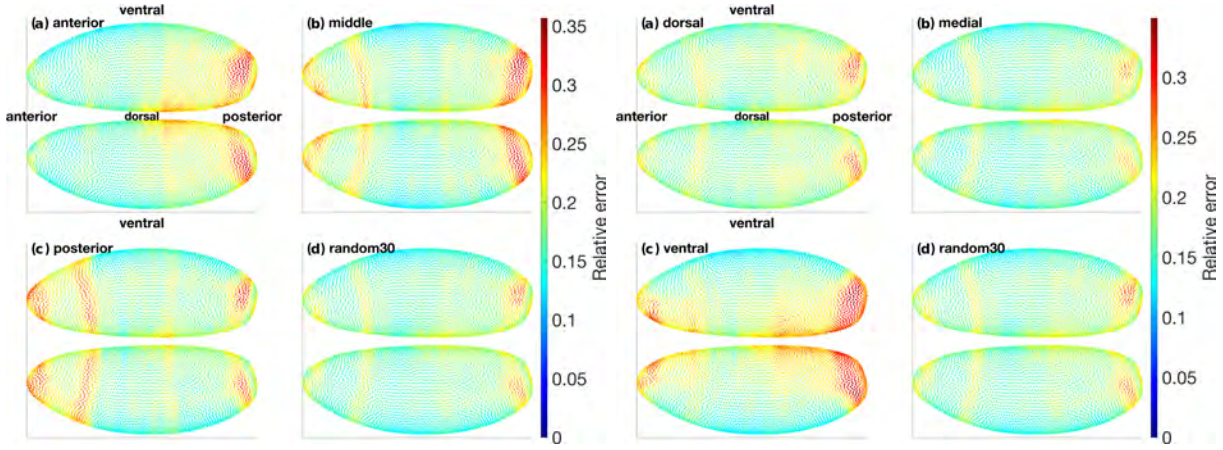

Figure S14: **Spatial distribution of Hill-model errors for different regional partitions.** Relative errors for the 27 fully measured genes are visualized on the embryo surface at latent dimension 10 with $\lambda = 10^{-4}$. Panels (a–c) show models trained on individual regions (anterior, middle, posterior for AMP; dorsal, medial, ventral for DMV). Panel (d) shows the random-30% baseline. In both partitions, random-30% yields a more spatially uniform and consistently lower error distribution, especially in posterior and ventral domains where regional training tends to overfit.

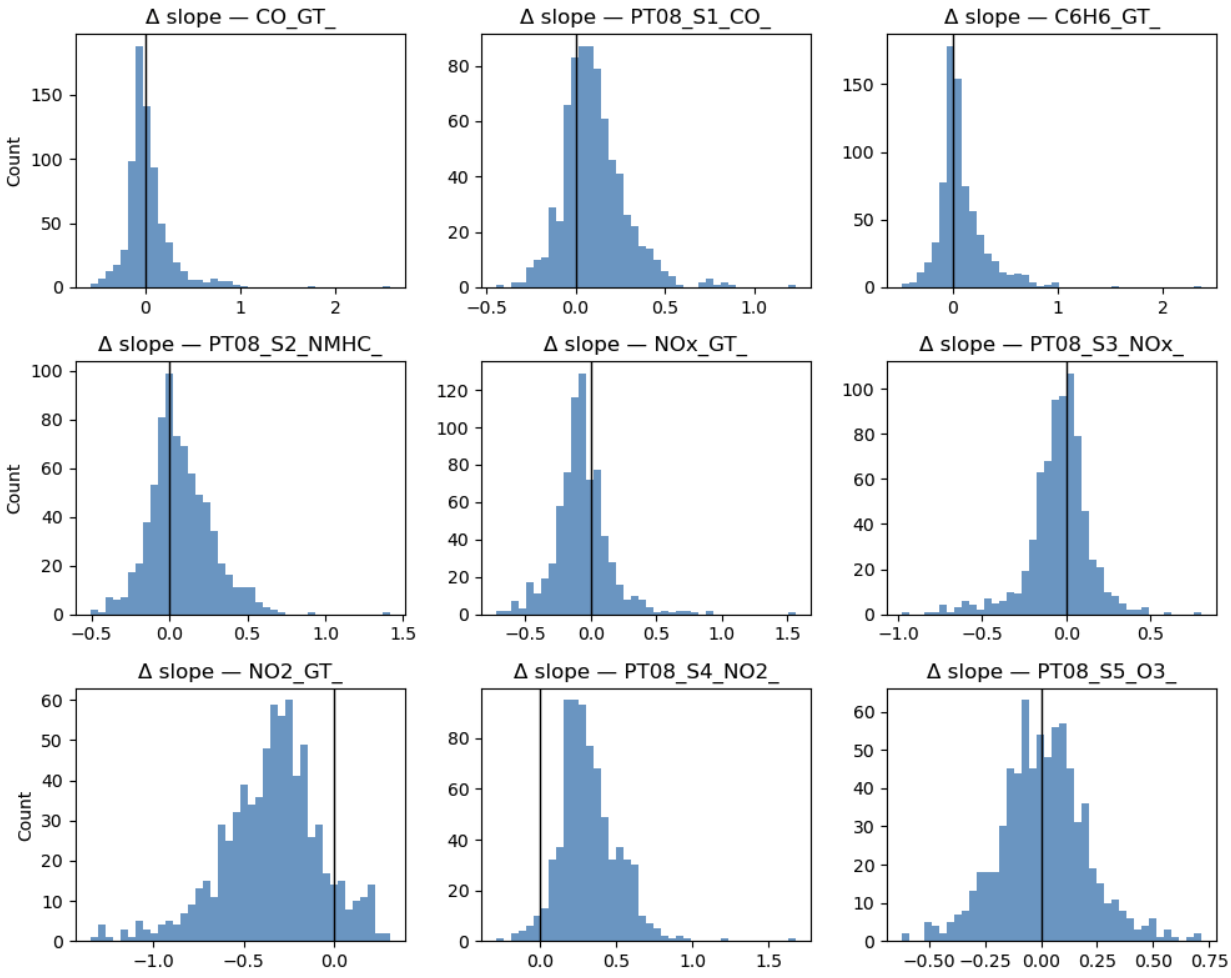

Figure S15: **Temperature-swap sensitivity of pushforward velocities (January → July).** Histograms show the change in instantaneous pushforward slope $\Delta\dot{x}$ per channel when replacing the temperature control in January with the July temperature at the same day and hour while keeping all other controls fixed. Vertical lines indicate $\Delta\dot{x} = 0$.

List of Tables

$$\mathcal{L}_{\text{recon}} = \frac{1}{B}\sum_{b=1}^{B} \frac{\left\| m_t^{(b)} \odot (\hat{g}_t^{(b)} - g_t^{(b)}) \right\|_2^2}{\sum_i m_{t,i}^{(b)} + \varepsilon}$$

$$\mathcal{L}_{\text{KL}} = \tfrac{1}{2}\sum_{j=1}^{d_z} \big(\mu_j^2 + \sigma_j^2 - \log \sigma_j^2 - 1\big), \qquad w_{\text{KL}} = \beta \cdot \min\Big(1, \frac{\text{epoch}+1}{\text{warmup}}\Big)$$

$$\mathcal{L}_{\text{Jac}} = \mathbb{E}_v \left\| J_{\text{Dec}}(z)\, v \right\|_2^2 \;\approx\; \frac{1}{P}\sum_{p=1}^{P} \left\| \frac{\text{Dec}(z + \varepsilon v_p) - \text{Dec}(z - \varepsilon v_p)}{2\varepsilon} \right\|_2^2$$

$$\mathcal{L}_{\text{dir}} = 1 - \cos\big(J_{\text{Dec}}(z)\, r,\ \Delta g\big), \quad r = \frac{\mu_{t+1} - \mu_t}{\|\mu_{t+1} - \mu_t\|},\ \Delta g = g_{t+1} - g_t$$

$$\mathcal{L}_{\text{sp}} = \sum_{\ell} \sum_{i,j} W_{ij}^{(\ell)}\, \tanh\big(W_{ij}^{(\ell)}/\tau\big)$$

Table S1: Loss components used in the VAE: (1) masked reconstruction $\mathcal{L}_{\text{recon}}$; (2) KL with warm-up $w_{\text{KL}}\mathcal{L}_{\text{KL}}$; (3) decoder Jacobian-size proxy $\mathcal{L}_{\text{Jac}}$; (4) optional directional alignment $\mathcal{L}_{\text{dir}}$; (5) smooth sparsity $\mathcal{L}_{\text{sp}}$. The total objective is $\mathcal{L} = \mathcal{L}_{\text{recon}} + w_{\text{KL}}\mathcal{L}_{\text{KL}} + \lambda_J\mathcal{L}_{\text{Jac}} + \lambda_{\text{dir}}\mathcal{L}_{\text{dir}} + \lambda_{\text{sp}}\mathcal{L}_{\text{sp}}$. Here $m_t^{(b)}$ denotes the binary observation mask for sample b, $\odot$ denotes elementwise multiplication, and $\varepsilon > 0$ is a small constant for numerical stability.

Category	Setting	Value / Notes
Inputs/targets	Encoder input	$[g_t \| s] \in \mathbb{R}^{G+3}$; nonnegativity of g_t preserved
	Decoder target	Genes only, $\hat{g}_t \in \mathbb{R}^G$; spatial coordinates are not reconstructed
	Masking	Per-row binary mask on g_t (99D: loaded; 27D: all ones). *The same mask is applied to reconstruction, Jacobian, and directional terms so unobserved/pre-imputed entries never affect gradients.*
Model	Depth	Compact MLP encoder/decoder (4 fully connected layers each)
	Activations	Smooth nonlinearities (e.g., `tanh` or `gelu`)
	Decoder dropout	$\approx$ 0.10 between hidden layers (encoder dropout usually 0)
	Output scaling	Sigmoid output scaled by 3.18 (matches training range)
	Numeric precision	`float64` throughout
Optimization	Optimizer	AdamW
	Learning-rate schedule	Short warmup to the peak learning rate, then cosine decay to **5%** of the peak
	Batch / epochs	Batch size = 500; up to 3000 epochs with early stopping
	Checkpointing	Save the model with the best validation loss
Defaults	Hidden widths	$(N_1, N_2, N_3) = (1024, 512, 256)$
	Latent prior noise	Standard-normal with scale 0.10
	KL regularization	Weight $\beta = 10^{-5}$; warmup $\approx$ 300 epochs
	Smooth sparsity	Weight decay surrogate: threshold 10^{-5}; weight in $[0, 10^{-6}]$ (gentle)
	Decoder Jacobian control	Penalty weight 10^{-4}; finite-difference step 0.10; Hutchinson probes = 1
	Directional alignment (optional)	Weight $\approx$ 0 by default (off); activates only when $\|\Delta g\| > 0.02$
	Peak LR and decoder LR ratio	Peak LR = 10^{-3}; decoder LR = 0.5$\times$ encoder LR
	Warmup length	1500 epochs to peak before cosine decay

Table S2: Exact VAE settings used in our experiments. Values are defaults unless stated otherwise in the main text.

Category	Setting	Value / Notes
Inputs	Data source	Full spatiotemporal atlas after imputation
	Pairing	Build $(\mu_t, \mu_{t+1}, \sigma_{t+1})$ from consecutive times $t \to t{+}1$
Model	Latent dim d_z	Inherited from VAE (typically 6–14)
	Velocity net	3 hidden layers, width N_s=256
	Activation	`tanh` nonlinearity
	Output	$\dot{z} \in \mathbb{R}^{d_z}$
ODE integration	Solver	Runge–Kutta (`dopri5`) via `torchdiffeq`
	Endpoints	Either only $t = 0, 1$ or N_{step} points in $[0, 1]$
	Tolerances	Relative tol 10^{-12}, absolute tol 10^{-12}
	Adjoint	Gradient via adjoint method by default
Loss	Objective	Relative squared error scaled by σ^2_{t+1}
	Formula	$\mathcal{L}_{\text{ODE}} = \frac{1}{d_z}\sum_j (z_{t+1,j} - \mu_{t+1,j})^2/(\sigma^2_{t+1,j} + \varepsilon)$
	Scale ε	10^{-8} (stability)
Optimization	Optimizer	Adam with maximum LR $= 10^{-3}$ (typical)
	LR schedule	Sigmoid-like warmup to epoch `ep_maxlr`, then slow decay
	Gradient clipping	Norm clip at 1.0
	Batch / epochs	Batch size = 1024, up to 3000 epochs
	Checkpoints	Save model with best validation loss

Table S3: Neural ODE settings. Symbols: μ_t latent mean at time t, σ_{t+1} encoder-predicted posterior standard deviation at time $t + 1$, d_z latent dimension, N_s hidden width, N_{step} number of integration points.

Supplemental References

References

[1] J. Han, S. Perera, Z. Wunderlich, V. Periwal, Mechanistic gene networks inferred from single-cell data with an outlier-insensitive method, Mathematical biosciences 342 (2021) 108722.