TEMPOBRIDGE: SOURCE-CONDITIONED FLOW MATCHING WITH OPTIMAL TRANSPORT COUPLINGS FOR SINGLE-CELL POPULATION TRANSITIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Destructive single-cell measurements provide unpaired population snapshots rather than observations of the same cells across conditions. Local cell states and transition requests may also be insufficient to distinguish responses across source populations. We introduce TempoBridge, a common source-conditioned transport formulation for temporal, genetic, and chemical population transitions. Source cells initialize latent transport and provide a fixed empirical population summary. The velocity field receives this summary alongside the evolving cell state, flow time, and a structured transition descriptor. Minibatch optimal transport (OT) supplies couplings only for conditional flow-matching training paths; inference requires neither target expression nor OT computation. On held-out donors, TempoBridge achieves an Energy distance of 0.129 versus 0.144 for scGen. Genetic mean-expression L_2 error is 2.261 versus 3.156 for scGPT-scratch under Seen 2/2. On held-out compounds, condition-averaged drug-effect correlation is 0.598 versus 0.561 for the CellFlow adapter. Temporal ablations show higher mean distributional error after removing source context, replacing optimal transport with random pairing, or replacing flow matching with static residual regression. Together, these results demonstrate the predictive utility of a common source-conditioned transport formulation across held-out donors, gene combinations, and compounds.

1 INTRODUCTION

Single-cell experiments measure populations at selected time points and under selected perturbations, leaving many differentiation states, gene combinations, and drug responses unobserved (Cuomo et al., 2020; Norman et al., 2019; Srivatsan et al., 2020). We consider the prediction of a target gene-expression distribution from source-cell profiles and a requested time point or perturbation. Because sequencing destroys measured cells, training provides unpaired source and target populations rather than trajectories of the same cells (Schiebinger et al., 2019). At inference, the requested target population is unavailable.

Methods for this setting use learned representations of cells and perturbations, mappings between populations, or conditional transport. scGen and CPA model perturbation effects in latent spaces (Lotfollahi et al., 2019; 2023); CellOT maps unpaired populations (Bunne et al., 2023); and CMonge and CellFlow condition transport on experimental descriptors (Driessen et al., 2026; Klein et al., 2025). scDFM uses control-cell expression, while SCALE incorporates a set-level control representation (Yu et al., 2026; Chen et al., 2026). These methods establish the value of source cells and experimental conditions. Here we ask how a transport field should use information about the particular source population as its cells move. Initializing a particle from a source cell gives the field that cell's state, but does not separately expose the population from which it was sampled.

The distinction becomes consequential when cells in similar states require different responses across source populations. Under a specified coupling and training-path construction, two populations can have different conditional mean velocities at the same latent state, flow time, and transition descriptor. A field conditioned only on these inputs receives the same input in both cases and cannot represent both velocities. We call this pointwise conflict *source-dependent transition ambiguity* (Section 3.2).

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.

Supplying a source-population summary can resolve the conflict when that summary retains the relevant difference between populations.

We introduce TempoBridge, a transport model that keeps an empirical summary of the initial source population available throughout integration. Each source cell initializes a latent particle; mean pooling the initial cell embeddings produces a summary shared by all particles from that population. The velocity field takes the evolving particle state, flow time, source summary, and transition descriptor as separate inputs. We train it with optimal-transport conditional flow matching (OT–CFM), using minibatch Sinkhorn couplings to construct paths between unpaired endpoints (Lipman et al., 2023; Tong et al., 2024). Prediction starts from source-cell latents and requires neither target expression nor OT computation. The temporal, genetic, and drug-response models follow this formulation but are trained independently.

We evaluate the complete models on held-out donors, gene combinations, and compounds. For intermediate populations of unseen donors, TempoBridge achieves an Energy distance of 0.129 versus 0.144 for scGen. On Seen 2/2 gene combinations, mean-expression L_2 error is 2.261 versus 3.156 for scGPT-scratch. On compounds excluded from training and validation, condition-averaged drug-effect correlation is 0.598 versus 0.561 for the CellFlow adapter. These comparisons measure the performance of the complete task models, but do not isolate the contribution of source context.

Targeted analyses separate source-context use from predictive benefit. In a controlled ambiguity task, context raises branch accuracy from 50.0% to 98.0%; donor-context swaps show that a fitted temporal model uses the summary. Matched temporal retraining lowers Energy from 0.139 to 0.135 and MMD from 0.397 to 0.378 with context, although effect metrics favor no context. Random pairing raises temporal Energy from 0.1291 to 0.1446; a static residual variant reaches 0.1350 but also changes the training objective. A separate drug conditioning analysis finds small gains in effect metrics but no consistent distributional benefit.

Our contributions are:

- **A source-dependent ambiguity.** Under fixed couplings and paths, we identify a pointwise conflict for a velocity field given only cell state, flow time, and transition descriptor, and demonstrate it in a controlled construction.
- **A source-conditioned transport field.** A fixed summary of the initial source population conditions each particle's velocity throughout transport; OT couplings construct training paths without being required at inference.
- **An assessment of the design.** Donor-context swaps and matched context ablations distinguish use from metric-dependent benefit; temporal component variants test the training recipe, while independently trained models are evaluated on held-out donors, gene combinations, and compounds.

2 Related Work

Optimal transport and flow matching. Optimal transport provides a distributional framework for learning from unpaired cell populations. CellOT learns neural mappings between cellular distributions (Bunne et al., 2023), while CMonge conditions transport mappings on treatment covariates (Driessen et al., 2026). Flow matching learns continuous transformations through velocity regression (Lipman et al., 2023). OT–CFM uses minibatch transport couplings to construct training paths (Tong et al., 2024), while GENOT uses flow matching to learn entropic Wasserstein and Gromov–Wasserstein transport plans (Klein et al., 2024). For cellular response prediction, CellFlow incorporates experimental conditions into flow-based transport (Klein et al., 2025), and scDFM combines control-conditioned generation with MMD-based distributional alignment (Yu et al., 2026). TempoBridge builds on OT–CFM and represents the evolving cell state, empirical source-population context, and structured transition descriptor as separate inputs to the velocity field. Source cells initialize the latent particles, while a fixed summary of their initial population conditions the field throughout integration.

Temporal population modeling. Temporal single-cell models describe changes in population distributions from snapshots collected at different times. Waddington-OT estimates temporal population

couplings (Schiebinger et al., 2019); PRESCIENT learns potential-based stochastic differentiation dynamics (Yeo et al., 2021). TrajectoryNet uses dynamic optimal transport (Tong et al., 2020), scNODE combines latent representations with neural differential equations (Zhang et al., 2024), and BranchSBM models branching populations through multiple velocity fields and growth processes (Tang et al., 2026). TempoBridge uses endpoint supervision to predict intermediate population distributions at requested times for held-out donors.

Genetic and chemical perturbation prediction. Predicting responses to unseen interventions requires representations of the requested perturbations. GEARS uses gene-relation graphs for multi-gene response prediction (Roohani et al., 2024); chemCPA encodes molecular information for novel-compound prediction (Hetzel et al., 2022); and MAP combines knowledge-informed representations with a pretrained single-cell model (Feng et al., 2026). Genetic-perturbation benchmarks also motivate strong linear and additive baselines (Ahlmann-Eltze et al., 2025). TempoBridge represents genetic interventions as unordered gene sets and chemical interventions through molecular and experimental descriptors.

3 Problem Setup

3.1 Unpaired Population Prediction

For task $q \in \{\text{time}, \text{gene}, \text{drug}\}$, a training record contains source cells $\mathcal{X}_s = \{x_i^s\}_{i=1}^{n_s}$, target cells $\mathcal{X}_t = \{x_j^t\}_{j=1}^{n_t}$, and a transition descriptor c_q. Source and target populations are associated within each training record but have no cell-level correspondences. At inference, only $\mathcal{X}_s$ and c_q are observed; the number of output cells is specified separately. Each cell is represented by an expression vector in $\mathbb{R}^{G_q}$, where G_q is the number of genes used for task q. Their empirical distributions are

$$\widehat{P}_s = \frac{1}{n_s}\sum_{i=1}^{n_s}\delta_{x_i^s}, \qquad \widehat{P}_t = \frac{1}{n_t}\sum_{j=1}^{n_t}\delta_{x_j^t}, \tag{1}$$

where δ_x denotes a point mass at x. Given the observed source distribution $\widehat{P}_s$ and a structured transition descriptor c_q, we aim to predict the corresponding target population distribution

$$P_{\theta_q}^{\text{tgt}} = \mathcal{T}_{\theta_q}(\widehat{P}_s, c_q) \approx \widehat{P}_t. \tag{2}$$

The descriptor specifies the requested target; in temporal prediction, the field is conditioned on the shared Day 0–Day 3 endpoint transition, while the requested day determines the integration horizon rather than entering the field directly (Section 4.2).

3.2 Source-Dependent Transition Ambiguity

Suppressing the task index, fix a source–target coupling and path construction for each training record, and let Z_τ and U_τ be the resulting latent position and path velocity. For source population $\widehat{P}_s$ and descriptor c, define the conditional mean velocity $m_s(z,\tau,c) = \mathbb{E}[U_\tau \mid Z_\tau = z, \tau, \widehat{P}_s, c]$. If two source populations with the same c have overlapping path states at $(z^\star, \tau^\star)$ but different values of $m_s(z^\star, \tau^\star, c)$, a deterministic field $v_\theta(z,\tau,c)$ receives identical inputs there and cannot match both means. We call this pointwise conflict *source-dependent transition ambiguity*. For the fixed coupling and path law, Appendix B.5 quantifies the excess squared-error risk when source-specific mean velocities differ on an overlap of positive probability.

Conditioning on an empirical summary of the source population gives

$$v_\theta\Big(z, \tau, h(\widehat{P}_s), c\Big). \tag{3}$$

This field can distinguish the two population-specific mean velocities if h retains the relevant source-population differences. We test this information restriction in controlled ambiguity experiments (Section 5.5.1).

4 TEMPOBRIDGE

For each task q, a shared encoder maps unpaired source and target cells into a latent space. Source latents initialize transport, while their mean provides fixed source-population context for the velocity field throughout integration. Minibatch OT supplies latent endpoint pairs for flow matching. At inference, the learned field transports source-seeded particles under the requested transition, and the decoder predicts expression without target cells or OT. Temporal, genetic, and drug-response models are trained separately.

4.1 CELL-STATE ENCODING AND SOURCE CONTEXT

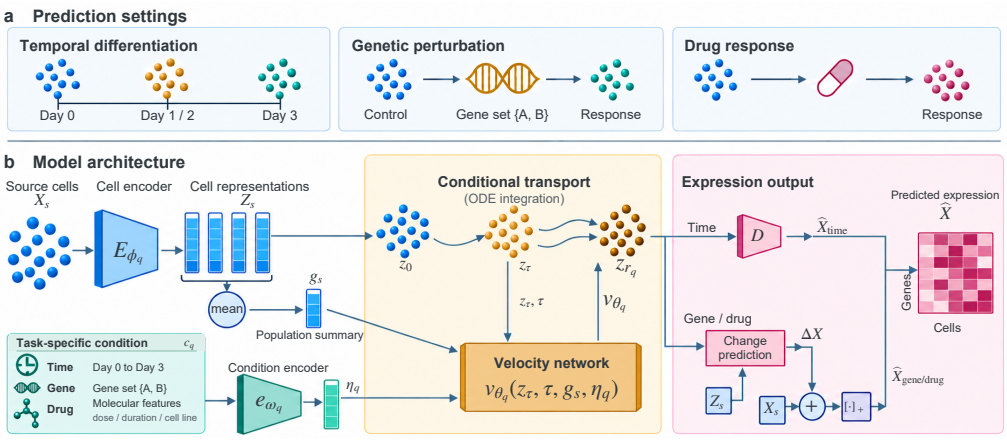

Figure 1: **TempoBridge overview. (a)** Temporal, genetic, and drug-response models use a common source-conditioned transport formulation and are trained independently. **(b)** Source cells initialize latent particles and provide a fixed empirical population summary. The velocity field takes the evolving latent state, flow time, source summary, and encoded transition descriptor as inputs. Each task model has its own expression decoder. Minibatch OT constructs training paths; inference requires no target expression or OT computation.

Figure 1 summarizes the common transport design. Within task q, an encoder $E_{\phi_q} : \mathbb{R}^{G_q} \to \mathbb{R}^{d_q}$ maps source and target cells into the same latent space:

$$z_i^s = E_{\phi_q}(x_i^s), \qquad z_j^t = E_{\phi_q}(x_j^t). \tag{4}$$

Target cells are encoded during training; source latents seed transport (Section 4.4). The source population is summarized by permutation-invariant mean pooling (Zaheer et al., 2017):

$$g_s = h(\widehat{P}_s) = \frac{1}{n_s} \sum_{i=1}^{n_s} z_i^s. \tag{5}$$

Computed from the valid source cells, g_s remains fixed during integration and is shared by all particles from that population.

4.2 STRUCTURED CONDITION INTERFACE

The transition condition is encoded separately from the source population:

$$\eta_q = e_{\omega_q}(c_q^{\text{field}}), \qquad v_{\theta_q}(z_\tau, \tau, g_s, \eta_q), \tag{6}$$

where c_q^{field} is the descriptor supplied to the velocity field. The field therefore receives the current cell state, flow time, source-population summary, and transition descriptor as separate inputs.

For temporal differentiation, $c_{\text{time}}^{\text{field}}$ specifies the Day 0–Day 3 endpoint transition. A request for day t sets the integration horizon to $r_{\text{time}}(t) = \sqrt{t/3}$; this mapping is fixed rather than learned. For genetic and drug prediction, $c_q^{\text{field}} = c_q$ and the integration horizon is one. Genetic conditions

represent one or two perturbation genes by order-invariant pooling of entries from a task-specific gene-identity embedding table. Drug conditions use SMILES-derived molecular features (Weininger, 1988), dose, treatment duration, and cell line, without a learned drug-identity lookup. Appendix B gives the task-specific encodings.

4.3 OT-Guided Conditional Flow Matching

Source and target cells have no observed correspondence. Following flow matching and minibatch OT–CFM (Lipman et al., 2023; Tong et al., 2024), we compute a coupling within each source–target population record. The cost is the squared Euclidean distance $C_{ij} = \|z_i^s - z_j^t\|_2^2$, normalized to obtain $\widetilde{C}$ as described in Appendix C.3. With uniform marginals $\alpha_i = 1/n_s$ and $\beta_j = 1/n_t$, the entropically regularized plan is

$$\pi_q^\star = \underset{\pi \in \Pi(\alpha,\beta)}{\arg\min} \left\{ \langle \pi, \widetilde{C} \rangle + \varepsilon_q \sum_{i,j} \pi_{ij}(\log \pi_{ij} - 1) \right\}. \tag{7}$$

Sinkhorn iterations approximate the plan (Cuturi, 2013) without gradient tracking. Endpoint pairs are sampled with replacement from its full joint distribution.

For a sampled pair (z_0, z_1) and $\tau \sim \mathcal{U}(0,1)$, the training path and target velocity are

$$z_\tau = (1-\tau)z_0 + \tau z_1, \qquad u_\tau = z_1 - z_0. \tag{8}$$

The flow-matching objective is

$$\mathcal{L}_{\mathrm{FM}}^{(q)} = \mathbb{E}_{(z_0,z_1)\sim\pi_q^\star,\tau} \left[\frac{1}{d_q} \left\| v_{\theta_q}(z_\tau, \tau, g_s, \eta_q) - (z_1 - z_0) \right\|_2^2 \right]. \tag{9}$$

The expectation is estimated from sampled pairs across the training minibatch. OT pairs define training paths; they are not observed cell correspondences.

The objective for the transport-training stage combines flow matching, reconstruction of the two observed endpoints, and task-specific auxiliary terms:

$$\mathcal{L}_{\mathrm{total}}^{(q)} = \lambda_{\mathrm{FM}}^{(q)}\mathcal{L}_{\mathrm{FM}}^{(q)} + \lambda_{\mathrm{src}}^{(q)}\mathcal{L}_{\mathrm{rec,src}}^{(q)} + \lambda_{\mathrm{tgt}}^{(q)}\mathcal{L}_{\mathrm{rec,tgt}}^{(q)} + \sum_{k\in\mathcal{A}_q} \lambda_k^{(q)}\mathcal{L}_k^{(q)}. \tag{10}$$

The reconstruction terms compare each observed endpoint with its own decoded latent representation, using a Huber loss (Huber, 1964) that balances zero and nonzero expression entries. The set $\mathcal{A}_q$ contains the auxiliary terms enabled for task q. Any representation or condition pretraining has a separate objective.

For differentiation, the auxiliary term orders predicted population means along the direction between the observed endpoints. It uses full-interval rollouts under virtual time descriptors; measured Day 1 and Day 2 expression does not enter this loss. Intermediate prediction instead retains the Day 0–Day 3 descriptor and changes the integration horizon. For genetic and drug prediction, auxiliary terms constrain population mean responses and latent or expression-space moments. In the genetic objective, the population-mean and mean-shift correlation losses supervise quantities also assessed by Mean L2 and Δr. Appendix C.4 specifies these terms and their coefficients.

4.4 Population Generation

Given source cells and a transition descriptor, the model computes Z_s, g_s, and η_q, then integrates the neural ordinary differential equation (Chen et al., 2018):

$$\frac{\mathrm{d}z_i(\tau)}{\mathrm{d}\tau} = v_{\theta_q}(z_i(\tau), \tau, g_s, \eta_q), \qquad z_i(0) = \widetilde{z}_i^s, \quad 0 \le \tau \le r_q(c_q). \tag{11}$$

Both g_s and η_q remain fixed during integration. The horizon is one for genetic and drug perturbations and follows the fixed mapping above for temporal requests. For temporal and genetic prediction, $\widetilde{z}_i^s = z_i^s$. For drug prediction, $\widetilde{z}_i^s = z_i^s + \epsilon_i - \bar{\epsilon}$, where the ϵ_i are independently sampled from $\mathcal{N}(0, I)$ and $\bar{\epsilon}$ is their mean within the generated population. The number of output cells follows the

task's inference protocol. For each output position i, (x_i^s, z_i^s) denotes the selected source cell and its encoded latent state used for initialization. Target expression and OT are not used at inference.

Let $\widehat{z}_i = z_i(r_q(c_q))$ denote the transported latent state at the requested horizon. Temporal prediction decodes the transported latent states as absolute expression. Genetic and drug prediction use source-anchored residual decoding:

$$\widehat{x}_i = \max\{0,\, x_i^s + D_q(\widehat{z}_i) - D_q(z_i^s)\}\,. \tag{12}$$

The evaluation protocols specify source-cell selection, numerical solvers, and checkpoint selection for each task.

5 Experiments

5.1 Experimental Protocol

We evaluate population prediction for held-out donors, unseen perturbation components or combinations, and unseen compounds. Controlled and matched comparisons assess source context and transport components. Within each task, methods share the evaluation split, gene space, sampling protocol, and evaluator. The genetic feature panel was selected before condition splitting (Section 5.3). Complete comparisons appear in Appendix A. Main comparisons use one validation-selected checkpoint per learned method and training setting, evaluated without retraining over three randomized test runs (seeds 42, 123, and 2026). Mean $\pm$ sample standard deviation (SD) summarizes run-level aggregates; genetic tables report means only. Retraining SD and donor- or compound-bootstrap intervals are reported separately. Drug main-table means weight conditions equally, whereas paired intervals weight compounds equally. Appendix D.6 defines aggregation, statistical units, and uncertainty estimators. Mean L2 and effect correlation Δr measure population mean-response agreement; Energy distance (Székely & Rizzo, 2013) and maximum mean discrepancy (MMD) (Gretton et al., 2012) compare distributions. Differential-expression (DE) recovery uses top-response-gene overlap (DE@100) or fold-change rank correlation on reference DE genes (DE ρ). Drug pEnergy and pMMD use a common projected expression space. Task-specific estimators are defined in Appendix D.6.

5.2 Differentiation in Held-Out Donors

The Cuomo iPSC time course (Cuomo et al., 2020) tests whether the learned transport predicts intermediate populations for unseen donors. All cells from a donor remain in one partition. Representation pretraining and transition training use only training-donor Day 0 and Day 3 populations; measured Day 1/2 expression is excluded from both stages. Validation Day 0→Day 3 predictions select checkpoints. Requiring at least 64 cells at Days 0, 1, and 2 yields 15 evaluation donors and 30 day 0→Day 1/2 requests. For metric comparison, the evaluator sets the output count to the observed target count; target expression is not provided to the model. TempoBridge achieved the lowest mean MMD and Energy and the highest mean Δr and DE@100 among the primary evaluated methods (Table 1). Energy was 0.1291 versus 0.1442 for scGen, and DE@100 was 0.5624 versus 0.5220 for PRESCIENT. Covariance recovery followed a different ranking: gene-covariance correlation was 0.4176 versus PRESCIENT's 0.4874. The auxiliary time-adapted scVIDR (Kana et al., 2023) reference also had higher DE@100 (0.694 ± 0.003; Appendix D.7). These comparisons assess full-model prediction of intermediate populations for held-out donors; Sections 5.5 and 5.6 examine the role of source context separately.

5.3 Held-Out Genetic Perturbations

The Norman K562 Perturb-seq benchmark (Norman et al., 2019) evaluates held-out intervention conditions using a fixed source panel of 128 control cells. It therefore tests perturbation prediction rather than the effect of varying source context. Seen 2/2 holds out combinations whose two components appear in training interventions; Seen 1/2 and Seen 0/2 expose one or neither component. Here "unseen" concerns intervention labels, not absence of the gene from expression inputs. Additive combines observed training single-gene responses and applies only to Seen 2/2. Seen 0/1 tests single genes held out in the Seen 0/2 split. The four test sets contain 20, 20, 16, and 11 conditions. Models are trained separately for the combination splits, with validation Mean L2 selecting checkpoints;

Table 1: Temporal differentiation on 15 held-out donors, covering 30 prediction requests in the 2,009-gene evaluation space. Scores are aggregated across donors and requested times within each run. Entries report mean ± standard deviation over three randomized test runs with fixed trained models.

Method	MMD ↓	Energy ↓	Δr ↑	DE@100 ↑
scGen	0.4300 ± 0.0072	0.1442 ± 0.0019	0.6168 ± 0.0049	0.5111 ± 0.0043
Conditional VAE (2015)	0.4243 ± 0.0023	0.1462 ± 0.0006	0.5826 ± 0.0011	0.4532 ± 0.0053
Conditional MLP	0.4302 ± 0.0032	0.1507 ± 0.0021	0.5742 ± 0.0040	0.4448 ± 0.0070
CellOT	0.4459 ± 0.0105	0.1521 ± 0.0022	0.5849 ± 0.0053	0.4579 ± 0.0082
PRESCIENT	0.4967 ± 0.0052	0.1635 ± 0.0035	0.6210 ± 0.0064	0.5220 ± 0.0035
scNODE	0.5628 ± 0.0018	0.2294 ± 0.0059	0.3543 ± 0.0449	0.3316 ± 0.0327
TrajectoryNet	0.5453 ± 0.0023	0.2785 ± 0.0042	-0.2845 ± 0.0189	0.3484 ± 0.0059
TempoBridge	$\mathbf{0.3626 \pm 0.0143}$	$\mathbf{0.1291 \pm 0.0047}$	$\mathbf{0.6284 \pm 0.0115}$	$\mathbf{0.5624 \pm 0.0110}$

Seen 0/1 reuses Seen 0/2 checkpoints without retraining or further selection. The gene panel includes highly variable genes selected from the full quality-controlled dataset before condition splitting (Appendix C.6). TempoBridge achieved Mean L2 2.2610 versus 3.1564 for scGPT-scratch (Cui et al., 2024), the lowest-L2 baseline; Additive obtained 3.6163 and $\Delta r = 0.8719$, versus 0.9161 for TempoBridge (Table 2). Energy was 0.4793 versus 0.7335 for scDFM, the lowest-Energy baseline in this setting (Table S2). Across all three combination holdouts, TempoBridge had the lowest Mean L2 and Energy and highest Δr among the evaluated methods. In Seen 0/1 it retained the lowest Mean L2 and highest Δr, while GEARS had lower Energy. Table S3 provides an auxiliary, equally weighted average of the four settings. Mean-response and Energy improvements were more consistent across

Table 2: Genetic prediction by holdout setting. Means use the test-repeat protocol in Section 5.1; each setting weights its conditions equally. Additive applies only to Seen 2/2; NA denotes an inapplicable comparison. Bold marks the best mean within each setting and metric. Complete metrics are in Table S2.

	Seen 2/2 20 conditions		Seen 1/2 20 conditions		Seen 0/2 16 conditions		Seen 0/1 11 conditions	
Method	L2↓	Δr ↑	L2↓	Δr ↑	L2↓	Δr ↑	L2↓	Δr ↑
Control	5.6384	0.0011	5.5351	0.0114	4.7860	0.0110	3.5682	0.0113
Additive	3.6163	0.8719	NA		NA		NA	
scGPT-scratch	3.1564	0.8635	4.1885	0.6614	4.1235	0.5237	3.4632	0.4203
GEARS	3.8881	0.7330	4.7681	0.4596	4.2786	0.5081	3.4356	0.3338
CPA	5.0754	0.4594	5.3329	0.3943	6.1414	0.2933	4.1867	0.3266
CellFlow	3.6758	0.8263	4.4758	0.6697	4.1562	0.5436	3.2767	0.4236
scDFM	3.7269	0.8964	4.7935	0.6095	5.5934	0.5014	3.9321	0.2359
TempoBridge	**2.2610**	**0.9161**	**3.9962**	**0.6741**	**3.9880**	**0.5589**	**3.2307**	**0.4330**

combination holdouts than DE ranking, whose strongest method varied by setting (Table S2). In the equally weighted four-setting summary, DE ρ was 0.6628 for TempoBridge versus 0.7474 for scGPT-scratch; TempoBridge led Seen 2/2 but ranked below the other trained methods in Seen 0/1. These results evaluate the complete training recipe, including auxiliary supervision, without isolating OT or flow contributions. Prediction with the ID-based descriptor also does not establish functional identification of unseen genes (Appendices B and C.4).

5.4 Response Prediction for Held-Out Compounds

The compound-disjoint sci-Plex3 split (Srivatsan et al., 2020) contains 150 training, 19 validation, and 19 test compounds across A549, K562, and MCF7 cells. Predictions start from vehicle controls matched by cell line and replicate. Replicates are pooled within cell-line–compound–dose–time conditions, and metrics weight the 248 test conditions equally. This task tests molecular descriptor conditioning for compounds absent from training and validation, without a learned training-compound identity. chemCPA, scDFM, and CellFlow (Hetzel et al., 2022; Yu et al., 2026; Klein et al., 2025)

are evaluated through project-specific adapters with different conditioning and output interfaces (Appendix C.7). Their expression predictions enter the same evaluator; treatment duration is not an explicit input to the chemCPA adapter. CellFlow's internal, train-fitted PCA is distinct from the common 50-dimensional Gaussian projection used for pEnergy and pMMD (seed 42); MMD bandwidth is calibrated on training populations. These comparisons concern the implemented adapters under this protocol. With conditions weighted equally, TempoBridge had the lowest Mean L2 and pEnergy and the highest Δr among the evaluated methods (Table 3). Mean L2 was 4.2478 versus 4.2875 for the CellFlow adapter, and Δr was 0.5977 versus 0.5608 for the same comparator. pEnergy was 0.2123 versus 0.2219 for CFM without OT and 0.2260 for Conditional MLP. In a separate compound-weighted paired analysis, the 95% confidence intervals for the pEnergy differences against CFM without OT and Conditional MLP both included zero (Appendix C.8). TempoBridge ranked second on pMMD (0.0327 versus 0.0317 for Conditional MLP) and DE@100 (0.5471 versus 0.5579 for the scDFM adapter).

Table 3: Unseen-compound response prediction on sci-Plex3 (248 equally weighted test conditions). Values are mean ± sample SD over fixed-checkpoint test repeats (Section 5.1). pEnergy and pMMD use a common 50-dimensional Gaussian random projection (seed 42), with MMD bandwidth calibrated on training data. Baseline adapters are project implementations; bold marks the best mean in each column.

Method	Mean L2↓	Δr ↑	pEnergy↓	pMMD↓	DE@100↑
Conditional MLP	4.8777±0.0455	0.2254±0.0190	0.2260±0.0095	**0.0317±0.0005**	0.3215±0.0228
CFM without OT	4.6357±0.0285	0.3848±0.0056	0.2219±0.0094	0.0345±0.0027	0.2626±0.0054
chemCPA adapter	5.5117±0.1670	0.3827±0.0122	0.3484±0.0358	0.0437±0.0029	0.5139±0.0098
scDFM adapter	7.7612±0.5215	0.2293±0.0220	0.8281±0.1396	0.0747±0.0079	**0.5579±0.0034**
CellFlow adapter	4.2875±0.0125	0.5608±0.0002	9.2228±0.0121	0.2714±0.0001	0.3465±0.0002
TEMPOBRIDGE	**4.2478±0.0470**	**0.5977±0.0047**	**0.2123±0.0193**	0.0327±0.0036	0.5471±0.0129

5.5 SOURCE-CONTEXT INFORMATION AND MODEL USE

The three task comparisons do not isolate source context. We next test its information content, its use by a fitted model, and its predictive benefit under matched retraining.

5.5.1 CONTROLLED SOURCE AMBIGUITY

We tested the information restriction in Section 3.2 using two synthetic source populations with exactly the same central cells and transition descriptor. Additional cells on the left or right distinguished the populations; their unpaired target populations required the central cells to reach opposite vertical branches. A parameter-matched no-context model replaced the pooled context with zeros. With identical shared initial states and descriptors, its deterministic field and solver produce identical shared-cell outputs for both populations. The additional source cells therefore supply the information needed to distinguish the two requests. Across three test runs of one fixed checkpoint per trained arm, branch accuracy was 50.00% without context, 98.04% with correct context, and 1.96% with swapped context (Table 4). Correct context also gave the lowest full-population Energy in this comparison. Appendix E gives the construction, evaluation protocol, and controls.

Table 4: Controlled source-dependent transition ambiguity. Results are mean ± sample SD across three test runs of one fixed checkpoint per trained arm. Branch accuracy evaluates shared synthetic cells; Energy evaluates the full population. Correct and swapped contexts use the same TempoBridge checkpoint.

Model or setting	Branch acc. (%) ↑	Energy ↓
No context	50.00 ± 0.00	0.4575 ± 0.0012
Swapped context	1.96 ± 0.40	3.3299 ± 0.1760
TempoBridge	98.04 ± 0.40	0.0054 ± 0.0004

5.5.2 DONOR-CONTEXT SWAPS

Using the held-out donors from Section 5.2, we split each donor's Day 0 cells into disjoint query and context subsets to test source-context use. With query cells, requested day, model parameters, and solver fixed, we replaced only the pooled context with each of the other 14 donors' contexts. Averaging incorrect contexts increased Energy by 0.00424 (donor-bootstrap 95% CI $[0.00180, 0.00678]$), with 12 of 15 donors favoring the correct context (Table S14). MMD increased by 0.00657 and Δr decreased by 0.01124, with intervals also excluding zero. This intervention indicates use of empirical source information by the fitted model (Appendix D.5); Section 5.6 separately tests predictive benefit under matched retraining.

5.6 TRANSPORT AND CONDITIONING ABLATIONS

Temporal transport variants. Temporal variants replace OT coupling with independent random pairing, or retain OT endpoint sampling while learning a single latent residual update. Random pairing increased mean Energy from 0.1291 to 0.1446; static OT residual increased it to 0.1350 (Table S4). Both variants also reduced Δr and DE@100 and increased MMD. The static formulation replaces pathwise velocity regression with residual regression and omits the flow-specific ordering term (Appendix C.5). These results favor the full temporal training formulation; the static comparison does not isolate numerical integration alone.

Source context with matched retraining. We retrained temporal models with and without source context to test predictive benefit separately from the fixed-model swaps in Section 5.5.2. Across three training seeds, the context model had lower mean MMD (0.378 vs. 0.397) and Energy (0.135 vs. 0.139), while the no-context model had higher Δr (0.639 vs. 0.625) and DE@100 (0.529 vs. 0.515; Table S13). Context therefore favored distributional distances, whereas no context favored effect correlation and DE@100.

Drug-conditioning variants. A matched 2×2 Stage-B analysis varies source-context and cell-line conditioning while sharing a frozen expression backbone and condition warm-start (Table S10). For Full minus No source context, the compound-weighted paired effects were 0.004 for Δr (95% CI $[0.001, 0.006]$) and 0.019 for DE ρ ($[0.008, 0.033]$). Under condition weights, pEnergy slightly favored No source context and pMMD was identical at the reported precision; removing cell-line conditioning reduced Δr from 0.425 to 0.414. These metric-specific findings concern Stage-B conditioning, separately from the main drug comparison and its test-repeat variability (Appendix C.8).

6 DISCUSSION AND CONCLUSION

The controlled construction identifies an information restriction when local state and the requested transition are identical. Frozen-model donor swaps show that the fitted field uses source context on biological data. Matched retraining separately tests whether supplying that information improves prediction; full-model comparisons alone do not isolate the contribution of source context.

Predictive benefit depends on the setting and metric. In matched temporal retraining, context favored distributional distances, whereas the no-context model favored effect correlation and DE@100; the drug conditioning analysis also showed metric tradeoffs. The fixed-source genetic benchmark evaluates conditional perturbation prediction rather than source-context benefit. Across the biological tasks, mean-response and population-distribution gains were more consistent than covariance recovery and DE ranking.

TempoBridge addresses missing source-population information in unpaired transition prediction by conditioning the transport field separately on evolving cell states, a fixed initial empirical source-population summary, and the requested transition. Temporal, genetic, and drug models are trained independently. OT couplings construct the flow-matching training paths, while inference requires only the source population and transition descriptor.

AI USE STATEMENT

Generative AI tools assisted with manuscript organization, language editing, and code development and debugging for method implementation, experiments, and figure formatting. Numerical figures were prepared from existing experimental records; these tools were not used to generate experimental observations. The authors take responsibility for the final text, claims, code, and other artifacts, including all material produced with AI assistance.

ETHICS STATEMENT

This study retrospectively analyzes previously collected single-cell population measurements. We do not claim individual-level clinical prediction, biological lineage identification, or causal effects of interventions. Dataset access and reuse follow the conditions specified by the original data providers.

REPRODUCIBILITY STATEMENT

The main text and appendix document task-specific splits, target-visibility boundaries, model settings, model-selection rules, and evaluation estimators. An anonymized code archive is provided as supplementary material. The reported comparisons are retrospective evaluations, as described in Section 5.1.

Each principal learned comparison uses one checkpoint selected by validation metrics per method and training setting. The selected checkpoint is evaluated using test random seeds 42, 123, and 2026, with no retraining or changes to its weights between runs. Reported means and, where applicable, sample standard deviations (SDs) are calculated from the three run-level aggregate scores. These SDs describe test-run variation conditional on the selected fitted models. Appendix D.6 defines the aggregation procedures and sample SD estimator.

The genetic tables report means only. Table 2 reports each setting separately, while Table S3 assigns equal weight to the four setting-level means. Genetic predictions use one fixed exponential moving average (EMA) parameter set per trained model, without combining predictions across training seeds.

The drug main table evaluates trained two-stage pipelines under the test-run protocol described above. Drug pMMD and pEnergy use a fixed Gaussian random projection, with the MMD bandwidth calibrated on training data. The fixed-backbone Stage-B training-seed analysis is reported separately in the appendix. The Gaussian and PCA sensitivity analyses in Appendix C.8 also concern the fixed-backbone setting. Project-specific adapters are identified in the corresponding drug comparison.

The controlled source-context experiment uses one fixed checkpoint per trained arm and three test runs. Supplementary analyses based on independently trained models report training-seed variability separately from these fixed-checkpoint test-run SDs. Projection seeds define fixed transforms and are not statistical replicates. Paired bootstrap intervals describe variation across biological units conditional on the fitted models and are reported separately from both test-run and training-seed SDs. Drug main-table means weight conditions equally, whereas the paired intervals in Appendix C.8 concern a comparison that weights compounds equally.

REFERENCES

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

Charlotte Bunne, Stefan G. Stark, Gabriele Gut, Jacobo Sarabia del Castillo, Mitch Levesque, Kjong-Van Lehmann, Lucas Pelkmans, Andreas Krause, and Gunnar Rätsch. Learning single-cell perturbation responses using neural optimal transport. *Nature Methods*, 20(11):1759–1768, 2023. doi: 10.1038/s41592-023-01969-x. URL https://www.nature.com/articles/s41592-023-01969-x.

Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David K. Duvenaud. Neural ordinary differential equations. In *Advances in Neural Information Processing Systems*,

volume 31, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/69386f6bb1dfed68692a24c8686939b9-Abstract.html.

Shuizhou Chen, Lang Yu, Xueqin Lin, Xinjie Mao, Songming Zhang, Xinyu Gu, Hao Wu, Sheng Xu, Kedu Jin, Lei Bai, Quan Qian, Qin Chen, Qiang Gao, Siqi Sun, and Zhangyang Gao. SCALE: Scalable conditional atlas-level endpoint transport for virtual cell perturbation prediction. *arXiv preprint arXiv:2603.17380*, 2026. doi: 10.48550/arXiv.2603.17380. URL https://arxiv.org/abs/2603.17380v3. Version 3, revised 31 August 2026; preprint.

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

Anna S. E. Cuomo, Daniel D. Seaton, Davis J. McCarthy, Iker Martinez, Marc Jan Bonder, Jose Garcia-Bernardo, Shradha Amatya, Pedro Madrigal, Abigail Isaacson, Florian Buettner, Andrew Knights, Kedar Nath Natarajan, HipSci Consortium, Ludovic Vallier, John C. Marioni, Mariya Chhatriwala, and Oliver Stegle. Single-cell RNA-sequencing of differentiating iPS cells reveals dynamic genetic effects on gene expression. *Nature Communications*, 11(1):810, 2020. doi: 10.1038/s41467-020-14457-z. URL https://www.nature.com/articles/s41467-020-14457-z.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, volume 26, 2013. URL https://proceedings.neurips.cc/paper/2013/hash/af21d0c97db2e27e13572cbf59eb343d-Abstract.html.

Alice Driessen, Dhruva Abhijit Rajwade, Benedek Harsanyi, Marianna Rapsomaniki, and Jannis Born. Conditional Monge gap enables generalizable single-cell perturbation modelling. *Nature Machine Intelligence*, 8(6):984–996, 2026. doi: 10.1038/s42256-026-01242-8. URL https://www.nature.com/articles/s42256-026-01242-8.

Jinghao Feng, Ziheng Zhao, Xiaoman Zhang, Mingfei Liu, Jingyi Chen, Xingran Quan, Boyang Fu, Jian Zhang, Yanfeng Wang, Ya Zhang, and Weidi Xie. A knowledge-driven framework for predicting single-cell responses for unprofiled drugs. *Nature Machine Intelligence*, 8(9):1478–1491, 2026. doi: 10.1038/s42256-026-01286-w. URL https://www.nature.com/articles/s42256-026-01286-w.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(25):723–773, 2012. URL https://jmlr.org/papers/v13/gretton12a.html.

Leon Hetzel, Simon Böhm, Niki Kilbertus, Stephan Günnemann, Mohammad Lotfollahi, and Fabian J. Theis. Predicting cellular responses to novel drug perturbations at a single-cell resolution. In *Advances in Neural Information Processing Systems*, volume 35, pp. 26711–26722, 2022. doi: 10.52202/068431-1937. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/aa933b5abc1be30baece1d230ec575a7-Abstract.html.

Peter J. Huber. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35(1):73–101, 1964. doi: 10.1214/aoms/1177703732. URL https://doi.org/10.1214/aoms/1177703732.

Omar Kana, Rance Nault, David Filipovic, Daniel Marri, Tim Zacharewski, and Sudin Bhattacharya. Generative modeling of single-cell gene expression for dose-dependent chemical perturbations. *Patterns*, 4(8):100817, 2023. doi: 10.1016/j.patter.2023.100817. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC10436058/.

Dominik Klein, Théo Uscidda, Fabian Theis, and Marco Cuturi. GENOT: Entropic (Gromov) Wasserstein flow matching with applications to single-cell genomics. In *Advances in Neural Information Processing Systems*, volume 37, pp. 103897–103944, 2024. doi: 10.52202/079017-3301. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/bc46e29f91e676747c584ca181cb0ea1-Abstract-Conference.html.

Dominik Klein, Jonas Simon Fleck, Daniil Bobrovskiy, Lea Zimmermann, Sören Becker, Alessandro Palma, Leander Dony, Alejandro Tejada-Lapuerta, Guillaume Huguet, Hsiu-Chuan Lin, Nadezhda Azbukina, Fátima Sanchís-Calleja, Theo Uscidda, Artur Szalata, Manuel Gander, Aviv Regev, Barbara Treutlein, J. Gray Camp, and Fabian J. Theis. CellFlow enables generative single-cell phenotype modeling with flow matching. *bioRxiv*, 2025. doi: 10.1101/2025.04.11.648220. URL https://www.biorxiv.org/content/10.1101/2025.04.11.648220v1. Preprint.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=PqvMRDCJT9t.

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

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

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

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

Geoffrey Schiebinger, Jian Shu, Marcin Tabaka, Brian Cleary, Vidya Subramanian, Aryeh Solomon, Joshua Gould, Siyan Liu, Stacie Lin, Peter Berube, Lia Lee, Jenny Chen, Justin Brumbaugh, Philippe Rigollet, Konrad Hochedlinger, Rudolf Jaenisch, Aviv Regev, and Eric S. Lander. Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming. *Cell*, 176(4):928–943.e22, 2019. doi: 10.1016/j.cell.2019.01.006. URL https://pubmed.ncbi.nlm.nih.gov/30712874/.

Kihyuk Sohn, Xinchen Yan, and Honglak Lee. Learning structured output representation using deep conditional generative models. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015. URL https://proceedings.neurips.cc/paper_files/paper/2015/file/8d55a249e6baa5c06772297520da2051-Paper.pdf.

Sanjay R. Srivatsan, José L. McFaline-Figueroa, Vijay Ramani, Lauren Saunders, Junyue Cao, Jonathan Packer, Hannah A. Pliner, Dana L. Jackson, Riza M. Daza, Lena Christiansen, Fan Zhang, Frank Steemers, Jay Shendure, and Cole Trapnell. Massively multiplex chemical transcriptomics at single-cell resolution. *Science*, 367(6473):45–51, 2020. doi: 10.1126/science.aax6234. URL https://doi.org/10.1126/science.aax6234.

Gábor J. Székely and Maria L. Rizzo. Energy statistics: A class of statistics based on distances. *Journal of Statistical Planning and Inference*, 143(8):1249–1272, 2013. doi: 10.1016/j.jspi.2013.03.018. URL https://www.sciencedirect.com/science/article/abs/pii/S0378375813000633.

Sophia Tang, Yinuo Zhang, Alexander Tong, and Pranam Chatterjee. Branched schrödinger bridge matching. In *International Conference on Learning Representations*, pp. 140186–140230, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/e332505c4c80ad1d9dc0af26103b672b-Abstract-Conference.html.

Alexander Tong, Jessie Huang, Guy Wolf, David Van Dijk, and Smita Krishnaswamy. TrajectoryNet: A dynamic optimal transport network for modeling cellular dynamics. In *Proceedings of the 37th*

International Conference on Machine Learning, volume 119 of *Proceedings of Machine Learning Research*, pp. 9526–9536. PMLR, 2020. URL https://proceedings.mlr.press/v119/tong20a.html.

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

David Weininger. SMILES, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of Chemical Information and Computer Sciences*, 28(1):31–36, 1988. doi: 10.1021/ci00057a005. URL https://doi.org/10.1021/ci00057a005.

Grace Hui Ting Yeo, Sachit D. Saksena, and David K. Gifford. Generative modeling of single-cell time series with PRESCIENT enables prediction of cell trajectories with interventions. *Nature Communications*, 12(1):3222, 2021. doi: 10.1038/s41467-021-23518-w. URL https://www.nature.com/articles/s41467-021-23518-w.

Chenglei Yu, Chuanrui Wang, Bangyan Liao, and Tailin Wu. scDFM: Distributional flow matching model for robust single-cell perturbation prediction. In *International Conference on Learning Representations*, pp. 72511–72535, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/75f029a1289a47aa99c86588239f0c12-Abstract-Conference.html.

Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabás Póczos, Ruslan Salakhutdinov, and Alexander J. Smola. Deep sets. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/f22e4747da1aa27e363d86d40ff442fe-Abstract.html.

Jiaqi Zhang, Erica Larschan, Jeremy Bigness, and Ritambhara Singh. scNODE: generative model for temporal single cell transcriptomic data prediction. *Bioinformatics*, 40(Supplement 2):ii146–ii154, 2024. doi: 10.1093/bioinformatics/btae393. URL https://doi.org/10.1093/bioinformatics/btae393.

A Complete Task and Prediction Results

The tables below retain the complete comparisons underlying the main-text figures, including all reported metrics and all evaluated methods. Table S1 summarizes the task interfaces. Figure panels show selected metrics; claims about other metrics refer to the corresponding tables.

Table S1: Task-specific instances of the conditional population prediction problem. Each task uses an independently trained model and its own gene space. The tasks share a common formulation, but not model parameters.

Task	**Source population**	**Transition descriptor**	**Target population and evaluation setting**
Temporal differentiation	Day 0 cells from a held-out donor	Requested target time	Day 1 or Day 2 population of the same donor
Genetic perturbation	Fixed K562 control population	Unordered set of one or two perturbed genes	Perturbed population under controlled component exposure
Drug response	Matched vehicle-control population	Molecular description, dose, treatment time, and cell line	Treated population for a compound excluded from training and validation

B Detailed Task Parameterization

The main comparisons follow the protocol in Section 5.1: each learned method and training setting contributes one selected checkpoint, evaluated with test random seeds 42, 123, and 2026 while its parameters remain fixed. Training-seed identifiers describe how checkpoints were obtained; test-repeat identifiers index repeated evaluation of a selected checkpoint. Supplementary analyses that aggregate independently trained checkpoints state this separately. The training-seed records below therefore do not define the three main-table test repeats. Projection seeds specify fixed evaluation or loss transforms and are not statistical replicates. Appendix D.6 gives the aggregation formulas.

B.1 Factorized conditioning and source context

The descriptor c_q contains the structured information needed to identify a requested target condition and is implemented separately for each task. In every TempoBridge instance, the descriptor encoding remains separate from the context computed from the observed source cells. Let E_{ϕ_q} be the task-specific expression encoder and let m_i mark valid cells in a padded population batch. The source context used by the velocity field is

$$h(\widehat{P}_s) = \bar{z}_s = \frac{1}{\sum_i m_i} \sum_i m_i E_{\phi_q}(x_i^s). \tag{13}$$

Pooling uses the valid source cells in the input population record, before OT-specific endpoint sampling; the record can itself be sampled by the data loader. OT subsampling retains this summary rather than recomputing it from the selected endpoints. For donor-context swaps, a disjoint context subset supplies the summary, as specified in Appendix D.5. This mean-pooled latent vector is a global context, not a complete encoding of the source distribution; cell-level variation remains in the individual transported latent states. The pooling operation is permutation invariant. For any permutation π of the valid source cells,

$$h\Big(\{x_{\pi(i)}^s\}_{i=1}^{n_s}\Big) = h(\{x_i^s\}_{i=1}^{n_s})\,. \tag{14}$$

Table S5 distinguishes empirical population context, explicit descriptor fields, and metadata that is not supplied to the predictor. Dataset and task identifiers may appear as implementation tokens, but they are constant within the separately trained task instances and do not create a shared-parameter multitask model.

Replicate labels are used to construct populations and check for leakage, not to specify a requested biological transition.

Table S2: Complete genetic metrics by holdout setting. For learned methods, each entry is the mean across test seeds 42, 123, and 2026 using the fixed checkpoint for that training setting, after condition-level averaging within each test run. Control and Additive are deterministic. Additive uses the observed single-gene responses and applies only to Seen 2/2. NA denotes an undefined metric for Control or an inapplicable comparison for Additive. Bold marks the best mean for each setting and metric. Seen 0/1 reuses the Seen 0/2 checkpoints. Control copies a target-sized prefix of the fixed source panel; its effect correlations exclude undefined values, as detailed in Appendix D.6.

Method	Mean L2↓	Δr ↑	Energy↓	DE ρ ↑
Seen 2/2: 20 test conditions				
Control	5.6384	0.0011	1.8371	0.0378
Additive	3.6163	0.8719	0.9871	0.8956
scGPT-scratch	3.1564	0.8635	7.8094	0.9034
GEARS	3.8881	0.7330	0.9807	0.7925
CPA	5.0754	0.4594	7.2051	0.6198
CellFlow	3.6758	0.8263	2.7345	0.8584
scDFM	3.7269	0.8964	0.7335	0.9160
TempoBridge	**2.2610**	**0.9161**	**0.4793**	**0.9233**
Seen 1/2: 20 test conditions				
Control	5.5351	0.0114	1.7694	0.0900
Additive	NA	NA	NA	NA
scGPT-scratch	4.1885	0.6614	8.4925	**0.7678**
GEARS	4.7681	0.4596	1.3841	0.5816
CPA	5.3329	0.3943	7.7166	0.5833
CellFlow	4.4758	0.6697	3.2712	0.7460
scDFM	4.7935	0.6095	1.2584	0.7474
TempoBridge	**3.9962**	**0.6741**	**1.0692**	0.6978
Seen 0/2: 16 test conditions				
Control	4.7860	0.0110	1.3573	0.0196
Additive	NA	NA	NA	NA
scGPT-scratch	4.1235	0.5237	8.0856	0.6873
GEARS	4.2786	0.5081	1.0833	**0.6927**
CPA	6.1414	0.2933	8.5650	0.5235
CellFlow	4.1562	0.5436	2.9558	0.6456
scDFM	5.5934	0.5014	1.8711	0.6637
TempoBridge	**3.9880**	**0.5589**	**0.9738**	0.6102
Seen 0/1: 11 test conditions				
Control	3.5682	0.0113	0.7880	NA
Additive	NA	NA	NA	NA
scGPT-scratch	3.4632	0.4203	7.3042	0.6311
GEARS	3.4356	0.3338	**0.7118**	**0.6725**
CPA	4.1867	0.3266	6.5700	0.5276
CellFlow	3.2767	0.4236	2.2575	0.5856
scDFM	3.9321	0.2359	0.9585	0.4509
TempoBridge	**3.2307**	**0.4330**	0.7133	0.4198

Table S3: Genetic perturbation prediction averaged over four settings: Seen 2/2 (two seen genes; 20 conditions), Seen 1/2 (one seen and one unseen gene; 20 conditions), Seen 0/2 (two unseen genes; 16 conditions), and Seen 0/1 (one unseen gene; 11 conditions). For each method and metric, the reported value is the arithmetic mean of the four setting-level scores, giving each setting equal weight. Each setting-level score first averages test conditions within each run, then averages test seeds 42, 123, and 2026 using one fixed checkpoint per learned method and training setting. Only means are reported; no cross-setting SD is computed. DE ρ denotes Spearman correlation between predicted and observed log-fold changes on reference DE genes. Control is deterministic; NA indicates an undefined four-setting average. Additive is reported in Table S2 for Seen 2/2; its four-setting average is not defined. Seen 0/1 reuses the Seen 0/2 checkpoints without retraining.

Method	Mean L2 ↓	Δr ↑	Energy ↓	DE ρ ↑
Control	4.8819	0.0087	1.4380	NA
scGPT-scratch	3.7329	0.6172	7.9229	**0.7474**
GEARS	4.0926	0.5086	1.0400	0.6848
CPA	5.1841	0.3684	7.5142	0.5636
CellFlow	3.8961	0.6158	2.8048	0.7089
scDFM	4.5115	0.5608	1.2054	0.6945
TEMPOBRIDGE	**3.3690**	**0.6455**	**0.8089**	0.6628

Table S4: Temporal transport-formulation comparisons under the same donor, gene-space, and evaluation protocol as Table 1. Entries report mean ± sample SD over three randomized test runs of one fixed checkpoint per variant (test seeds 42, 123, and 2026), using the same run-level aggregation as the temporal main comparison.

Variant	MMD ↓	Energy ↓	Δr ↑	DE@100 ↑
Random-pairing flow	0.4145 ± 0.0340	0.1446 ± 0.0063	0.6080 ± 0.0102	0.5316 ± 0.0116
Static OT residual	0.3848 ± 0.0207	0.1350 ± 0.0049	0.6159 ± 0.0105	0.5071 ± 0.0105
Full flow + OT	$\mathbf{0.3626 \pm 0.0143}$	$\mathbf{0.1291 \pm 0.0047}$	$\mathbf{0.6284 \pm 0.0115}$	$\mathbf{0.5624 \pm 0.0110}$

Table S5: Conditioning interface used by the three task instances.

Task	Context from source cells	Structured descriptor fields	Not supplied as predictor input
Temporal	Day 0 population from the requested donor	Endpoint times; requested time through the integration horizon	Donor identity; replicate label
Genetic	Matched K562 control population	Perturbed genes; control/single/combination type	Combination identity hash; replicate label
Drug	Matched vehicle-control population	Molecular feature, dose, treatment time, cell line	Drug identity in the unseen-compound setting; replicate label

B.2 TEMPORAL DESCRIPTOR AND INTEGRATION REQUEST

At the formulation level, the temporal descriptor contains an endpoint transition and a requested time,

$$c_{\text{time}} = (c_{\text{end}}, t_r), \qquad c_{\text{end}} = (t_s, t_e, \Delta t), \qquad \Delta t = t_e - t_s. \tag{15}$$

The temporal instance implements these two parts through different interfaces. For endpoint flow matching and evaluation, the condition encoder receives the fixed descriptor $c_{\text{end}} = (0, 3, 3)$, whereas t_r determines how far the endpoint-conditioned field is integrated. The auxiliary ordering branch uses virtual time descriptors. Appendix C.5 specifies the request-to-horizon mapping, ordering regularizer, target-visibility boundary, and numerical solver.

B.3 GENETIC-PERTURBATION DESCRIPTOR

For the Norman instance, the descriptor is

$$c_{\text{gene}} = (\mathcal{G}, \rho), \qquad \mathcal{G} = \{g_k\}_{k=1}^{K}, \qquad \rho \in \{\text{control}, \text{single}, \text{combination}\}. \tag{16}$$

Here $K = 1$ for a single-gene intervention and $K = 2$ for the combinatorial conditions. Gene identifiers are mapped to the expression gene space, padded positions contribute no mass, and the component embeddings are mean-pooled,

$$e_{\mathcal{G}} = \frac{1}{K} \sum_{g \in \mathcal{G}} e_g. \tag{17}$$

The pooled representation is combined with a learned intervention-type token. The condition encoder uses a separate 512-dimensional gene-embedding table, initialized from a zero-mean normal distribution with standard deviation 0.02; the ObjectiveEMA configuration does not load pretrained gene embeddings. An unseen gene here is absent from training intervention labels, not from the expression input space. Its embedding row receives no intervention-specific supervision. Shared downstream parameters are learned from training conditions, and optimizer weight decay can still change that row. Permutation invariance therefore does not imply an external functional prior for an unseen gene. This construction is invariant to the order of genes in a combination and does not use a combination-identity hash. A dedicated control token represents the unperturbed condition. Target expression for held-out combinations is not an inference input.

B.4 DRUG-RESPONSE DESCRIPTOR

For the compound-disjoint sci-Plex3 instance, the task descriptor is

$$c_{\text{drug}} = (f_m, a, t_d, \ell), \qquad f_m \in \mathbb{R}^{256}, \tag{18}$$

where f_m is a SMILES-derived character n-gram representation, a is the dose, t_d is treatment time, and ℓ is the cell-line category encoded from the training vocabulary. The molecular representation is not a Morgan/ECFP fingerprint. Because validation and evaluation compounds are disjoint from training compounds, the discrete drug-identity embedding is disabled. Compounds absent from training are represented through f_m and the remaining structured fields. Cell-line metadata in c_{drug} is distinct from $h(\widehat{P}_s)$, which is computed from the actual matched vehicle-control cells. Replicate labels are used for grouping rather than conditioning. The two-stage training and source-anchored decoding rules are given in Appendix C.6.

B.5 EXPECTED RISK UNDER SOURCE-DEPENDENT AMBIGUITY

Fix the distribution of training records, couplings, paths, and flow time. Let S denote the source population, $X = (Z_\tau, \tau, C)$, and $U = U_\tau$, with $\mathbb{E}\|U\|_2^2 < \infty$. Define $R^*(V) = \inf_f \mathbb{E}\|U - f(V)\|_2^2$, where the infimum is over all measurable square-integrable predictors. Writing $m(X, S) = \mathbb{E}[U \mid X, S]$ and $\bar{m}(X) = \mathbb{E}[U \mid X]$ gives

$$R^*(X) - R^*(X, S) = \mathbb{E}\|m(X, S) - \bar{m}(X)\|_2^2.$$

To prove the identity, expand $U - \bar{m} = (U - m) + (m - \bar{m})$. The cross term vanishes after conditioning on (X, S), and conditional means minimize squared error. Thus the risk gap is positive exactly when $m(X, S) \neq \bar{m}(X)$ with positive probability.

For a quantitative bound, suppose $S \in \{0,1\}$, let $w(X) = \Pr(S = 1 \mid X)$, and write $m_j(X) = \mathbb{E}[U \mid X, S = j]$. Assume an X-measurable event A has probability $p > 0$ and, almost surely on A, $a \le w(X) \le 1 - a$ and $\|m_1(X) - m_0(X)\|_2 \ge \delta$, where $0 < a \le 1/2$ and $\delta > 0$. Then

$$\begin{aligned} R^*(X) - R^*(X,S) &= \mathbb{E}\big[w(X)(1-w(X))\|m_1(X) - m_0(X)\|_2^2\big] \\ &\ge p\, a(1-a)\delta^2. \end{aligned}$$

The equality is the conditional variance of a two-point vector; restricting its expectation to A proves the bound. Divide by latent dimension for dimension-normalized FM loss.

For a deterministic summary $H = h(S)$, the same argument gives

$$R^*(X,H) - R^*(X,S) = \mathbb{E}\|m(X,S) - \mathbb{E}[U \mid X, H]\|_2^2.$$

Consequently, a summary attains the full-source optimum exactly when these conditional means agree almost surely. These claims concern the fixed path law; shared initial states in the synthetic construction alone do not establish the required positive-probability path overlap.

C TempoBridge Implementation Details

C.1 Expression encoders and decoders

Each task uses its own expression encoder and decoder. The temporal and genetic encoders remain trainable during transition learning. The drug instance first trains an expression-reconstruction backbone on the Task 3 training split and then freezes the source encoder and gene embedding during transition training; its decoder remains trainable. The temporal decoder predicts an absolute expression state, whereas the genetic and drug decoders predict source-anchored expression changes. Final expression predictions are constrained to be nonnegative. Gene-space and latent dimensions and the corresponding inference settings are listed in Table S7. Appendix D.2 gives the full temporal architecture and optimization record.

C.2 Descriptor encoders

All task instances preserve the factorization introduced in Section 4: source context is computed from observed source cells, whereas the descriptor encoder receives structured information about the target condition. The temporal instance encodes the Day 0–Day 3 endpoint and supplies the requested day through the integration horizon. The genetic instance mean-pools the embeddings of the requested genes and combines the result with an intervention-type token. The drug instance encodes a 256-dimensional SMILES character n-gram vector, dose, treatment time, and cell line. Its drug-identity embedding is disabled for compound-disjoint evaluation. Appendix B gives the exact descriptor tuples and the metadata excluded from prediction. For batched computation, populations are padded and a validity mask excludes padded positions from the latent mean used as source context.

C.3 OT coupling

For numerical stability, squared latent distances are normalized within each minibatch:

$$M_{ij} = \frac{\|z_i^s - z_j^t\|_2^2}{\mathrm{median}_{i,j}\|z_i^s - z_j^t\|_2^2}. \tag{19}$$

The entropic OT problem used to construct the regression paths uses uniform marginals, after which endpoint pairs are sampled from the joint plan. The temporal protocol uses $\varepsilon = 0.03$ and 200 Sinkhorn iterations; the genetic protocol uses $\varepsilon = 0.05$ and 30 iterations; and the drug protocol uses $\varepsilon = 0.05$ and 100 iterations. For genetics, the plan is computed from detached latents separately within each record, and 128 pairs are sampled with replacement. Gradients pass through the gathered latent values but not through the plan or sampled indices. Random-pairing controls independently sample the two endpoints, and no-OT controls remove the transport plan while retaining the conditional flow architecture.

C.4 RECONSTRUCTION AND AUXILIARY LOSSES

Endpoint reconstruction balances zero and nonzero expression entries. For valid positions Ω, let Ω_0 and Ω_+ denote entries with zero and nonzero targets. With Huber threshold one, the loss for one endpoint is

$$\mathcal{L}_{\text{rec,end}} = \frac{1}{2|\Omega_0|} \sum_{r \in \Omega_0} \ell_{\text{Huber}}(\hat{x}_r, x_r) + \frac{1}{2|\Omega_+|} \sum_{r \in \Omega_+} \ell_{\text{Huber}}(\hat{x}_r, x_r). \tag{20}$$

Source and target endpoints are reconstructed from their own latent states, so this term does not create row-wise supervision across populations.

The genetic instance uses the complete objective

$$\begin{aligned}\mathcal{L}_{\text{gene}} = \mathcal{L}_{\text{FM}} + \mathcal{L}_{\text{rec,src}} + \mathcal{L}_{\text{rec,tgt}} + 100\mathcal{L}_{\text{mean}} + \mathcal{L}_{\Delta\text{-corr}} \\ + 0.1\mathcal{L}_{\text{flow-moment}} + 0.1\mathcal{L}_{\text{direction}} + 0.05\mathcal{L}_{\text{magnitude}} \\ + 0.1\mathcal{L}_{\text{var}} + 0.05\mathcal{L}_{\text{proj-cov}}.\end{aligned} \tag{21}$$

These coefficients are shared by all nine reported genetic training runs. Here $\mathcal{L}_{\text{mean}}$ is Smooth L1 loss between predicted and observed population means in log-expression space, and $\mathcal{L}_{\Delta\text{-corr}}$ is one minus Pearson correlation between their mean shifts relative to the same source population. The flow-moment term sums Smooth L1 losses on the means and log variances of the generated and observed target latents. If $\hat{d}$ and d are their latent mean displacements from the source mean, the direction and magnitude terms are $1 - \cos(\hat{d}, d)$ and Smooth L1 loss between $\log \|\hat{d}\|_2$ and $\log \|d\|_2$, respectively. The source reference and observed displacement are detached for these two terms.

The variance loss compares log gene variances. The covariance loss compares covariance matrices after a fixed 5041×128 Gaussian projection with seed 42, scaling both matrices by the observed projected standard-deviation products before applying Smooth L1. Variances and normalization denominators use the implementation's 10^{-6} floor. Smooth L1 uses threshold one and averages over valid entries; all population statistics exclude padded cells. Reconstruction uses Equation 20, with an empty zero or nonzero group contributing zero. Paired-expression supervision, separate latent variance/covariance regularizers, population separation, and temporal ordering are disabled for the genetic instance. The genetic results evaluate this complete training recipe; the reported benchmark comparisons do not isolate the contributions of OT pairing, continuous transport, or individual auxiliary objectives.

Table S6: Shared configuration for the nine genetic runs. Each holdout regime is trained independently with seeds 42, 123, and 2026.

Component	Setting
Expression encoder; cell latent	MLP $5041 \to 512 \to 256$
Decoder	MLP $256 \to 512 \to 5041$; Softplus
Perturbation representation	Separate 512-dimensional gene embeddings; mean pooling
Dropout	Encoder/decoder 0.1; vector field 0
OT	Median-normalized squared latent cost; $\varepsilon = 0.05$; 30 iterations
Batch	8 conditions; 128 source and target cells per record
Optimizer; maximum updates	AdamW, learning rate 10^{-4}; 50,000
Checkpoint selection	Validation condition-mean L2; EMA decay 0.99
Inference	8 Euler steps; no added initial noise; source-anchored decoding

The drug instance uses

$$\begin{aligned}\mathcal{L}_{\text{drug}} = \mathcal{L}_{\text{FM}} + \mathcal{L}_{\text{mean}} + 0.3\mathcal{L}_{\Delta\text{-corr}} + 0.2\mathcal{L}_{\text{var}} + 0.05\mathcal{L}_{\text{proj-cov}} \\ + 0.1\mathcal{L}_{\text{flow-moment}} + 0.05\mathcal{L}_{\text{rec,src}} + 0.05\mathcal{L}_{\text{rec,tgt}} + 0.1\mathcal{L}_{\text{target-rec-moment}}.\end{aligned} \tag{22}$$

The projected-covariance term uses a fixed 2000×128 random projection with seed 42. These coefficients were selected using validation data before the drug evaluation.

C.5 TEMPORAL ORDERING AND INTERMEDIATE REQUESTS

The temporal flow-matching term uses Day 0 and Day 3 endpoint batches with the descriptor $c_{\text{end}} = (0, 3, 3)$. An auxiliary term evaluates virtual time conditions without reading measured intermediate expression. Let $\bar{z}_0$ and $\bar{z}_3$ be endpoint population means and define

$$d = \frac{\bar{z}_3 - \bar{z}_0}{\|\bar{z}_3 - \bar{z}_0\|_2}. \tag{23}$$

For $t_k \in \{1, 2, 3\}$, the auxiliary branch substitutes $(t_s, t_k, t_k - t_s)$ into the time fields and integrates the corresponding vector field over the full interval $[0, 1]$; the Day 3 prediction reuses the endpoint rollout. It then mean-pools the predicted latent population $\widehat{\bar{z}}_{t_k}$. Its progress score is

$$s_k = (\widehat{\bar{z}}_{t_k} - \bar{z}_0)^\top d, \tag{24}$$

and adjacent scores are separated by margin $m = 0.02$:

$$\mathcal{L}_{\text{traj}} = \frac{1}{2} \sum_{k=1}^{2} \max(0, m - (s_{k+1} - s_k)). \tag{25}$$

The full temporal flow objective is

$$\mathcal{L}_{\text{time}} = \mathcal{L}_{\text{FM}} + 0.1\mathcal{L}_{\text{rec,src}} + 0.1\mathcal{L}_{\text{rec,tgt}} + 0.1\mathcal{L}_{\text{traj}}. \tag{26}$$

The static OT residual comparator retains OT endpoint sampling and the transition MLP dimensions. It evaluates the transition network at the source latent with its scalar time input fixed to one, defining $r_\theta(z_0, g_s, \eta) = v_\theta(z_0, 1, g_s, \eta)$, and regresses this output against $z_1 - z_0$. A requested horizon h gives the single update $\widehat{z}_h = z_0 + h\, r_\theta(z_0, g_s, \eta)$. It retains endpoint reconstruction but does not use interpolated-state velocity regression or the flow-specific ordering term. Thus, this comparator changes the training formulation as well as the generation rule; it is not an integrator-only ablation. The random-pairing flow variant retains the flow formulation and substitutes independently paired endpoints for OT samples.

Intermediate prediction keeps the endpoint descriptor fixed and maps request time to an integration horizon:

$$r(t_r) = \frac{t_r - t_s}{t_e - t_s}, \qquad h(t_r) = r(t_r)^\gamma, \qquad \gamma = 0.5. \tag{27}$$

For the Day 0–Day 3 protocol, $h(1) = 0.577$, $h(2) = 0.816$, and $h(3) = 1$. The endpoint-conditioned field is integrated over $[0, h(t_r)]$ with a 32-step Heun solver. The square-root warp is fixed by the task protocol and is not learned biological time.

Representation pretraining and transition-module training both use only Day 0 and Day 3 populations from training donors, as specified in the main temporal protocol. Measured Day 1 and Day 2 expression is excluded from both training stages and from transition checkpoint selection. Transition checkpoints are selected by raw RBF MMD on validation Day 0→Day 3 endpoint predictions. At inference, each held-out donor supplies only Day 0 expression and the requested time; measured intermediate populations are used only after generation for evaluation. Section D.9 reports interventions that exchange or fix the requested horizons.

C.6 TASK-SPECIFIC INFERENCE

Table S7: Task-specific representation and inference settings.

Task	Genes	Latent	Source encoder	Solver	Steps
Temporal	2,009	256	Trainable	Heun	32
Genetic	5,041	256	Trainable	Euler	8
Drug	2,000	512	Frozen in stage B	Heun	16

Temporal requests use partial integration as defined above. The genetic model takes and returns expression in the same ordered 5,041-gene space. Mean L2, perturbation-effect correlation, and

Energy are evaluated on this full model gene set. DE ρ is evaluated on condition-specific reference DE genes within this space. The 5,041 genes comprise 5,000 highly variable genes selected from the full quality-controlled dataset before condition splitting and 41 additional measured perturbation targets. The drug model is trained in two stages: expression reconstruction followed by conditional transition learning. Stage B freezes the source encoder and gene embedding, trains the descriptor encoder, velocity field, and decoder, and uses source-anchored residual decoding. The descriptor encoder is warm-started by response-aware pretraining on training-population responses, with validation used for selection. Only its condition modules are transferred; the temporary response head is discarded. This initialization does not make held-out target expression an inference input. Transition learning continues to update the descriptor encoder. The transition solver adds standard-normal cell-level latent noise, centered over valid source cells so that their latent mean is unchanged. The same initial-noise rule applies to auxiliary generated rollouts during training and to inference. OT pairing and the FM straight-line paths use the encoded endpoints before this rollout noise is added. In every task, target expression from the held-out unit is absent from the inference inputs.

C.7 Drug adapter and evaluation interfaces

The chemCPA adapter (Hetzel et al., 2022) uses ComPert with fixed 256-dimensional SMILES-derived features, an amortized doser, and cell-line covariates. Predictions are sampled from the predicted Gaussian expression distribution. Treatment duration is not an explicit input to this adapter. The scDFM adapter (Yu et al., 2026) uses a continuous molecular-feature, dose, duration, and cell-line interface around the PAD-Transformer backbone. The CellFlow adapter (Klein et al., 2025) uses the same types of condition inputs with an OT flow-matching model and a 50-component PCA representation fitted on training cells. All three are project adaptations, not unmodified reference benchmarks.

The common evaluator receives predictions on the ordered 2,000-gene expression axis. Its distribution distances use a separate Gaussian random projection with 50 columns and seed 42; this is not CellFlow's internal PCA space. The multiscale RBF bandwidth is calibrated on training populations. This evaluation projection is also distinct from the 128-dimensional projection used by the drug model's covariance training loss. Adapter-specific generation procedures are retained; outputs are not forced to share a common sampling model.

C.8 Additional drug-response analyses

The supplementary Stage-B analysis uses training seeds 42, 123, and 2026 with a shared frozen seed-42 expression backbone and condition warm-start. Table S8 reports condition-weighted scores for each trained checkpoint. The aggregate pMMD and pEnergy are 0.031 ± 0.001 and 0.197 ± 0.010; these SD describe training-seed variation, separately from the fixed-model test repeats in the main comparison.

Table S8: Seed-level Stage-B results for TempoBridge on the compound-disjoint sci-Plex3 split, using a shared frozen seed-42 expression backbone and condition warm-start. Checkpoints were selected on validation data before evaluation on the compound-disjoint split. Each seed's score weights conditions equally; SD is the sample SD across the three training seeds, not a compound-bootstrap interval.

Metric	Seed 42	Seed 123	Seed 2026	Mean	SD
L2↓	4.515	4.498	4.513	4.509	0.010
MSE↓	0.011	0.011	0.011	0.011	< 0.001
MAE↓	0.074	0.074	0.075	0.074	< 0.001
Gene r ↑	0.976	0.976	0.977	0.976	< 0.001
Δr ↑	0.418	0.417	0.441	0.425	0.014
DE ρ ↑	0.340	0.285	0.277	0.301	0.034
DE@100↑	0.261	0.257	0.282	0.267	0.013

The main comparison evaluates the full two-stage drug pipeline and uses one fixed checkpoint per method with test seeds 42, 123, and 2026. Its condition-weighted pEnergy means are 0.2123 for TempoBridge, 0.2260 for Conditional MLP, and 0.2219 for CFM without OT (Table 3). The separate paired analysis first averages conditions within each compound, then weights the 19 compounds

equally. Its pEnergy intervals against both baselines cross zero. These compound-weighted intervals concern a different average from the condition-weighted table means, so the main comparison supports a mean ranking without establishing a compound-level advantage. Appendix D.6 specifies the bootstrap; it holds the fitted models fixed and does not represent training-seed variability.

Table S9: Drug-response main evaluation and supplementary Stage-B analysis. All rows weight biological conditions equally. The Stage-B row averages transition modules fitted with training seeds 42, 123, and 2026 on a shared frozen seed-42 expression backbone and condition warm-start. The main-pipeline and baseline rows evaluate one fixed checkpoint per method with test seeds 42, 123, and 2026, as in Table 3; the main-pipeline row additionally reports variance and covariance correlations. Their SDs describe training-seed and test-run variation, respectively. Cross-row differences do not isolate an OT component effect.

Method	pEnergy↓	pMMD↓	L2↓	Δr ↑	DE@100↑	Var. r ↑	Cov. r ↑
Stage-B TempoBridge	0.197±0.010	0.031±0.001	4.509±0.010	0.425±0.014	0.267±0.013	0.909±0.003	0.219±0.002
Main-pipeline TempoBridge	0.212±0.019	0.033±0.004	4.248±0.047	0.598±0.005	0.547±0.013	0.908±0.004	0.221±0.001
Conditional MLP	0.226±0.010	0.032±0.001	4.878±0.046	0.225±0.019	0.322±0.023	0.937±0.000	0.343±0.001
Conditional CFM (no OT)	0.222±0.009	0.034±0.003	4.636±0.029	0.385±0.006	0.263±0.006	0.904±0.002	0.217±0.019

The matched 2×2 conditioning analysis varies source context and cell-line information within the fixed-backbone Stage-B setting. Its three training seeds share the frozen seed-42 expression backbone and response-aware condition initialization; the descriptor encoder, field, and decoder are then trained for each seed. Table S10 separates condition-weighted means and training-seed SD from compound-weighted paired effects. Full-minus-No-context differences were 0.004 for Δr (95% CI $[0.001, 0.006]$) and 0.019 for DE ρ ($[0.008, 0.033]$). Under condition weights, pEnergy favored No source context and pMMD was unchanged at the reported precision; removing cell-line information reduced Δr from 0.425 to 0.414. The compound-weighted effects and condition-weighted changes summarize different averages.

Table S10: Matched 2×2 conditioning ablation on compound-disjoint sci-Plex3 in the fixed-backbone Stage-B setting. (a) Equal condition weights within each training seed; mean±sample SD over seeds 42, 123, and 2026. (b) Compound-weighted paired differences, Full minus No source context, with their 95% intervals; positive values favor Full. These differences use compound weights and are not obtained by subtracting the displayed condition-weighted means. Removing a condition replaces its vector with zeros without changing model capacity.

a) Condition-weighted means and training-seed SD

Arm	pEnergy↓	pMMD↓	L2↓	Δr ↑	DE ρ ↑	DE@100↑
Full	0.197±0.010	0.031±0.001	4.509±0.010	**0.425±0.014**	**0.301±0.034**	0.267±0.013
No source context	**0.193±0.005**	0.031 (SD < 0.001)	**4.506±0.017**	0.422±0.004	0.289±0.046	0.261±0.003
No cell line	0.195±0.003	0.031 (SD < 0.001)	4.539±0.006	0.414±0.003	0.295±0.055	0.260±0.005
Neither	0.196±0.004	0.031 (SD < 0.001)	4.538±0.008	0.415±0.004	0.295±0.052	**0.270±0.003**

b) Compound-weighted paired effects: Full minus No source context

Metric	Paired mean difference	95% interval
Δr	0.004	$[0.001, 0.006]$
DE ρ	0.019	$[0.008, 0.033]$
Condition-PDS	0.005	$[0.002, 0.008]$

Condition-PDS evaluates response identification among the full set of biological conditions. For each predicted mean response, candidates are the observed condition-specific mean responses relative to their source means, ranked by mean absolute (L1) distance. With K candidates and correct-condition rank r, the score is $1 - (r-1)/(K-1)$; average ranks handle ties, and larger scores are better. For the Stage-B paired analysis, condition-level metric scores are first averaged across the fitted training-seed models. Full-minus- No-context differences are then averaged within compound. A percentile bootstrap resamples whole compounds 10,000 times (seed 20260824), holding the fitted models fixed. These intervals describe compound variation after seed averaging and are distinct from the training-seed SD in panel (a).

The fixed-backbone projection analysis evaluates Gaussian dimensions 20, 50, 100, and 200 with five train-only projection seeds each, followed by PCA-50 and PCA-100. TempoBridge has lower

Energy than Conditional MLP in all 20 Gaussian spaces and 21 of the 22 spaces overall. Under PCA-50, TempoBridge's Energy and MMD are approximately 2.74% and 7.23% higher than MLP's, respectively. Under PCA-100, its Energy is approximately 1.41% lower and its MMD 3.36% higher. MMD favors TempoBridge in 8 of 22 spaces. The magnitude and direction of the distributional comparisons therefore depend on the evaluation projection.

Table S11: Fixed-backbone TempoBridge projection-space sensitivity relative to Conditional MLP. Each Gaussian dimension uses five train-only projection seeds. The final row also includes PCA-50 and PCA-100.

Evaluation space	Energy win rate	Mean Energy reduction	MMD win rate
Gaussian 20D	100.000%	7.530%	20.000%
Gaussian 50D	100.000%	8.280%	40.000%
Gaussian 100D	100.000%	9.640%	60.000%
Gaussian 200D	100.000%	10.800%	40.000%
All 22 spaces	95.500% (21/22)	8.180%	36.400% (8/22)

Finally, a scaffold-disjoint split tests stricter chemical extrapolation. This supplementary analysis uses seed 42 only.

Table S12: Supplementary scaffold-disjoint evaluation for seed 42. The test split contains 17 compounds and 220 biological conditions. This single-seed analysis is not used for the main compound-disjoint claim.

Method	L2↓	Δr ↑	DE ρ ↑	pMMD↓	pEnergy↓
TempoBridge	**4.677**	**0.414**	0.161	0.034	**0.223**
Conditional CFM (no OT)	4.793	0.387	0.171	0.035	0.242
Conditional MLP	5.060	0.239	0.361	0.034	0.259
Fingerprint KNN	5.187	0.225	**0.388**	0.037	0.279
chemCPA adapter	5.620	0.368	0.386	0.046	0.386

D Temporal Differentiation Details

D.1 Dataset and preprocessing

The data split contained 21,533 training cells from 77 donors, 3,556 validation cells from 17 donors, and 10,955 cells from 32 donors in the evaluation partition. Requiring at least 64 cells at each of Days 0, 1, and 2 yielded the fixed 15-donor evaluation cohort used by every method. Cells were retained if they had at least 500 total counts, 200 detected genes, and a mitochondrial fraction below 0.2; genes had to occur in at least three cells. Counts were normalized to a library size of 10,000 and transformed with $\log(1 + x)$. Highly variable genes were selected without using the evaluation cohort. The nine prespecified genes were added to the 2,000-gene set before model training.

D.2 Architecture and optimization

The expression encoder has widths $2009 \to 1024 \to 256$; the decoder mirrors this mapping with a Softplus output. The velocity MLP uses two hidden transformations after a 769-dimensional concatenated input. The complete model contains 9,797,593 parameters. Three-seed training used weights 1.0, 0.1, 0.1, and 0.1 for flow matching, source reconstruction, target reconstruction, and trajectory consistency, respectively. The trajectory term evaluated virtual target times $\{1, 2, 3\}$ with a ranking margin of 0.02; it did not use measured Day 1 or Day 2 expression. Optimization used AdamW with a base learning rate of 2×10^{-4}, a learning rate of 2×10^{-5} for pretrained parameters, weight decay 0.005, gradient clipping at 1.0, a batch size of 256, and bfloat16 arithmetic. Mean training cost was 2.467 ± 0.896 GPU-hours. With 32-step Heun integration, inference took 0.032 ± 0.001 seconds per population or 0.374 ± 0.170 seconds per 1,000 cells in the timing experiment.

Checkpoint selection used validation Day 0→Day 3 endpoints. Models were evaluated every 500 steps and selected by raw single-bandwidth RBF MMD, with patience 8 after two warm-up evaluations and a minimum improvement of 0.001. The selected steps for seeds 42, 123, and 2026 were 8,500,

10,500, and 4,000. These records identify independently trained checkpoint candidates; the main comparison uses one selected checkpoint for all three test repeats. Main-table uncertainty follows the fixed-model randomized-test-run definition in Section 5.1, rather than treating checkpoint identifiers as test-repeat identifiers.

D.3 Representation and gene-space sensitivity

The temporal task was rescored on the full 2,009-gene training space, the subset of 2,000 highly variable genes, and the nine-gene prespecified panel. This analysis checks whether the conclusion depends on one favorable feature set. A scratch-encoder control was also trained for three seeds. The pretrained initialization did not produce a stable advantage and is not presented as a contribution.

D.4 Source-context ablation protocol

The temporal source-context ablation uses the same 2,009-gene space, 15-donor evaluation cohort, and 30 Day 0→Day 1/2 requests as the main comparison. This separate retraining diagnostic uses training seeds 42, 123, and 2026, with both arms trained independently of the main-table models. Within each training seed, the Full and No-context arms use the same training configuration except for source-context conditioning. The No-context arm replaces the pooled vector $h(\widehat{P}_s)$ with a zero vector while retaining source-cell latents, the expression encoder, transition descriptor, OT construction, flow objective, decoder, seed, and training protocol. Aggregate descriptive results are reported in Table S13.

The matched Full arm has Formal MMD 0.378 ± 0.021 and Energy 0.135 ± 0.005, compared with the independently trained main-table Full Flow values 0.363 ± 0.014 and 0.129 ± 0.005. The context-free arm has larger between-seed standard deviations for Formal MMD (0.048 versus 0.021), Energy (0.013 versus 0.005), and variance correlation (0.047 versus 0.010). The ablation compares the two matched arms descriptively; the main-table Full Flow checkpoint is not part of this matched comparison. The supplementary between-training-seed standard deviations and the main-table test-run standard deviations describe different sources of variability and should not be compared as if they estimated the same uncertainty.

Table S13: Separately matched source-context ablation on the 15-donor temporal cohort in the 2,009-gene space. Both arms were retrained under the same training configuration and seeds, independently of the main-table models; the only difference is whether the pooled source context is supplied. Values are mean ± sample standard deviation over three training seeds. Relative change is signed so that positive values favor the full-context arm.

Metric	Full context	No context	Relative change
Formal MMD ↓	**0.378±0.021**	0.397±0.048	**4.720%**
Energy ↓	**0.135±0.005**	0.139±0.013	**3.100%**
Δ Pearson ↑	0.625±0.005	**0.639±0.010**	-2.170%
Variance Pearson ↑	**0.622±0.010**	0.613±0.047	**1.620%**
Covariance Pearson ↑	0.419±0.006	**0.421±0.005**	-0.470%
DE@100 ↑	0.515±0.005	**0.529±0.007**	-2.730%

D.5 Temporal donor context-swap intervention

The matched retraining experiment asks whether the context channel has a net effect. A separate inference-time intervention asks whether a frozen model uses the context of the requested donor. For each of the 15 donors, Day 0 cells were split into nonoverlapping query and context subsets. Each Day 1 and Day 2 request was evaluated with the correct donor context and with contexts from each of the other 14 donors in each of three seeded evaluations (42, 123, and 2026). Within each donor/day/seed request, the same checkpoint, query cells, and targets are used for all context choices. The checkpoint is held fixed throughout each paired intervention. This gives $15 \times 2 \times 15 \times 3 = 1{,}350$ evaluated inference cases. No parameter was updated and no donor was excluded after scoring.

The correct context reduced Energy by 0.00292 at Day 1 (95% CI $[0.00080, 0.00511]$) and by 0.00556 at Day 2 (95% CI $[0.00259, 0.00871]$). Both effects and intervals are absolute Energy differences. The

Table S14: Temporal donor context-swap intervention. The correct context is compared with the mean over the 14 incorrect donor contexts while query cells, requested time, checkpoint, decoder, and target remain fixed within each donor/day/seed comparison. Effects average the two days and evaluations with seeds 42, 123, and 2026, and are reported in the direction favoring the correct context; intervals use donor-level paired bootstrap conditional on the fitted checkpoints and context bank.

Metric	Correct	Incorrect mean	Effect	95% CI	Donors won
Energy↓	0.138	0.142	0.004	[0.002,0.007]	12/15
Formal MMD↓	0.385	0.392	0.007	[0.004,0.009]	12/15
Δr ↑	0.627	0.616	0.011	[0.003,0.021]	11/15
DE@100↑	0.512	0.503	0.009	[0.004,0.014]	12/15

correct-context retrieval rate was 17.780% at rank one and 45.560% within the top three, compared with a 6.670% random rank-one rate. These diagnostics indicate use of donor-specific population information; they do not establish donor identity classification or a causal biological effect.

D.6 Metric definitions and statistical unit

For genetic evaluation, Control copies the first n cells of the fixed 128-cell source panel, with n matching the target cardinality. Effect metrics use the full 128-cell source mean as their reference. Hence Control has a zero mean effect, and undefined effect correlation, when $n = 128$; smaller target populations can yield finite correlations from the prefix/reference difference. Summary correlations exclude undefined conditions. These values do not represent learned perturbation responses.

The following definitions distinguish the three real-data evaluators. Let $X = \{x_i\}_{i=1}^n$ and $Y = \{y_j\}_{j=1}^m$ denote generated and observed target cells, and let S be the observed source population. Mean L2 is $\|\bar{x} - \bar{y}\|_2$ in the declared log-expression space; Δr is Pearson correlation across genes between $\bar{x} - \bar{s}$ and $\bar{y} - \bar{s}$. The temporal, genetic, and drug main comparisons use 2,009, 5,041, and 2,000 genes, respectively. Metrics are computed within a biological condition before aggregation.

Energy distance. For representations $a_i = T(x_i)$ and $b_j = T(y_j)$, we estimate Energy distance (Székely & Rizzo, 2013) with the following U-statistic:

$$\widehat{\mathcal{E}}_U = \frac{2}{nm}\sum_{i,j}\|a_i - b_j\|_2 - \frac{1}{n(n-1)}\sum_{i\neq i'}\|a_i - a_{i'}\|_2 - \frac{1}{m(m-1)}\sum_{j\neq j'}\|b_j - b_{j'}\|_2. \tag{28}$$

The temporal evaluator uses $T(x) = x/\sqrt{G_{\text{eval}}}$, where G_{eval} is the number of scored genes (2,009 for the full-space comparisons), and the V-statistic: the within-population sums include diagonal entries and have denominators n^2 and m^2. Its final estimate is clipped at zero. The genetic evaluator uses $T(x) = x$ and Equation 28, retaining potentially negative finite-sample estimates. Drug projected Energy uses $T(x) = xR$, where $R \in \mathbb{R}^{2000\times 50}$ has independent $\mathcal{N}(0, 1/50)$ entries generated with seed 42. It uses the U-statistic followed by a nonnegative clamp. None of these Energy scores takes an outer square root. Their absolute magnitudes are not comparable across tasks or gene panels.

MMD, bandwidths, and clipping. With $k_\sigma(a, b) = \exp[-\|a - b\|_2^2/(2\sigma^2)]$, the U-statistic estimator of squared MMD (Gretton et al., 2012) is

$$\widehat{M}_U^2(\sigma) = \frac{\sum_{i\neq i'} k_\sigma(a_i, a_{i'})}{n(n-1)} + \frac{\sum_{j\neq j'} k_\sigma(b_j, b_{j'})}{m(m-1)} - \frac{2\sum_{i,j} k_\sigma(a_i, b_j)}{nm}. \tag{29}$$

Temporal Formal MMD uses the same scaled expression space as temporal Energy and bandwidth factors $\{0.25, 0.5, 1, 2, 4\}$. Each population is capped at 512 cells by lexicographically sorting expression rows and taking evenly spaced indices. The base bandwidth is the square root of the median positive squared distance among pooled within- and cross-population pairs. Kernels are accumulated in blocks of 256 cells. The reported value is $\sqrt{\frac{1}{5}\sum_a \max\{0, \widehat{M}_U^2(a\sigma_0)\}}$: the implementation clips each bandwidth's estimate before averaging. Raw `mmd_rbf`, used for temporal endpoint checkpoint selection, instead takes the square root of a nonnegative-clipped, single-bandwidth V-statistic. The

main-table Formal MMD field is `formal_mmd_rbf_multiscale`; it is distinct from that selection metric.

Genetic MMD uses unscaled log expression and factors $\{0.5, 1, 2, 4\}$. For each regime, the base bandwidth is fixed from 2,048 pooled validation target cells before test evaluation; the base values for Seen 2/2, Seen 1/2, and Seen 0/2 are 17.76577, 16.78579, and 18.10347, respectively. Drug pMMD uses the fixed Gaussian-50 projection and the same four factors, with its base bandwidth calibrated on training cells. These two evaluators report $\sqrt{\max\{0, \frac{1}{4}\sum_a \widehat{M}_U^2(a\sigma_0)\}}$, clipping after averaging over bandwidths. The 128-dimensional projection used by the covariance training loss is separate from the drug evaluation projection. A quantity explicitly labeled MMD^2 has no outer square root.

Differential-expression recovery. Temporal DE@100 selects the 100 largest absolute changes in the mean log-expression vector, separately for prediction and observation relative to the same Day 0 mean, using `torch.topk`. It reports the intersection size divided by 100; it does not filter genes by a significance test. Drug DE@100 uses the same overlap denominator but ranks absolute pseudobulk $\log_2$ fold changes. Ties use stable sorting in the fixed gene order. The genetic evaluator's auxiliary `de_overlap_100` field instead reports the Jaccard index of its top-100 fold-change sets; it is not the temporal DE@100 metric and is not the DE ρ reported in Tables S2 and S3.

For genetic and drug DE ρ, the fold change of population X is

$$\text{LFC}_g(X, S) = \log_2 \frac{\frac{1}{n}\sum_i \max\{\text{expm1}(x_{ig}), 0\} + \eta}{\frac{1}{|S|}\sum_{s\in S} \max\{\text{expm1}(s_g), 0\} + \eta}. \tag{30}$$

The pseudocount η is 10^{-8} for genetics and float64 machine epsilon for drugs. Observed source and target cells alone determine the DE gene set: genetics uses two-sided Mann–Whitney U tests with asymptotic p-values and continuity correction, whereas drugs use Welch t-tests on log-expression values. Both apply Benjamini–Hochberg correction with $q < 0.05$. DE ρ is Spearman correlation of predicted and observed LFCs on that set, using average ranks for ties and requiring at least 10 genes. Undefined correlations remain missing and are excluded from the valid metric denominator, rather than replaced by zero.

Aggregation and uncertainty. In the temporal main and daywise tables, each randomized test run first averages the 15 donor-specific scores at each requested day, then averages Day 1 and Day 2 equally. Genetic evaluation averages equally across retained perturbation conditions within each setting and test run. Table 2 reports setting-specific means; Table S3 additionally reports the equally weighted average of the four settings. Drug evaluation pools replicates within each cell-line/compound/dose/time condition before scoring and averages the 248 conditions equally within each test run. For each learned method and training setting, the same selected checkpoint is used in all three test runs, with random seeds 42, 123, and 2026. These seeds control the randomized evaluations, not three independently trained models. Let a_r denote the aggregated metric from run $r \in \{1, 2, 3\}$. The reported mean is $\bar{a} = (a_1 + a_2 + a_3)/3$, and its sample SD is $s = \sqrt{\sum_{r=1}^{3}(a_r - \bar{a})^2/(3-1)}$. The SD is computed from run-level aggregates, rather than pooling cells, conditions, or checkpoints across runs. It quantifies repeated-evaluation variability conditional on the selected checkpoint; it is neither a standard error nor a biological-unit confidence interval. The genetic tables report means without SD columns. Separately labeled supplementary training-seed analyses aggregate one score per independently trained checkpoint using the training-seed denominator stated in the corresponding caption. Those checkpoints do not contribute to the main-table test-run SD. Cells are not independent replicates for these uncertainty summaries.

For context-swap intervals, the 14 incorrect contexts are first averaged within each donor/day/seed request. Their paired difference from the correct context is averaged across the three seeded evaluations within a donor; the overall analysis additionally averages the two days. The 95% percentile interval uses 10,000 bootstrap resamples of the 15 donors, with seed 20260825. For the drug main-pipeline comparisons, condition scores averaged across the three test runs are first averaged within each compound, then 19 whole compounds are resampled 10,000 times, with seed 20260825. Thus those intervals weight compounds equally, whereas the drug main table weights conditions equally. They estimate compound-weighted effects and are not confidence intervals for the condition-weighted

table means. These intervals describe biological-unit variation conditional on the fitted models; they do not resample training seeds or individual cells.

D.7 AUXILIARY TIME-ADAPTED SCVIDR REFERENCE

Table S15: Auxiliary 2,009-gene result for the time-adapted scVIDR implementation on the fixed 15-donor cohort with target-matched cardinality. Values follow the main temporal test-repeat protocol: mean ± sample standard deviation over test seeds 42, 123, and 2026.

Method	Energy ↓	Formal MMD ↓	Sink. ↓	Gene r ↑
scVIDR (time-adapted)	0.294 ± 0.005	0.476 ± 0.006	0.195 ± 0.002	0.726 ± 0.005

Method	Δr ↑	Var. r ↑	Cov. r ↑	DE@100 ↑
scVIDR (time-adapted)	0.604 ± 0.006	0.546 ± 0.015	0.333 ± 0.017	0.694 ± 0.003

The scVIDR implementation adapts a perturbation-response model (Kana et al., 2023) to the temporal endpoint protocol and is reported separately from the primary external-method ranking. The row uses the fixed 15-donor cohort, the 2,009-gene space, target-cardinality matching, and the three test repeats.

D.8 NON-DEPLOYABLE ENDPOINT REFERENCES

Table S16: Non-deployable endpoint references. Both methods read held-out Day 3 expression and are excluded from the external-method ranking.

Method	Energy ↓	Formal MMD ↓	Gene r ↑	Var. r ↑	DE@100 ↑
Endpoint interpolation	0.225	0.364	0.755	0.632	0.699
PCA interpolation	0.290	0.466	0.724	0.598	0.688

Both interpolation references use held-out Day 3 expression and do not satisfy the Day 0-only inference protocol. They are not included in the external-method ranking. Their values provide endpoint-informed reference points, not universal upper bounds: an interpolation rule can be favorable for this four-day schedule without representing the correct transition mechanism.

D.9 SUPPLEMENTARY TEMPORAL ANALYSES

Table S17: Request-time decomposition in the 2,009-gene space on the fixed 15-donor cohort. Within each randomized test run, metrics are averaged equally across the 15 donor-specific requests at each day. Values are mean ± sample standard deviation over three randomized test runs of one fixed checkpoint per variant (test seeds 42, 123, and 2026), matching the main temporal evaluation.

	Energy ↓		Formal MMD ↓		DE@100 ↑	
Method	Day 1	Day 2	Day 1	Day 2	Day 1	Day 2
TempoBridge Full Flow	**0.131±0.006**	**0.128±0.004**	**0.373±0.015**	**0.353±0.014**	**0.501±0.013**	0.624±0.009
Random-pairing Flow	0.149±0.007	0.140±0.005	0.429±0.037	0.401±0.031	0.426±0.014	**0.637±0.010**
Static OT residual	0.131±0.002	0.139±0.008	0.385±0.017	0.385±0.024	0.401±0.016	0.614±0.009

Table S17 uses the same 2,009-gene space and target-cardinality matching as the main temporal comparison. GENOT is not included because its request-time evaluation uses source-matched output cardinality. Full Flow leads Formal MMD and Energy at both request times among the three reported transport variants. For DE@100, Full Flow has the highest mean at Day 1, whereas Random-pairing Flow has the highest mean at Day 2. The aggregate scores are the equal mean of the two 15-donor request-time strata within each randomized test run; Full Flow has the highest aggregate DE@100 among these variants.

The known-transport synthetic experiment is separate from the shared-state source-context construction in Appendix E. It uses independently trained Full Flow and Static OT models with seeds 42, 123, and 2026, selected using endpoint data. Generation reads only Day 0 cells from four validation donors, with 512 cells per donor and requests at Days 1 and 2. Model input and distribution scoring use the same 2,009-gene synthetic space. Scores are averaged over the eight donor/time populations within each seed and then across seeds. Ground-truth intermediate cells and correspondences are available only to the evaluator. This diagnostic uses half the Energy V-statistic (unscaled expression) and a single-bandwidth RBF MMD^2 V-statistic, with median pooled squared distance as the squared bandwidth. Path and coupling metrics use intrinsic coordinates recovered by the fixed generator-loading pseudoinverse; coupling top-1 is the fraction whose nearest true intrinsic state is the known matching cell. These scores are not directly comparable with the real-data main table. Full Flow had lower intermediate-distribution Energy (0.158 ± 0.016 versus 0.189 ± 0.024), lower MMD^2 (0.012 ± 0.002 versus 0.014 ± 0.002), and higher coupling top-1 accuracy (0.339 ± 0.015 versus 0.223 ± 0.036). Static OT residual had lower path RMSE, velocity error, and branch error. The Full Flow training formulation yielded better distribution recovery than the Static OT residual formulation in this experiment, without lowering every pathwise error.

Integration-horizon intervention. We evaluated horizon interventions on the fixed 2,009-gene panel for three validation donors, using the seed-42 checkpoint selected at update 8,500. All interventions used 32 Heun steps and identical source inputs, with generated population sizes matched to the observed targets. Energy was averaged equally over the six donor–day requests. Keeping the Day 0–Day 3 endpoint condition fixed, the prescribed Day 1/2 horizons gave Energy 0.261. Swapping the horizons increased Energy to 0.457; applying the Day 1 or Day 2 horizon to both requests gave 0.361 and 0.358, and permutation gave 0.385. This intervention tests dependence on the requested integration progress, not whether the condition encoder learned biologically calibrated Day 1 and Day 2 time.

Table S18: Heun solver sensitivity for a frozen seed-42 Full Flow checkpoint on three validation donors (six requests), scored on the fixed 2,009-gene panel with target-matched cardinality. Scores average the six requests. One Heun step uses two velocity-network evaluations. Timing is seconds per 1,000 generated cells.

Steps / calls	Energy ↓	Formal MMD ↓	Time (s) ↓
4 / 8	0.259	0.405	0.087
8 / 16	0.259	0.405	0.126
16 / 32	0.259	0.405	0.211
32 / 64	0.259	0.405	0.363

Solver sensitivity. On validation seed 42, 8-step and 32-step Heun integration produced similar Energy (0.259 and 0.259) and Formal MMD (0.405 and 0.405). Their measured inference times were 0.126 and 0.363 seconds per 1,000 generated cells. Timing covers source encoding, transport, and decoding from Day 0 input, with one warmup and three repeats; it excludes disk I/O, host–device transfer, serialization, and metric computation. These diagnostics use three validation donors; the main table uses 15 held-out donors. Because Heun evaluates the velocity network twice per step, the 32-step setting makes 64 velocity-network calls.

D.10 Additional protocol scope

In the daywise comparison, Full Flow and Random-pairing Flow had larger Energy distances at Day 1 than at Day 2, whereas Static OT residual showed the opposite ordering. Main-table scores use target-matched output cardinality rather than source-matched cardinality.

E CONTROLLED SOURCE-CONTEXT AMBIGUITY

E.1 CONTROLLED CONSTRUCTION

The experiment is confined to a two-dimensional latent space and uses identity encoders and decoders. For each replicate, contexts A and B contain the exact same 128 shared source coordinates sampled from $\mathcal{N}((0,0), 0.15^2 I)$. Each also contains 128 context-specific source points, centered at $(-2,0)$ for A and $(2,0)$ for B. The corresponding shared target components are sampled independently around $(0,2)$ and $(0,-2)$, whereas the context-specific target components remain centered at $(-2,0)$ and $(2,0)$. Each source and target population has 256 points. Targets are shuffled after generation, and component labels are retained only for analysis. Both contexts receive the fixed descriptor $c = (0,0)$.

The generator uses 64 training, 16 validation, and 32 test replicates per context. Within a replicate, the maximum absolute difference between the A/B shared source supports is exactly zero.The target branches differ only in the population associated with those shared states. Their ideal centers are separated by four latent units.

E.2 TRAINING PROTOCOL

The experiment directly instantiates `OTCFMTransition` with latent and condition dimensions of 2, hidden width 128, no dropout, LayerNorm, and no BatchNorm. All trainable arms contain 18,050 parameters. Training uses the linear flow path, an entropic Sinkhorn coupling with $\varepsilon = 0.05$ and 50 iterations, 64 OT cells per population, and batches of 16 populations. The objective is flow-matching MSE only. AdamW uses learning rate 0.002, weight decay 10^{-5}, and gradient clipping at 1.0.

The maximum training length is 2,500 steps. Validation is performed every 250 steps using mean full-population Energy over validation replicates. Early stopping begins after 500 steps, uses patience 6, and requires an improvement of 10^{-4}. Test predictions use a 32-step Heun solver with no initial noise. Each trained arm contributes one validation-selected checkpoint, which is held fixed across three test runs. No retraining or weight updates occur between test runs. Hyperparameters were fixed before training, and no post-result tuning was performed.

E.3 EXPERIMENTAL ARMS

The descriptor-only arm retains the same model and source particles as TempoBridge but sets the pooled source-context slot to zero. TempoBridge uses the mean of all 256 empirical source particles; the 64-point OT subsample does not redefine this context. Swapped context uses the frozen TempoBridge checkpoint and exchanges A/B empirical context vectors only at inference, without retraining. The oracle arm is trained with fixed vectors $(-1,0)$ and $(1,0)$ for contexts A and B; it provides a diagnostic reference with explicit context labels.

E.4 EVALUATION METRICS AND TEST-RUN AGGREGATION

For each test replicate, D_{ctx} is the mean Euclidean distance between the two predicted endpoints of each identical shared source point under contexts A and B. Shared error is the mean of the distances from the predicted shared-component centroids to the ideal centers $(0,2)$ and $(0,-2)$. Branch accuracy counts a shared prediction as correct when its second coordinate is strictly positive for A or strictly negative for B, then averages over shared points and contexts; zero is counted as incorrect. Energy evaluates the full predicted and target populations.

Table 4 reports the mean and sample SD across three test runs of one fixed checkpoint per trained arm. Each run contributes one aggregate score per metric. The sample SD uses the three run-level scores and denominator $3-1$; it describes test-run variability conditional on the selected checkpoint. The swapped-context setting uses the same TempoBridge checkpoint as the correct-context setting, changing only the context input.

E.5 VELOCITY FIELDS AND MODEL TRAJECTORIES

Figures S1 and S2 visualize the conditional fields and shared-state trajectories in the controlled synthetic system.

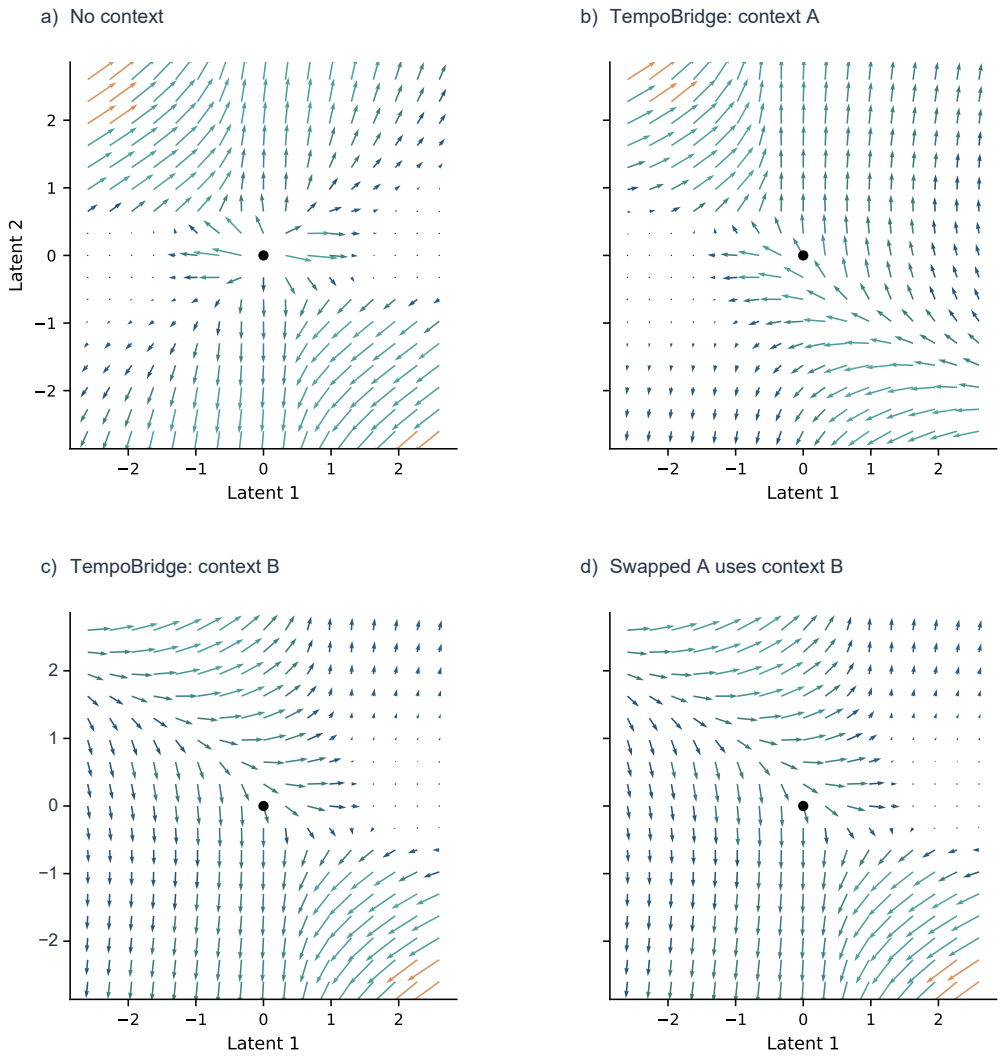

Figure S1: Velocity fields at $\tau = 0.5$ for seed 42. The context-free field is identical for A and B at a shared state; empirical contexts produce distinct fields, and exchanging the context input exchanges the field. These are diagnostics in the controlled synthetic latent system.

Table S19: Checkpoint-level results for the controlled source-context experiment. Seed indexes five training checkpoints, separately from the fixed-model randomized test repeats in Table 4. Branch accuracies are tabulated as fractions. S_{ctx} is a post hoc directional metric computed from fixed predictions without retraining or checkpoint reselection.

Arm	Seed	D_{ctx}	S_{ctx}	Shared error	Branch acc.	Energy	MMD
Descriptor-only	42	0.000	0.000	2.001	0.500	0.457	0.273
	123	0.000	0.000	2.001	0.500	0.456	0.273
	2026	0.000	0.000	2.001	0.500	0.459	0.272
	31415	0.000	0.000	2.001	0.500	0.458	0.272
	27182	0.000	0.000	2.000	0.500	0.457	0.272
Swapped context	42	3.899	-3.882	3.941	0.013	3.430	0.779
	123	3.867	-3.825	3.913	0.023	3.080	0.708
	2026	3.883	-3.849	3.925	0.020	3.215	0.754
	31415	3.797	-3.754	3.878	0.023	3.504	0.775
	27182	3.896	-3.869	3.935	0.019	3.419	0.778
TempoBridge	42	3.899	3.882	0.079	0.987	0.006	0.000
	123	3.867	3.825	0.120	0.977	0.005	0.000
	2026	3.883	3.849	0.092	0.980	0.006	0.000
	31415	3.797	3.754	0.144	0.977	0.005	0.000
	27182	3.896	3.869	0.091	0.981	0.006	0.000
Oracle context	42	3.901	3.884	0.079	0.987	0.006	0.000
	123	3.883	3.840	0.118	0.975	0.006	0.000
	2026	3.875	3.841	0.096	0.979	0.006	0.000
	31415	3.929	3.894	0.100	0.987	0.005	0.000
	27182	3.893	3.865	0.092	0.981	0.006	0.000

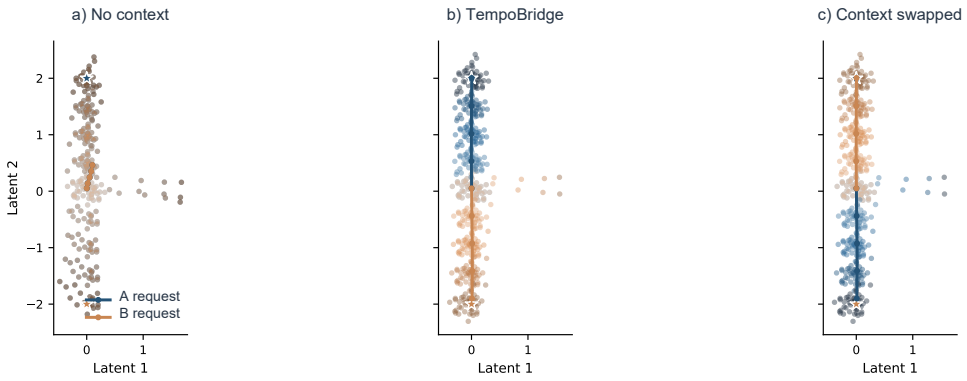

Figure S2: Model trajectories of the identical shared source states for seed 42. TempoBridge separates the requested branches, while the context-free model overlaps and swapped context reverses the separation. The curves are synthetic ODE trajectories, not reconstructed biological lineages.

E.6 VELOCITY PROBE

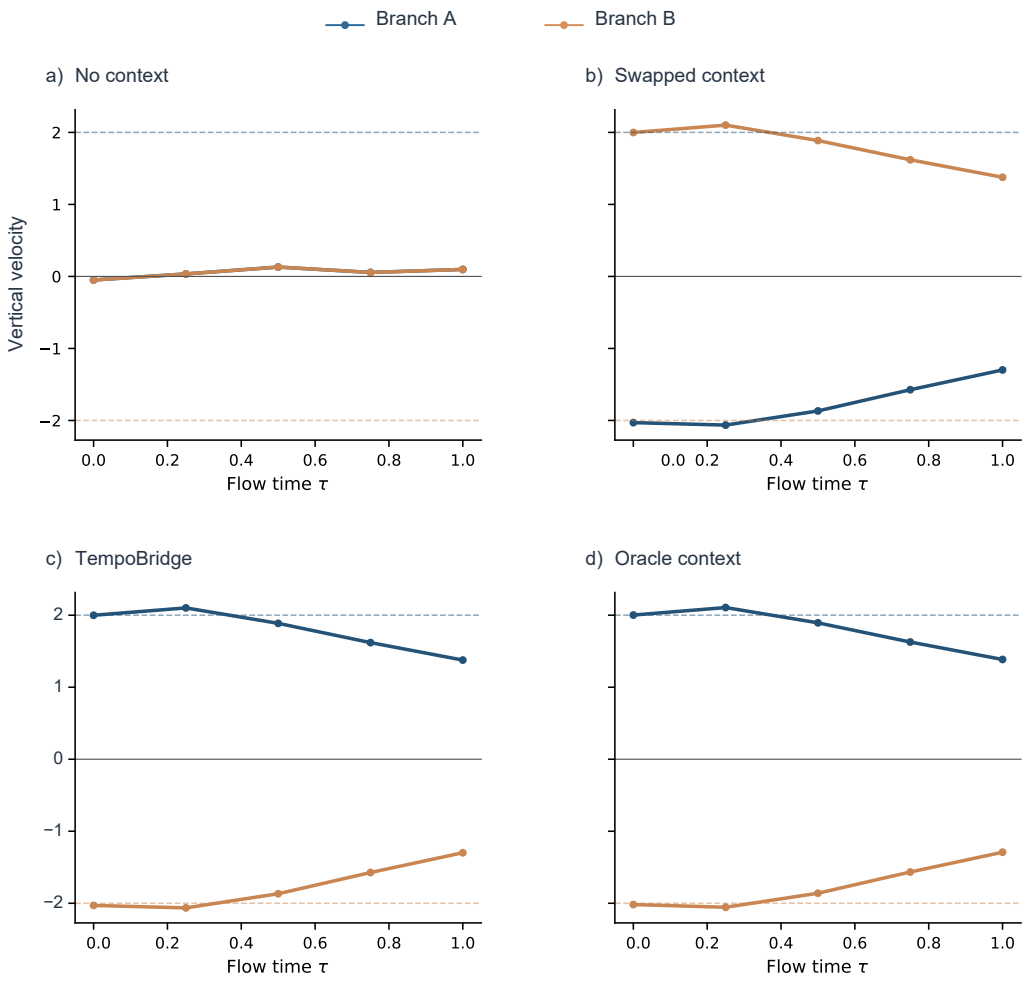

Figure S3: Velocity y-component at the identical state $z^\star = (0, 0)$ across flow time for seed 42. The context-free field returns identical velocities for A and B. TempoBridge assigns opposite directions, and swapping the empirical contexts reverses them. Dashed lines mark reference shared-component centroid displacements $+2$ and -2 over the unit transport interval, not exact conditional velocities at every probed time.

At $\tau = 0$, the context-free velocity has $y = -0.050$ for both branches. TempoBridge gives 1.999 for A and -2.030 for B; the swapped-context intervention exchanges these values. This fixed-state probe complements the population-level metrics by checking the velocity interface directly.

E.7 CONTROLS FOR THE AMBIGUITY CONSTRUCTION

The A/B shared-support difference, condition difference, no-context pointwise output difference, and maximum no-context D_{ctx} are all zero. Parameter counts match across the descriptor-only, TempoBridge, and oracle trainable arms, and no arm contains BatchNorm. The comparisons in Table 4 assess shared-cell branch accuracy and full-population Energy with no context, correct context, and swapped context. All three test runs use one fixed checkpoint per trained arm. Swapping context changes neither the TempoBridge parameters nor the shared source coordinates.