LEARNING INDIVIDUAL DYNAMICS FROM SPARSE CROSS-SECTIONAL SNAPSHOTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Forecasting how an individual unit, such as a patient, a community or a machine, will evolve usually requires repeated measurements of that unit, yet many datasets are cross-sectional: many units, each observed once. Sequence models need the per-unit sequences such data lack, while optimal transport and flow matching describe the population rather than the individual. We introduce CADENCE, which forecasts individual continuous-time trajectories from a single observation per unit by anchoring latent dynamics to static context. A score-based encoder maps observations into a latent space with a fixed Gaussian marginal, and a context-routed mixture of experts blends a shared basis of dynamical archetypes into the per-unit parameters of a latent neural ODE, trained by matching per-subgroup marginals across units. We show that marginal matching alone cannot identify individual dynamics, and give structural conditions under which the per-subgroup dynamics are identifiable within a fixed encoder frame; these conditions motivate the design. On six synthetic benchmarks, CADENCE attains the lowest individual-level error among published cross-sectional methods on five and the highest per-unit skill on all six, while conditional transport trained on full trajectories remains more accurate where such data exist. On real-world lineage-traced haematopoiesis data, trained on day-2 cells and unlabelled later-day populations without lineage or fate labels, CADENCE achieves strongest clonal fate predictions with AUROC 0.82, above published unsupervised dynamics methods (0.71–0.78).

1 Introduction

Understanding unit-level dynamics (how individuals age, diseases progress, or engineering systems degrade) is central across the quantitative sciences, and dense per-unit longitudinal data supervises it directly. Often such data cannot be obtained: in biomedical cohorts and plasma diagnostics each unit is observed *once*, or a few times irregularly. We study individual continuous-time forecasting when no unit has longitudinal measurements.

Method	Persistent dynamics	Context-based dynamics	Cross-sectional
OT / flow matching	✗	✗	✓
Conditional transport	✗	✓	✓
Latent ODE	✓	✗	✗
Context models	✓	✓	✗
CADENCE (ours)	✓	✓	✓

Table 1: Capabilities of representative method families. "Persistent dynamics": one integrated trajectory per unit, with per-unit parameters (measured by the increment-error ratio, App. I.10).

Inferring an individual vector field from unpaired cross-sectional data is an ill-posed inverse problem, and existing methods either model the population from single observations or model individuals from repeated observations: sequence models such as Latent ODEs need long per-unit sequences, cross-sectional transport moves aggregate distributions without per-unit information, and contextual dynamics models need dense sequences to fit their routing. We call a forecast *individual* when it carries per-unit dynamical parameters rather than being a per-unit sample from a conditional law; **the method families of Tab. 1 do not combine all three capabilities under that definition.**

Mechanism. A cross-section holds one observation per unit but many units: **a late-stage unit's snapshot constrains the future of a contextually similar early-stage unit**, so units with similar context share information that no single record contains. A learned router maps static context to convex weights over a shared basis of dynamical regimes, so similarity follows from the learned map,

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

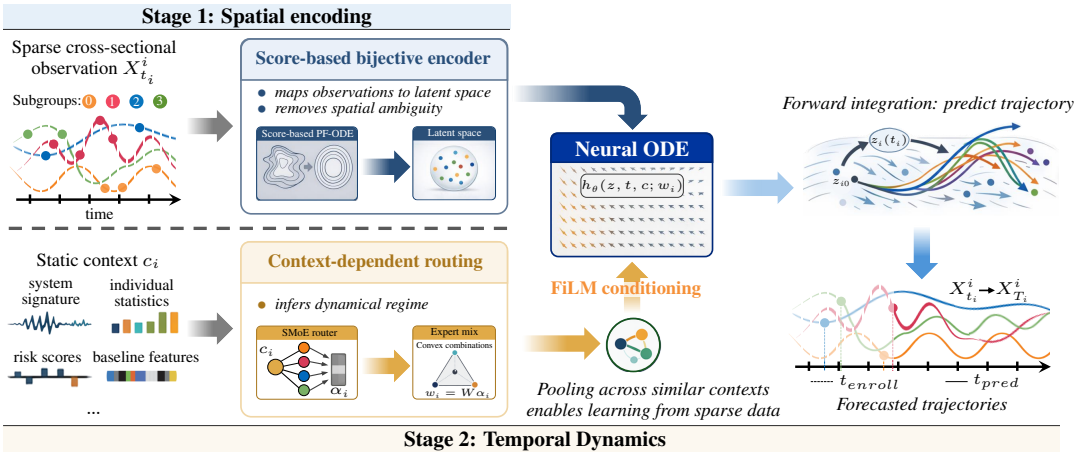

Figure 1: CADENCE overview. **Stage 1** maps each high-dimensional observation $\mathbf{x}^i_{t_i}$ to a latent state $\mathbf{z}^i_{t_i}$ via a score-based bijective PF-ODE, Gaussian-pinning the latent space. **Stage 2** routes the unit's static context $\mathbf{c}_i$ through the SMoE gating network to produce a convex expert mixture $\mathbf{w}_i$, which conditions the Neural ODE. Forward integration yields individual future trajectories.

not a pairwise search. We call this framework CADENCE (Contextual Archetypes and Diffusion ENCodings for Dynamics Estimation; Fig. 1). Our contributions:

1. **A method for individual trajectories from single snapshots.** CADENCE combines a score-based bijective encoder, a context-routed mixture of ODE experts and a per-subgroup cross-sectional objective, so that late-stage snapshots supervise contextually similar early-stage units without any pairing (Sec. 4).
2. **A conditional identifiability analysis that motivates the design.** Within a fixed encoder frame it reduces per-leaf identifiability to an assumed flow observability (FOA) and gives a Liouville sufficient condition for FOA (Theorems 1–2, Prop. 4); Sec. 5 states how the deployed system differs from the analysed class.
3. **Evidence across seven benchmarks.** Trained on one snapshot per unit, CADENCE attains the lowest individual-level error among published cross-sectional methods trained on the same data on five of six synthetic benchmarks, the lowest mean population-level error on all six, and the highest per-unit skill on all six, against below 0.30 for every published method in observation space and at most 0.74 for a state-conditioned sampler.

2 Problem Formulation

We first formalise the generative model and discuss non-identifiability. We consider N units, each observed once at $t_i \in \mathcal{T}$ with measurement $\mathbf{x}^i_{t_i} \in \mathbb{R}^p$ and static context $\mathbf{c}_i \in \mathbb{R}^{r_s}$ (Tab. 7). Dynamics evolve in a latent space $\mathbb{R}^q$ under a neural ODE h_θ, encoded by a bijection $\mathrm{Enc}_\psi : \mathbb{R}^p \to \mathbb{R}^q$ with decoder $\mathrm{Dec}_\psi := \mathrm{Enc}^{-1}_\psi$ ($q = p$ except on Benchmark BM4, App. I.2). Given a baseline time t_0 and an initial latent state $\mathbf{z}^i_0$, individual trajectories satisfy

$$\frac{d\mathbf{z}^i_t}{dt} = h_\theta\big(\mathbf{z}^i_t,\, t,\, \mathbf{c}_i;\, \mathbf{w}_i\big)\,, \qquad \mathbf{z}^i_{t_0} = \mathbf{z}^i_0, \tag{1}$$

where $\mathbf{w}_i \in \mathbb{R}^d$ is a unit-specific parameter vector (detailed in Sec. 3). Each unit in the cross-sectional dataset $\mathcal{D} = \{(\mathbf{x}^i_{t_i}, t_i, \mathbf{c}_i)\}_{i=1}^N$ contributes one draw from the time-t_i marginal μ_{t_i} and never a pair of times, so the data identify the marginals $\{\mu_t\}$ but not how individual units move between them, so h_θ is not identified from the marginals alone (Proposition 1).

Proposition 1 (Non-identifiability from marginals)**.** *Fix any frame. If $q \geq 2$, $\rho_0 = \mathcal{N}(\mathbf{0}, \mathbf{I})$ and $A \neq 0$ is skew-symmetric, the fields $h_w(z) = wAz$ with scalar w generate identical marginals for every w and t, yet distinct trajectories; marginal matching alone cannot identify individual dynamics.*

Proof is given in App. B. We frame cross-sectional training as a measure-theoretic inverse problem: we seek a latent vector field h_θ whose flow transports the baseline measure $p(\mathbf{z}_0)$ to match empirical marginals across enrolment times. Prop. 1 is the *temporal* ambiguity; a *spatial* one arises, since any

bijective re-encoding fits the data. We reduce the latter by choosing Enc_ψ as a score-based PF-ODE, whose Gaussian pinning of the aggregate latent law constrains the residual symmetry group:

Proposition 2 (Residual Spatial Symmetry). *Let $\nu := E_\# p_{\mathrm{data}}$, where $E : \mathbb{R}^p \to \mathbb{R}^q$ is a (lossless) compressor (the identity, $q = p$, on BM1–BM3 and BM5–BM7; BM4's headline run uses the lossy PCA-10 pre-compressor of App. I.2 and so lies outside Theorems 1–2) and let $P, \tilde{P} : \mathbb{R}^q \to \mathbb{R}^q$ be two* Gaussianising *encoders for ν, i.e. diffeomorphisms with $P_\#\nu = \tilde{P}_\#\nu = \gamma_q := \mathcal{N}(\mathbf{0}, \mathbf{I})$. Then $\varphi := \tilde{P} \circ P^{-1}$ lies in $G := \{\varphi \in \mathrm{Diff}(\mathbb{R}^q) : \varphi_\#\gamma_q = \gamma_q\}$, a* proper *subgroup of* $\mathrm{Diff}(\mathbb{R}^q)$. *For $q \geq 2$, G strictly contains the orthogonal group and is non-compact and infinite-dimensional in the C^1 topology (proof and structure: App. B.2).*

Denoising score matching pushes the data bijectively to $\mathcal{N}(\mathbf{0}, \mathbf{I})$ (exactly only in the terminal-law idealisation), so two Gaussianising encoders agree up to $\varphi \in G$ ($\varphi = \mathrm{id}$ in that idealisation; App. B.2). The *temporal* problem remains; Sec. 3 states when it is identifiable within the analysed class.

3 Structural Assumptions and Heterogeneity Model

The assumptions below make the pooling precise: dynamics are shared within subgroups, membership is encoded in static context, and the parameter space is finite-dimensional and regular.
Terminology: a *leaf* L_λ is a set of contexts sharing one parameter $w(\lambda)$; *subgroup* is the population of units in a leaf; an *expert* (or *archetype*) is a column w_k of the learned basis $\mathbf{W}$; *regime* refers to the dynamics a parameter produces; and *fate* is used only for BM7's terminal cell fates.

Assumption 1 (Dynamical Foliation Assumption, DFA). *There exists a measurable partition $\{L_\lambda\}_{\lambda\in\Lambda}$ of the context space $\mathcal{C}$, indexed by Λ, such that any two units i, j with $c_i, c_j \in L_\lambda$ share the same dynamical parameter: $w_i = w_j = w(\lambda)$.*

DFA partitions the ensemble into leaves that follow the same ODE parameters, which realises the surrogate supervision above (measure-theoretic basis: App. E; formal conditions: App. B.3).

Assumption 2 (Context Observability, COA′). *There exists a measurable function $g^* : \mathcal{C} \to \Lambda$ such that $\lambda(i) = g^*(c_i)$ for all $i \in \{1, \ldots, N\}$, known in one of two senses: **(a)** g^* is directly computable from context; or **(b)** g^* exists but is unknown; under per-context flow observability (FOA, Sec. 5) and the leaf separation assumption (SEP; Assumption 6 in App. B.3) it is recovered from the joint law of (x_{t_i}, t_i, c_i) (Theorem 1).*

The two readings impose different proof requirements, since existence of g^* alone does not guarantee leaf *recovery*. g^* gives only the leaf index, not $w(\lambda)$: under reading (a) the partition is supplied and Theorem 1 recovers the per-leaf parameters, carrying the auxiliary-variable identifiability of iVAE (Khemakhem et al., 2020) to the dynamical setting; under reading (b) the leaves must be recovered as well, which the theorem covers only under SEP (App. B.3). The trained router is unsupervised (App. C.4), so the routing accuracies we report are empirical evidence about reading (b). BM1–BM4 satisfy reading (a) by construction; on BM5–BM7 no deterministic g^* exists and only its Bayes-classifier relaxation holds (Tab. 7).

Two data assumptions are consumed by Thms. 1–2. **BAS (baseline observability):** t_0 is a right-adherent point of each leaf's window $\overline{T}_\lambda$, and ρ_0^λ is the t_0-marginal (one-sided limit) of $\{\rho_t^\lambda : t \in T_\lambda\}$, a data-generating property, not supplied by Gaussian pinning. Both conditions are read per context, so enrolment must charge T_λ for a.e. $\mathbf{c}$ (positivity, POS; App. B.4). **ENR (enrolment ignorability):** the enrolment time is independent of the unit's latent state given its context, $t_i \perp \mathbf{z}_{t_0}^i \mid \mathbf{c}_i$, so that $p(\mathbf{z}_{t_i} \mid \mathbf{c}, t_i) = \rho_{t_i,\mathbf{c}}^\lambda$. Without it the conditional law is a state-biased version of the leaf marginal and Step 1 of Thm. 1 fails; it holds by construction on BM1–BM6 and by BM7's fixed culture-day design.

Assumption 3 (Parameter-Set Regularity, MRA). *The parameter image $\mathcal{W} = \{w(g^*(c)) : c \in \mathcal{C}\} \subset \mathbb{R}^d$ is compact. If Λ is finite, $\mathcal{W}$ is finite; if Λ is a continuum and $c \mapsto w(g^*(c))$ is Lipschitz on the compact context domain $\mathcal{C} \subset \mathbb{R}^{r_s}$, then $\dim_{\mathrm{H}} \mathcal{W} \leq r_s$. $\mathcal{W}$ is contained in the ϵ-neighbourhood of the convex hull of K points $\{w_1, \ldots, w_K\} \subset \mathcal{W}$:*

$$w_i \approx \textstyle\sum_{k=1}^K \alpha_{i,k}\, w_k, \quad \alpha_i \in \Delta^{K-1}, \quad \textit{error} \leq \epsilon. \tag{2}$$

MRA casts individual dynamics as a blend of "archetypal" behaviours, and constrains rather than fixes the expert count by a case split proved in App. B.5: for finite Λ, $K = |\Lambda|$ represents $\mathcal{W}$ exactly and minimally under affine independence; for continuous Λ, a Lipschitz image admits a finite ϵ-net. The analysis concerns parameter recovery rather than forecast error: on BM1 both are minimised at

K=3, but forecast error varies $1.29\times$ over $K \in \{2, \dots, 5\}$ while the recovery error of the archetype parameters varies $3.6\times$ (0.08 against 0.29 at K=2 and 0.13 at K=5; App. J.10).

4 Architectural Design and Training Objective

CADENCE instantiates Eq. 1 with three components: a contextual SMoE that computes $\mathbf{w}_i$ from static context alone (DFA, COA′) as a convex combination of experts (MRA), FiLM conditioning that injects $\mathbf{w}_i$ into h_θ, and a score-based PF-ODE as Enc_ψ, reducing the spatial ambiguity (Prop. 2).

Contextual SMoE. The context router is an MLP $g_\omega : \mathbb{R}^{r_s} \to \mathbb{R}^K$. To satisfy MRA constraints, routing weights are computed as $\boldsymbol{\alpha}_i = \mathrm{Softmax}(g_\omega(\mathbf{c}_i)) \in \Delta^{K-1}$. This smoothly blends a shared expert basis $\mathbf{W} = [\mathbf{w}_1, \dots, \mathbf{w}_K] \in \mathbb{R}^{d \times K}$ to produce individual trajectory parameters $\mathbf{w}_i = \mathbf{W}\boldsymbol{\alpha}_i$.

FiLM Conditioning. The expert mixture $\mathbf{w}_i$ conditions the ODE via FiLM (Pérez et al., 2018), with FiLM scale s and shift b: $h_\theta(\mathbf{z}, t, \mathbf{c}; \mathbf{w}_i) = s(\mathbf{w}_i) \odot h_{\mathrm{base}}(\mathbf{z}, t, \mathbf{c}) + b(\mathbf{w}_i)$. App. D discusses hypernetwork alternatives.

Score-Based PF-ODE. Instantiating Enc_ψ as a score-based PF-ODE (Song et al., 2021) maps the aggregate data law to $\mathcal{N}(\mathbf{0}, \mathbf{I})$, which we call *Gaussian pinning* (Prop. 2). The unconditional encoder is bijective and dimension-preserving, so Stage 2 works in one latent frame against a reference measure, without a variational encoder's reconstruction noise. On BM1–BM6 the deployed encoder is a context-indexed family $P_{\mathbf{c}}$, giving one frame per context, and h_{base} an augmented ODE with five scratch coordinates; BM7 runs the unconditional encoder and the plain ODE (Tab. 8).

Decoupled Training Strategy. End-to-end backpropagation through cascaded spatial and temporal ODEs costs $O(\mathrm{Steps}_{\mathrm{spatial}} \times \mathrm{Steps}_{\mathrm{temporal}})$, so we split training in two. *Stage 1* trains Enc_ψ by denoising score matching on the training split alone, then freezes it and encodes all observations offline. *Stage 2* trains $(h_\theta, \mathbf{W})$ entirely in $\mathcal{Z}$-space at $O(\mathrm{Steps}_{\mathrm{temporal}})$ per step (a count of loop iterations) by forward integration, $\hat{\mathbf{z}}_t^i = \mathbf{z}_{t_i}^i + \int_{t_i}^t h_\theta(\mathbf{z}_\tau^i, \tau, \mathbf{c}_i; \mathbf{w}_i)\, d\tau$.

Cross-sectional distribution objective. Consistency is enforced *per subgroup*: with routing weights $\alpha_{i,k}$, the predicted marginal is $\hat{\mu}_t^{(k)} \propto \sum_{i:t_i \le t} \alpha_{i,k} \delta_{\hat{\mathbf{z}}_t^i}$ and the kernel-smoothed reference is $\mu_t^{\mathrm{ref},(k)} \propto \sum_j \alpha_{j,k} \kappa_h(t_j - t) \delta_{\mathbf{z}_{t_j}^j}$, so in expectation over $t \sim \mathcal{U}(t_0, t_{\mathrm{end}})$ the loss sums the per-subgroup MMD:

$$\mathcal{L}_{\mathrm{cross}} = \mathbb{E}_t \big[\textstyle\sum_{k=1}^K \mathrm{MMD}^2(\hat{\mu}_t^{(k)}, \mu_t^{\mathrm{ref},(k)}) \big]. \tag{3}$$

All headline results on BM1–BM6 use this objective with a single snapshot per unit; BM7's later-day cells carry no context, so there the reference stratification is inferred by a coupling with a mass-balance guard (Eq. 4, App. C.2). The MMD kernel is Gaussian with median-heuristic bandwidth (Tab. 9); with sparse follow-ups the exact latent MSE $\mathcal{L}_{\text{semi-long}}$ of App. I.6 applies.

The marginals must change over time: between enrolment-time bins, MMD^2 exceeds a within-bin floor by at least $119\times$ wherever measurable (App. B.9); this is necessary, not sufficient, and identification rests on FOA (Prop. 4).

Full objective. $\mathcal{L} = \lambda_{\mathrm{dyn}}\mathcal{L}_{\mathrm{dyn}} + \lambda_{\mathrm{router}}\mathcal{L}_{\mathrm{router}} + \lambda_{\mathrm{feat}}\mathcal{L}_{\mathrm{feat}}$ with $\mathcal{L}_{\mathrm{dyn}} = \mathcal{L}_{\mathrm{cross}}$ in all headline runs: $\mathcal{L}_{\mathrm{feat}}$ fits a scalar readout to an auxiliary enrolment measurement and is active on BM2 alone (every method receives it), and the unsupervised entropy term $\mathcal{L}_{\mathrm{router}}$ runs with the guard on BM7 only; elsewhere the objective is $\mathcal{L}_{\mathrm{cross}}$ (weights: Tab. 8; App. C.2). Annealing the routing temperature to $\tau_{\mathrm{end}} = 0.1$ sharpens assignments towards DFA's hard partition (App. B.8; pseudocode: App. C).

5 Identifiability Analysis

Proposition 2 leaves a residual spatial ambiguity $\varphi \in G$ that is harmless for forecasts, since conjugate models make identical decoded predictions (App. B.2), but not for training: the objective is not G-invariant, and the encoder ablation shows that the choice of latent frame determines whether routing succeeds (App. I.4). What remains is temporal: the leaf partition and the parameters $w(\lambda)$. Theorem 1 identifies both within a fixed frame, and Theorem 2 lifts this to the quotient $\mathcal{M}/G$. FOA, displayed below, is the identifiability assumption itself; Prop. 4 gives a sufficient condition in terms of the baseline density and the field difference, and its separation statistic between trained expert heads serves as a diagnostic (Tab. 5).

Assumptions consumed. From Sec. 3: DFA and COA′ on structure, and BAS (with its positivity condition POS) and ENR on the data. Two more concern the model class: parametric restriction

(PRW, Assumption 4), a finite-dimensional parametric family with the leaf parameter w separate from the shared backbone θ; and FOA, per context: for every $\mathbf{c} \in L_\lambda$, $w \mapsto \{(\Phi_{t,t_0}^{w,\mathbf{c}})_{\#}\rho_{0,\mathbf{c}}^{\lambda}\}_{t\in T_\lambda}$ is injective (App. B.3). Theorem 1 uses this per-context form; the consistency result of App. B.9 uses the leaf-marginal form, which it does not imply.

Theorem 1 (Parameter identifiability within a fixed frame)**.** *Fix a* frame*: a lossless compressor E, a Gaussianising encoder P and a backbone θ, either known or identified jointly with w under FOA$_{\mathrm{j}}$, the joint form of FOA over (θ, w). Assume DFA, COA′, BAS, POS, ENR, PRW and FOA, plus SEP under COA′(b). Then the observable joint law of $(\mathbf{x}_{t_i}^i, \mathbf{c}_i, t_i)$ determines the per-leaf parameters $\{w^*(\lambda)\}$ up to leaf-label permutation. Under reading (a) the context supplies the leaf partition; under reading (b) FOA and SEP recover it. Another admissible frame differs by some $\varphi \in G$ (Theorem 2).*

Key mechanism (full steps in App. B.4). Latents are unique given the frame; conditioning on $(\mathbf{c}, t)$ identifies each leaf's marginal family under ENR; COA′(a) supplies the partition and BAS identifies ρ_0^λ; the pushforward relation, read per context within a leaf, plus FOA injectivity then uniquely determines $w^*(\lambda)$. At population level each context is identified from its own conditional law; pooling similar contexts is a finite-sample device (stress test in App. J.1). For the idealised estimator, consistency follows from reference consistency under the leaf-marginal form of FOA; it extends to the trained objective, whose source is also empirical, given a uniform law of large numbers for the empirical leaf risk, which App. B.9 states as a hypothesis.

Theorem 2 (Composition identifiability)**.** *Under the conditions of Theorem 1, POS included, any two models in the class with unconditional encoders that exactly reproduce the cross-sectional data law differ only by a measure-preserving $\varphi \in G$ and a leaf-label permutation π: $\widetilde{\mathrm{Enc}} = \varphi \circ \mathrm{Enc}$, leaf maps agree up to π, and the flows are conjugate as measure transports on the observed windows and baseline supports. Conversely, any such conjugate reproduces the same law, by encoder bijectivity alone. If also $\varphi \in G_H$, the subgroup keeping every leaf's conjugated field in $\mathcal{H}$, and FOA$_{\mathrm{j}}$ holds, the fields themselves are conjugate, giving the finer quotient $\mathcal{M}/G_H$ (App. B.3).*

Scope. The theory motivates CADENCE's design; as with any analysis, the trained system departs from the analysed class in places. Tab. 37 (App. J) lists every difference between the class it covers and the system we train, with the measurement that bounds it: each measured difference changes MAE by at most 19% and routing accuracy by at most 0.04, and for one (hard, sample-split routing) the idealised variant is the worse. BM1 and BM2 lie within the theorems' scope; the other benchmarks extend beyond it, through BM3's continuous router (App. D.3), BM4's PCA-10 pre-compressor, the Bayes form of COA′ on BM5–BM6 and BM7's single enrolment day (POS), and show how the method behaves outside the idealised setting.

6 Related Work

Population-Level Transport. Reconstructing dynamics from unpaired snapshots is traditionally framed as optimal transport (Schiebinger et al., 2019), with scalable objectives from continuous normalising flows (Tong et al., 2023) and unbalanced OT (Sha et al., 2024). These learn a shared population transport; started from an observed unit they give a state-conditioned path, but they infer no persistent unit-specific dynamical parameters. A distinct *conditional* family (CVFM (Generale et al., 2026), COT-FM (Kerrigan et al., 2024), cMonge (Driessen et al., 2026)) maps to $P(x_t \mid c, t)$, but its per-unit output remains a draw from that law, with identity coming from source noise rather than the observed unit.

Individualised Latent Dynamics. Latent state-space models such as Latent ODEs (Rubanova et al., 2019) and context-adaptive extensions (CoDA (Kirchmeyer et al., 2022), LaDID (Lagemann et al., 2023), Neural ODE Processes (Norcliffe et al., 2021), BrainODE (Park et al., 2026)) model individualised trajectories and handle irregular sampling, but require repeated observations of each unit to constrain initial conditions or context; Soft Mixture-of-Experts (Puigcerver et al., 2024) likewise demands sequential data. CADENCE extends these routing paradigms to static covariates and sparse snapshots alone.

Identifiability and Control Theory. A central technical issue is latent identifiability (Locatello et al., 2019), resolved via auxiliary variables (Khemakhem et al., 2020), temporal constraints (Yao et al., 2021), or causal structure (Yao et al., 2024), but not for *cross-sectional* dynamics. CADENCE addresses this gap with a design motivated by nonlinear observability (Hermann & Krener, 1977), with Prop. 4 giving an explicit sufficient condition in terms of the baseline density and field differences, and a score-based encoder (Song et al., 2021) that bounds the latent ambiguity.

7 Experiments and Baselines

Benchmark suite. Validating individual trajectory inference needs continuous longitudinal ground truth, which real ensembles rarely provide, so we evaluate seven benchmarks scaling in dimensionality ($p \in [2, 500]$) and complexity. BM1–BM6 are synthetic or semi-synthetic, with simulated ground truth that permits exact evaluation of subgroups, parameters and trajectories. BM1, BM3 and BM5 are low-dimensional systems isolating routing, continuous MRA (BM3, with the continuous router of App. D.3) and the near-bifurcation limit; BM4 tests the encoder at p=500, every method receiving the same PCA-10 pre-compressor; BM2's enrolment readout is given to every method, so access is matched on all six; BM6 uses SERGIO (Dibaeinia & Sinha, 2020); BM7 is real LARRY data (Weinreb et al., 2020): each day-2 cell is a unit with the PCA-50 projection of its expression as context (state-derived routing), later-day cells form pooled, context-free references (App. H.5), and fate labels only score unsupervised fate prediction on 1,488 held-out cells. Details: Tab. 7, App. F.

Baselines. Four groups (Apps. G, I.10): (1) cross-sectional ensemble estimators (scNODE, TIGON, OT-CFM), which forecast a unit by transporting its own encoded snapshot along a single population field and so carry no per-unit parameters; (2) conditional-transport methods (CVFM, COT-FM, cMonge), whose per-unit output is a draw from $P(x_t \mid c_i, t)$ rather than an inferred state; (3) contextual dynamics models (NDP, CoDA); and (4) latent continuous-time models (Latent ODE, ODE^{2}VAE, LaDID). Groups (3) and (4) require longitudinal sequences and form the oracle block of Tab. 2. Groups (1)–(2) run cross-sectionally on CADENCE's inputs, with a full-trajectory arm in App. I.10. Every baseline is re-run and tuned on the same 3×3 grid by validation MAE, which for CVFM and COT-FM selects their published configuration (App. H.1; on BM7 under the protocol of App. H.5).

Metrics. MAE is the mean absolute error between each unit's decoded prediction and its true state, averaged over coordinates, times and units; $\mathbf{SW}_2$ is the sliced 2-Wasserstein distance between the predicted and true populations, averaged over evaluation times (true: noise-free simulator states on BM1–BM6, observed later-day cells on BM7). **Skill** against the leaf-conditional mean, $\mathrm{SS}_\lambda = 1 - \sum_{i,t} \|\mathbf{x}_t^i - \hat{\mathbf{x}}_t^i\|^2 / \sum_{i,t} \|\mathbf{x}_t^i - \bar{\mathbf{x}}_t^\lambda\|^2$ over the units of leaf λ, is the share of squared error removed beyond knowing each unit's subgroup (App. I.7). **Routing accuracy** is the fraction of units whose $\arg\max_k \alpha_{ik}$ matches the true leaf after a validation-fitted Hungarian alignment of expert indices, on every benchmark including BM7 (App. C.4). H counts steps of a shared 40-point evaluation grid spanning each benchmark's window, and each unit is scored on the grid points after its own enrolment, so no forecast leaves the evaluation window; we tabulate H=20 and H=40. Baselines are scored on one prediction per unit: a deterministic transport from its enrolment state (group (1); ground-truth initial state for oracles), one conditional draw (group (2)) or one Monge map (cMonge); App. J.2 adds median- and mean-based scoring of the samplers.

8 Results

Quantitative comparison across benchmarks. With the same cross-sectional data as every baseline, CADENCE ranks first, or level with the first, among published methods on every benchmark and both metrics (Tab. 2). On MAE it is lowest among published methods on five of six (BM1, BM3–BM6), by large margins that are significant on every one: the next-best published method's error is 5.8× CADENCE's on BM1, 4.4× on BM3, 2.0× on BM4 and BM5 and 1.6× on BM6. On SW_2 it is lowest among published methods on all six against baselines tuned for MAE (also after re-selecting the three strongest by validation SW_2), by 6–48%, with margins over CVFM on BM1 and over CVFM and COT-FM on BM3 within seed variation (App. H.2). On BM2, CADENCE and OT-CFM split the metrics: OT-CFM has the lower MAE (0.021 vs. 0.025) and CADENCE the lower SW_2 (0.050 vs. 0.055), with uncorrected p=0.013 and 0.017 that do not survive a Holm correction; CADENCE's SW_2 there depends on the BM2-specific feature loss (0.077 without it; App. F.3).

Per-unit skill beyond the leaf-conditional mean. The clearest separation is per unit. Against the leaf-conditional mean, a null that conditions on the true subgroup, skill on BM2 is 0.11 for OT-CFM, 0.24 for CVFM and 0.66 for CADENCE; no published cross-sectional baseline exceeds 0.30 on any benchmark in observation space (Tab. 41), and none exceeds 0.66 when re-fitted on CADENCE's latents (Tab. 23). The conditional samplers come close only on the population metric their objective targets: on SW_2, on BM1–BM3 and BM5 CVFM tracks CADENCE to within 6.0–18.2% and COT-FM to within 6.0–36.5%; on BM4 the gaps are 28% and 39%, on BM6 48% and 61%. On MAE, CVFM's error is 5.8× CADENCE's on BM1 and 1.2× on BM2. Scored at their point-optimal predictor instead of one draw, the samplers' MAE decreases by a factor of 1.26–1.30 and their

Table 2: MAE and SW_2 at horizon $H=20$, means over 5 seeds. **Left:** cross-sectional methods, trained on the same sparse snapshots as CADENCE with access matched on every benchmark (Apps. F.3, I.2); bold marks the best cross-sectional mean per row among published methods (after a Holm correction no contested margin is significant; App. H.2). **Adapt.:** CVFM-S, our state-conditioned CVFM (App. I.10), not a published method and not ranked. **Right (oracle):** baselines trained on full trajectories, *italic*, best per row *underlined*; three more in Tab. 11. Transport baselines are scored on one rollout per unit (Sec. 7); tuning: App. H.1; point-optimal arm: App. J.2.

		Cross-Sectional Data							Adapt.	Full Trajectory Data (Oracle)				
BM	**Metric**	**CADENCE**	OT-CFM	TIGON	scNODE	CVFM	COT-FM	cMonge	CVFM-S	OT-CFM	TIGON	scNODE	NDP	CoDA
1	MAE	**2.05**	14.82	18.64	17.93	11.97	13.97	13.41	3.41	*3.11*	*9.86*	*14.86*	*5.84*	*7.92*
	SW_2	**0.070**	0.318	0.561	0.604	0.075	0.095	0.412	0.088	*0.042*	*0.094*	*0.487*	*0.092*	*0.331*
2	MAE	0.025	**0.021**	0.108	0.131	0.031	0.036	0.162	0.024	*0.005*	*0.008*	*0.128*	*0.155*	*0.109*
	SW_2	**0.050**	0.055	0.119	0.842	0.054	0.062	0.704	0.047	*0.037*	*0.043*	*0.556*	*0.462*	*0.508*
3	MAE	**0.097**	1.402	1.427	1.118	0.424	0.524	0.581	0.168	*0.271*	*0.463*	*1.061*	*0.964*	*1.284*
	SW_2	**0.050**	0.091	0.093	0.612	0.053	0.053	0.431	0.058	*0.034*	*0.041*	*0.594*	*0.196*	*0.378*
4	MAE	**0.0222**	0.0441	0.0607	0.0913	0.0508	0.0624	0.0941	0.0281	*0.0139*	*0.0197*	*0.0864*	*0.0571*	*0.0913*
	SW_2	**0.0626**	0.0912	0.1287	0.2241	0.0801	0.0869	0.2604	0.0713	*0.0186*	*0.0281*	*0.2213*	*0.1628*	*0.8264*
5	MAE	**0.103**	0.204	0.538	0.261	0.260	0.358	0.604	0.169	*0.068*	*0.228*	*0.312*	*0.301*	*0.254*
	SW_2	**0.137**	0.441	1.284	0.712	0.162	0.187	0.831	0.158	*0.033*	*0.079*	*0.601*	*0.294*	*0.462*
6	MAE	**0.434**	0.688	1.412	1.244	0.722	0.822	1.283	0.508	*0.466*	*1.318*	*0.874*	*0.571*	*0.802*
	SW_2	**0.139**	0.291	0.702	0.708	0.206	0.224	0.471	0.171	*0.163*	*0.624*	*0.631*	*0.371*	*0.501*

SW_2 increases by a factor of 2.0–2.3; no MAE leader changes at either horizon, although at $H=20$ CVFM's point-optimal MAE on BM2 also falls below CADENCE's (App. J.2).

Access and representation controls. A state-conditioned CVFM that also receives the unit's own snapshot stays below CADENCE on skill on all six benchmarks (at most 0.74), and because it samples each time point independently its increment error relative to marginal error is 3.1–$6.3\times$ CADENCE's; on MAE it closes 58–86% of the gap on five benchmarks and matches CADENCE on BM2 (0.024 vs. 0.025, within one seed s.d.), where its SW_2 is also lower (0.047 vs. 0.050) (App. I.10).

Given full trajectories. The oracle block shows what full trajectories, which CADENCE never sees, give the baselines. OT-CFM has lower MAE and SW_2 than CADENCE on BM2, BM4 and BM5 and lower SW_2 on BM1 and BM3; CADENCE is lower on both metrics on BM6 only (Tab. 2, oracle block). CVFM and COT-FM reduce MAE by 4.2 to $31.5\times$ on BM1, BM2 and BM4 and by 3.2–$5.3\times$ on BM5, and overtake CADENCE on all four; on BM3, where the difficulty is per-unit state inference, full trajectories reduce COT-FM's MAE by 57% (0.524 to 0.227), and CADENCE's single-snapshot MAE (0.097) remains $2.3\times$ lower (App. I.10). CADENCE has lower MAE than every context-blind sequence-trained baseline of groups (3)–(4) on all six benchmarks (given the true leaf, a Latent ODE is lower on BM2; App. I.1), and their SW_2 is 1.3 to 21.5 times CADENCE's.

Where the advantage comes from. On BM1 the best sequence model, NDP, reaches 5.84 in raw observation space and 3.12 when re-fitted on CADENCE's Stage-1 latents, against 2.05; a field without router on those latents reaches 3.25, but BM1's context-indexed encoder already passes context into them (11.62 with an unconditional encoder, App. H.1), and routing lowers it by a further 37% (23–37% across BM1, BM2, BM3 and BM5; App. I.1). The cross-sectional baselines re-fitted on the same latents keep that ordering on all six benchmarks (App. I.1). On BM1, adding the router raises skill against the leaf-conditional mean from 0.13 to 0.61 and yields routing accuracy 0.937 (chance 0.333) (App. I.7).

BM1: Trajectory recovery and subgroup routing. Fig. 2(a) shows CADENCE's per-expert forecasts against ground truth in all three regimes; among the plotted methods it has the lowest RMSE over the horizon (panel c); and the SMoE router, trained without subgroup labels, assigns units to the true subgroups with 93.7% accuracy (panel b) (0.93–0.97 across retrained models; App. I.5).

Individual-level content. Against the leaf-conditional mean, skill is positive on every leaf: 0.47–0.83 on the five discrete-leaf benchmarks other than BM4 and 0.95–0.97 on BM4 (BM3: 0.96 against the weaker cohort null); within-leaf Spearman correlation is 0.53–0.95, which one trajectory per leaf could not give; and MAE is 2.6–16.2% of persistence (App. I.7).

BM6–BM7: High-dimensional and real-world stochastic dynamics. BM6 models the nonlinear evolution of a gene regulatory network ($p=100$, $K=2$). CADENCE attains the lowest MAE and SW_2 of any entry in Tab. 2 (0.434/0.139; RMSE over time in Fig. 3c), ahead of the conditional-transport baselines (CVFM 0.722/0.206, COT-FM 0.822/0.224) and every oracle model trained on

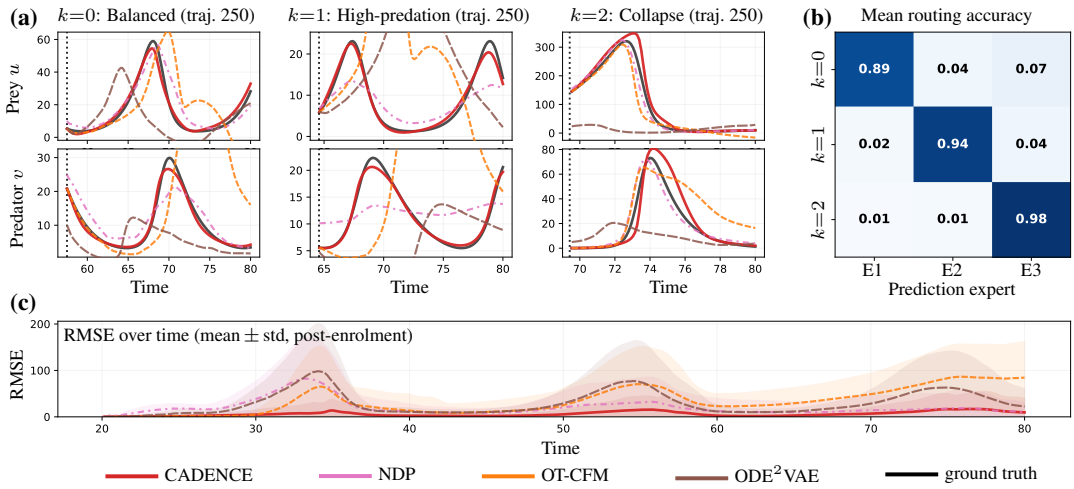

Figure 2: BM1 results on the 750 test units. CADENCE against the re-trained baselines of Tabs. 2 and 11. **(a)** Per-expert predicted trajectories vs. ground truth (axes: prey u, predator v; k indexes the subgroup/expert, named as in App. F.2). **(b)** Mean routing accuracy matrix, row-normalised; its mean diagonal (per-class recall) is 0.937 (chance 0.333; rows: ground-truth subgroup, columns: assigned expert). **(c)** Root-mean-square error per evaluation time, the Euclidean norm of each unit's decoded error over coordinates (not averaged over them, unlike MAE), mean ± s.d. over units; OT-CFM is the cross-sectional arm, NDP and ODE^{2}VAE are trained on full trajectories. The legend is shared; ground truth appears in (a) only.

full longitudinal trajectories, including the strongest of those, OT-CFM (0.466/0.163), although the MAE margin over it is within seed variation (p=0.62); the SW_2 margin has p=0.011 uncorrected and does not survive a Holm correction across the contested family (App. H.2). Given full trajectories, CVFM reaches MAE 0.415, below CADENCE's 0.434 by less than one seed s.d.; CADENCE keeps the lowest SW_2 at either access level. It recovers both subgroup trajectories rather than the ensemble mean (Fig. 3a), with 85% routing accuracy.

BM7 evaluates CADENCE on lineage-traced primary haematopoietic cells from LARRY (Fig. 4). Its unit is the day-2 cell, one snapshot per unit as elsewhere; the barcode supplies only the label, the by-clone folds, and the later-day target for the per-unit MAE, the one BM7 metric that cannot be per-cell (App. F.8). CADENCE's two experts align with the two terminal fates although no fate label enters training, and only two comparators, both given information it never uses, score higher.

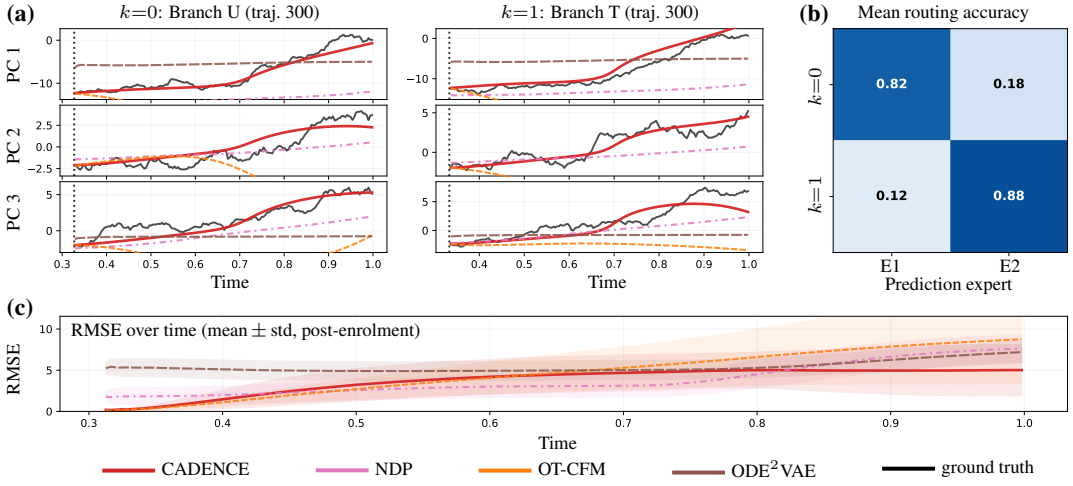

Figure 3: BM6 results on the 600 test units. CADENCE against the re-trained baselines of Tabs. 2 and 11. **(a)** Per-expert predicted trajectories vs. ground truth, shown in principal components 1 to 3 of the 100-gene expression space, **(b)** Mean routing accuracy matrix, row-normalised; mean diagonal 0.850 from static context alone (chance 0.50), below the Bayes accuracy 0.865 that the context overlap allows (App. F.7). **(c)** RMSE per evaluation time, as in Fig. 2(c).

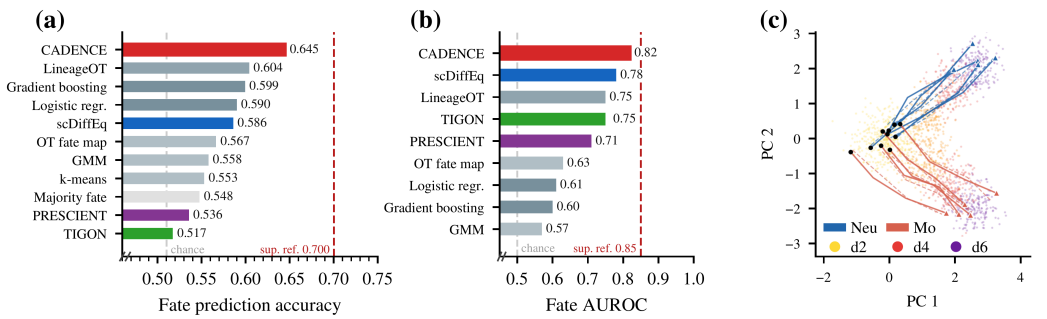

Figure 4: BM7 (LARRY haematopoiesis) at the single-cell unit: all 1,488 day-2 cells of 986 clones, cross-fitted over five by-clone folds, every method re-run by us. **(a)** Fate accuracy, x-axis from 0.46; **(b)** fate AUROC, from 0.45; slashes: truncated axes; dotted: chance; dashed red: supervised reference; black ticks and italics: published values. Accuracy s.e. (clone-clustered) at most 0.015. **(c)** Observed (dashed) and predicted (solid) clone paths in PCA space (see Fig. 12; values: App. H.5).

On the day-2 cells, the protocol the published single-cell results use, cross-fitted over five by-clone folds so that all 1,488 cells of the 986 clones are scored once, it reaches AUROC 0.82 ± 0.01 against 0.60–0.61 for classifiers trained on fate labels from the same day-2 context, which do not see the later-day populations, and 0.71–0.78 for published unsupervised dynamics methods on the same cells, all separated from it (clone-bootstrap $p \leq 0.004$) (App. H.5). The accuracy gap over the classifiers is 0.046 (0.645 vs. 0.599); accuracy uses the router's untuned $\arg\max$ rule, whereas AUROC is threshold-free. Our re-implementations reproduce the published fate results on both protocols their authors use, to within 0.010 where a point value is published (Tab. 14); published two-class accuracies do not exist, so accuracies are compared between re-runs. Per cell CADENCE reaches 0.645 against 0.586 for scDiffEq, the strongest published method without lineage input (exact McNemar $p < 10^{-6}$ at the clone unit, 680 against 617 of 986), and 0.604 for LineageOT, which uses the training clones' barcode couplings (p=0.004); given the lineage of the evaluated clones as well, LineageOT reaches 0.720, above CADENCE. The later-day populations carry the margin: removing the day-4 and day-6 reference marginals drops it to 0.588, the level of those supervised classifiers; this control does not separate the dynamics model from the added data. A supervised reference given the labels *and* the dynamics reaches 0.700/0.85 (App. H.5). BM7 is also evaluated on held-out-day forecasting: one later day is withheld from training entirely and then predicted (App. H.5). On pooled interpolation CADENCE is 23.5% below the published scDiffEq value. A model-free OT interpolant is 17.3% below the published value and 8% above CADENCE; with its step length tuned on the target day it reaches 1.16× the split-half noise level against CADENCE's 1.14×, level with it there. In the other five columns the oracle-tuned model-free bound is 1.20–1.31× CADENCE's divergence.

Ablation Study. App. I isolates each architectural choice. On BM4, pre-compressors improve the forecast metrics, but routing accuracy depends on the PF-ODE: 0.920–0.936 with it across the linear, PCA and VAE families and 0.343–0.348 without it, against chance 0.333 (three seeds, App. I.4).

9 Discussion and Limitations

CADENCE extracts longitudinal constraints from a cross-section by pooling units with similar context. On the synthetic benchmarks CADENCE has lower individual-level error than published aggregate transport on five of six, by 1.6–5.8×, and than sequence models trained on one snapshot per unit (tested on BM1), and the highest per-unit skill on all six; on LARRY it predicts fate without labels, ahead of all but two better-informed comparators.

Limitations (App. K). DFA, COA′ and FOA concern the data-generating process and its interaction with the model class, so they cannot be verified from a trained model; only COA′ failure is characterised, on synthetic data (App. I.5), and a fixed K leaves an MRA residual $\varepsilon > 0$ whenever the parameter image is not contained in the convex hull of K points (Prop. 8). The generative model assumes neural-ODE dynamics and enrolment independent of state given context, which the synthetic benchmarks satisfy by construction and BM7 by its fixed culture-day design; App. J.1 measures what violating it costs. It gives point forecasts rather than the predictive intervals clinical use requires, and its real-data evidence is one dataset with a two-class fate.

AI use statement

We used generative AI tools throughout the manuscript to assist with language editing, including improving grammar, wording, clarity, and readability. The scientific content, methodology, theoretical development, experimental design, analysis, and interpretation of results were developed by the authors. Generative AI tools were not used to generate data or to independently produce scientific conclusions. All AI-assisted text was reviewed, revised where necessary, and verified by the authors for scientific accuracy and consistency. All AI-assisted suggestions concerning mathematical arguments were independently verified by the authors. We take responsibility for the final content of this work, including all text, claims, and results.

Ethics statement

This work develops and evaluates a methodological framework using synthetic systems and publicly available, previously published datasets, including publicly released single-cell transcriptomic data. We conducted no new human- or animal-subject experiments and use no private or personally identifiable data. CADENCE produces individual-level point forecasts from sparse cross-sectional observations, which could be inappropriate for consequential decision making without application-specific validation. In particular, context-dependent models may reproduce biases present in the populations and covariates on which they are trained. We therefore regard the reported results as methodological evidence within the settings and assumptions evaluated in this work, rather than as validation for clinical or other consequential deployment. The scope and limitations of the individual-level forecasts are discussed in Sec. 9 and App. K.

Reproducibility statement

We provide the information required to reproduce the theoretical and empirical results throughout the paper and appendices. The assumptions and complete proofs of the theoretical results are given in App. B. Benchmark definitions, data generation and preprocessing, and evaluation protocols are described in Apps. F–H, including dataset sizes, splits, and baseline tuning. Complete benchmark-specific model, optimisation, and solver configurations are reported in the corresponding configuration tables in App. H, while Apps. I–J document the ablations, controls, and robustness analyses supporting the reported results. Code will be released upon acceptance.

References

Michael Albergo, Nicholas M Boffi, and Eric Vanden-Eijnden. Stochastic interpolants: A unifying framework for flows and diffusions. *Journal of Machine Learning Research*, 26(209):1–80, 2025.

Ilze Amanda Auzina, Çağatay Yıldız, Sara Magliacane, Matthias Bethge, and Efstratios Gavves. Modulated neural odes. *Advances in Neural Information Processing Systems*, 36:44572–44594, 2023.

Carles Balsells-Rodas, Yixin Wang, and Yingzhen Li. On the identifiability of switching dynamical systems. In *Proceedings of the 41st International Conference on Machine Learning*, Proceedings of Machine Learning Research. PMLR, 2024.

Markus W Baumgartner, Anson Lei, Joe Watson, and Ingmar Posner. Disentangling dynamical systems: Causal representation learning meets local sparse attention. *arXiv preprint arXiv:2603.14483*, 2026.

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.

Paidamoyo Chapfuwa, Sherri Rose, Lawrence Carin, Edward Meeds, and Ricardo Henao. Capturing actionable dynamics with structured latent ordinary differential equations. In *Uncertainty in Artificial Intelligence*, pp. 286–295. PMLR, 2022.

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.

Maurice D. Craig. Minimum-volume transforms for remotely sensed data. *IEEE Transactions on Geoscience and Remote Sensing*, 32(3):542–552, 1994. doi: 10.1109/36.297973.

Valentin De Bortoli, James Thornton, Jeremy Heng, and Arnaud Doucet. Diffusion Schrödinger bridge with applications to score-based generative modeling. In *Advances in Neural Information Processing Systems*, volume 34, pp. 17695–17709, 2021.

Edward De Brouwer, Jaak Simm, Adam Arany, and Yves Moreau. GRU-ODE-Bayes: Continuous modeling of sporadically-observed time series. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Payam Dibaeinia and Saurabh Sinha. Sergio: a single-cell expression simulator guided by gene regulatory networks. *Cell systems*, 11(3):252–271, 2020.

Alice Driessen, Dhruva Abhijit Rajwade, Benedek Harsanyi, Marianna Rapsomaniki, and Jannis Born. Conditional monge gap enables generalizable single-cell perturbation modelling. *Nature Machine Intelligence*, pp. 1–13, 2026.

Emilien Dupont, Arnaud Doucet, and Yee Whye Teh. Augmented neural ODEs. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Hubert M. Fonteijn, Martin Fillmer, Jonathan D. Rohrer, and Daniel C. Alexander. An event-based model for disease progression and its application in familial alzheimer's disease and huntington's disease. *NeuroImage*, 60(4):1880–1889, 2012.

Aden Forrow and Geoffrey Schiebinger. LineageOT is a unified framework for lineage tracing and trajectory inference. *Nature Communications*, 12(1):4940, 2021. doi: 10.1038/s41467-021-25133-1.

Marco Fraccaro, Simon Kamronn, Ulrich Paquet, and Ole Winther. A disentangled recognition and nonlinear dynamics model for unsupervised learning. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Marta Garnelo, Dan Rosenbaum, Christopher Maddison, Tiago Ramalho, David Saxton, Murray Shanahan, Yee Whye Teh, Danilo Rezende, and SM Ali Eslami. Conditional neural processes. In *International conference on machine learning*, pp. 1704–1713. PMLR, 2018a.

Marta Garnelo, Jonathan Schwarz, Dan Rosenbaum, Fabio Viola, Danilo J Rezende, SM Eslami, and Yee Whye Teh. Neural processes. *arXiv preprint arXiv:1807.01622*, 2018b.

Adam P. Generale, Andreas E. Robertson, and Surya R. Kalidindi. Modeling stochastic conditional dynamics from sparse observations via kernel-stabilized flow matching. *Transactions on Machine Learning Research*, 2026. arXiv:2411.08314.

David Ha, Andrew Dai, and Quoc V Le. Hypernetworks. *arXiv preprint arXiv:1609.09106*, 2016.

Hermanni Hälvä, Sylvain Le Corff, Luc Lehéricy, Jonathan So, Yongjie Zhu, Elisabeth Gassiat, and Aapo Hyvarinen. Disentangling identifiable features from noisy data with structured nonlinear ica. *Advances in Neural Information Processing Systems*, 34:1624–1633, 2021.

Tatsunori Hashimoto, David Gifford, and Tommi Jaakkola. Learning population-level diffusions with generative rnns. In *International Conference on Machine Learning*, pp. 2417–2426. PMLR, 2016.

Robert Hermann and Arthur J Krener. Nonlinear controllability and observability. *IEEE Transactions on Automatic Control*, 22(5):728–740, 1977.

Çağlar Hızlı, Çağatay Yıldız, Matthias Bethge, ST John, and Pekka Marttinen. Identifying latent state transition in non-linear dynamical systems. *arXiv preprint arXiv:2406.03337*, 2024.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Guillaume Huguet, Daniel Sumner Magruder, Alexander Tong, Oluwaseun Fasina, Manik Kuchroo, Guy Wolf, and Smita Krishnaswamy. Manifold interpolating optimal-transport flows for trajectory inference. In *Advances in Neural Information Processing Systems*, volume 35, pp. 29903–29920, 2022.

Aapo Hyvärinen, Hiroaki Sasaki, and Richard E Turner. Nonlinear ICA using auxiliary variables and generalized contrastive learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 859–868. PMLR, 2019.

Valerii Iakovlev, Cagatay Yildiz, Markus Heinonen, and Harri Lähdesmäki. Latent neural ODEs with sparse Bayesian multiple shooting. In *International Conference on Learning Representations*, 2023.

Alberto Isidori. *Nonlinear control systems: an introduction*. Springer, 1985.

Robert A Jacobs, Michael I Jordan, Steven J Nowlan, and Geoffrey E Hinton. Adaptive mixtures of local experts. *Neural computation*, 3(1):79–87, 1991.

Gavin Kerrigan, Giosuè Migliorini, and Padhraic Smyth. Dynamic conditional optimal transport through simulation-free flows. In *Advances in Neural Information Processing Systems*, 2024.

Ilyes Khemakhem, Diederik Kingma, Ricardo Monti, and Aapo Hyvarinen. Variational autoencoders and nonlinear ica: A unifying framework. In *International conference on artificial intelligence and statistics*, pp. 2207–2217. PMLR, 2020.

Patrick Kidger, James Morrill, James Foster, and Terry Lyons. Neural controlled differential equations for irregular time series. In *Advances in Neural Information Processing Systems*, volume 33, pp. 6696–6707, 2020.

Matthieu Kirchmeyer, Yuan Yin, Jérémie Donà, Nicolas Baskiotis, Alain Rakotomamonjy, and Patrick Gallinari. Generalizing to new physical systems via context-informed dynamics model. In *International conference on machine learning*, pp. 11283–11301. PMLR, 2022.

Dominik Klein, Giovanni Palla, Marius Lange, Michal Klein, Zoe Piran, Manuel Gander, Laetitia Meng-Papaxanthos, Michael Sterr, Lama Saber, Changying Jing, Aimée Bastidas-Ponce, Perla Cota, Marta Tarquis-Medina, Shrey Parikh, Ilan Gold, Heiko Lickert, Mostafa Bakhti, Mor Nitzan, Marco Cuturi, and Fabian J. Theis. Mapping cells through time and space with moscot. *Nature*, 638(8052):1065–1075, 2025. doi: 10.1038/s41586-024-08453-2.

David Klindt, Lukas Schott, Yash Sharma, Ivan Ustyuzhaninov, Wieland Brendel, Matthias Bethge, and Dylan Paiton. Towards nonlinear disentanglement in natural data with temporal sparse coding. *arXiv preprint arXiv:2007.10930*, 2020.

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

Kai Lagemann, Christian Lagemann, and Sach Mukherjee. Invariance-based learning of latent dynamics. In *The Twelfth International Conference on Learning Representations*, 2023.

Nan M Laird and James H Ware. Random-effects models for longitudinal data. *Biometrics*, pp. 963–974, 1982.

Hugo Lavenant, Stephen Zhang, Young-Heon Kim, and Geoffrey Schiebinger. Towards a mathematical theory of trajectory inference. *arXiv preprint arXiv:2102.09204*, 2021.

Chia-Hsiang Lin, Wing-Kin Ma, Wei-Chiang Li, Chong-Yung Chi, and ArulMurugan Ambikapathi. Identifiability of the simplex volume minimization criterion for blind hyperspectral unmixing: The no-pure-pixel case. *IEEE Transactions on Geoscience and Remote Sensing*, 53(10):5530–5546, 2015.

Yaron Lipman, Ricky T Q Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023.

Lennart Ljung. *System Identification: Theory for the User*. Prentice Hall, 2nd edition, 1999.

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *international conference on machine learning*, pp. 4114–4124. PMLR, 2019.

Maximilian Mauel, Johannes R Hübers, David Berghaus, Patrick Seifner, and Ramses J Sanchez. Foundation inference models for ordinary differential equations. *arXiv preprint arXiv:2602.08733*, 2026.

Kirill Neklyudov, Rob Brekelmans, Daniel Severo, and Alireza Makhzani. Action matching: Learning stochastic dynamics from samples. In *International conference on machine learning*, pp. 25858–25889. PMLR, 2023.

Alexander Norcliffe, Cristian Bodnar, Ben Day, Jacob Moss, and Pietro Liò. Neural {ode} processes. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=27acGyyI1BY.

Roussel Desmond Nzoyem, David AW Barton, and Tom Deakin. Neural context flows for meta-learning of dynamical systems. *arXiv preprint arXiv:2405.02154*, 2024.

Neil P Oxtoby, Alexandra L Young, David M Cash, Tammie LS Benzinger, Anne M Fagan, John C Morris, Randall J Bateman, Nick C Fox, Jonathan M Schott, and Daniel C Alexander. Data-driven models of dominantly-inherited alzheimer's disease progression. *Brain*, 141(5):1529–1544, 2018.

Wonjung Park, Suhyun Ahn, Maria Hernandez, Susana Maniega, and Jinah Park. Brainode: Neural shape dynamics for age-and disease-aware brain trajectories. *Advances in Neural Information Processing Systems*, 38:118550–118581, 2026.

Ethan Pérez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. FiLM: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Konpat Preechakul, Nattanat Chatthee, Suttisak Wizadwongsa, and Supasorn Suwajanakorn. Diffusion autoencoders: Toward a meaningful and decodable representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10619–10629, 2022.

Joan Puigcerver, Carlos Riquelme, Basil Mustafa, and Neil Houlsby. From sparse to soft mixtures of experts. In *International Conference on Learning Representations*, 2024.

Yulia Rubanova, Ricky T Q Chen, and David K Duvenaud. Latent ordinary differential equations for irregularly-sampled time series. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

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

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

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *International Conference on Learning Representations*, 2017.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021.

Bharath K. Sriperumbudur, Arthur Gretton, Kenji Fukumizu, Bernhard Schölkopf, and Gert R. G. Lanckriet. Hilbert space embeddings and metrics on probability measures. *Journal of Machine Learning Research*, 11 (50):1517–1561, 2010. URL http://jmlr.org/papers/v11/sriperumbudur10a.html.

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

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. *arXiv preprint arXiv:2302.00482*, 2023.

Cole Trapnell, Davide Cacchiarelli, Jonna Grimsby, Prapti Pokharel, Shuqiang Li, Michael Morse, Niall J Lennon, Kenneth J Livak, Tarjei S Mikkelsen, and John L Rinn. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nature Biotechnology*, 32(4):381–386, 2014.

Arash Vahdat, Karsten Kreis, and Jan Kautz. Score-based generative modeling in latent space. In *Advances in Neural Information Processing Systems*, volume 34, pp. 11287–11302, 2021.

A. W. van der Vaart. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1998. ISBN 9780521784504. doi: 10.1017/CBO9780511802256.

Ayesha Vermani, Josue Nassar, Hyungju Jeon, Matthew Dowling, and Il Memming Park. Meta-dynamical state space models for integrative neural data analysis. *arXiv preprint arXiv:2410.05454*, 2024.

Michael E. Vinyard, Anders W. Rasmussen, Ruitong Li, Allon M. Klein, Gad Getz, and Luca Pinello. Learning cell dynamics with neural differential equations. *Nature Machine Intelligence*, 7(12):1969–1984, 2025. doi: 10.1038/s42256-025-01150-3. URL https://doi.org/10.1038/s42256-025-01150-3.

Rui Wang, Robin Walters, and Rose Yu. Meta-learning dynamics forecasting using task inference. *Advances in Neural Information Processing Systems*, 35:21640–21653, 2022.

Xiaoda Wang, Yuji Zhao, Kaiqiao Han, Xiao Luo, Sanne van Rooij, Jennifer Stevens, Lifang He, Liang Zhan, Yizhou Sun, Wei Wang, et al. Conditional neural ode for longitudinal parkinson's disease progression forecasting. *arXiv preprint arXiv:2511.04789*, 2025.

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

Dingling Yao, Caroline Muller, and Francesco Locatello. Marrying causal representation learning with dynamical systems for science. *Advances in Neural Information Processing Systems*, 37:71705–71736, 2024.

Weiran Yao, Yuewen Sun, Alex Ho, Changyin Sun, and Kun Zhang. Learning temporally causal latent processes from general temporal data. *arXiv preprint arXiv:2110.05428*, 2021.

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.

Cagatay Yildiz, Markus Heinonen, and Harri Lahdesmaki. Ode2vae: Deep generative second order odes with bayesian neural networks. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Yuan Yin, Ibrahim Ayed, Emmanuel de Bézenac, Nicolas Baskiotis, and Patrick Gallinari. Leads: Learning dynamical systems that generalize across environments. *Advances in Neural Information Processing Systems*, 34:7561–7573, 2021.

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

A Claims and Supporting Evidence

Tab. 3 maps each claim of the abstract and introduction to its evidence and the limit of that evidence; Tab. 4 pairs each part of the analysis with the ablation that tests it.

Table 3: Claims, supporting evidence, and the scope of that evidence.

Claim	Evidence	Scope
Individual continuous-time point forecasts from one snapshot plus static context	Tab. 2; skill 0.47–0.97 against the leaf-conditional mean, Wilcoxon $p < 10^{-4}$, within-leaf Spearman 0.53–0.95 (App. I.7); App. I.8 (App. I.9 is a graded-coordinate check with μ given as context)	Point forecasts, no predictive intervals; skill is relative to the leaf-mean null. BM3's 0.96 uses the weaker cohort null; BM4's 0.95–0.97 is under PCA-10 (0.72–0.83 uncompressed)
Under explicit assumptions leaf parameters are identifiable	Thms. 1, 2; App. B	Population level, fixed encoder frame, under DFA, COA′, ENR, BAS, POS, PRW, FOA (+SEP, +FOA_j); not about a trained network (Tab. 37: twelve differences); not claimed on BM7 (POS fails)

Tab. 3, continued

Claim	Evidence	Scope
Every measurable difference between analysed class and deployed system is measured; none breaks the predicted behaviour	Sec. 5, Tab. 37; Apps. J.5, J.6, J.7, J.8, I.4, J.1	Nine of twelve measured: none costs over 19% MAE or 0.04 routing, eight under 10% and 0.02; the two theory bridges (learned θ, pooled matching) at most 4% and 0.006 on BM1/BM5; hard sample-split idealisation 14% worse. Three scope exclusions (BM4 pre-compressor, BM3 continuous router, BM7 single enrolment day). Empirical bounding, not an extension of the theorems
A separation statistic related to FOA's sufficient condition is evaluable between trained heads	Prop. 4, Cor. 1; Tab. 5	$\geq$ 1.4 orders above a resampling floor on every discrete-leaf benchmark, leaf-wise and unaugmented (2.4 above the probe floor); measures head distinctness (bears on SEP), does not verify FOA
Lowest cross-sectional error among published methods on the same data: MAE on five of six, mean SW_2 on six of six (BM1, BM3 margins within seed variation)	Tab. 2: BM2 readout and BM4 PCA-10 given to every method; baselines tuned on validation MAE, so SW_2 is against baselines tuned for the other metric; re-selecting by SW_2 leaves the ordering (App. H.2); snapshot-conditioned CVFM-S in App. I.10	BM2: OT-CFM lower MAE (0.021 vs. 0.025, uncorrected p=0.013), CADENCE lower SW_2 (uncorrected p=0.017), neither survives Holm, SW_2 lead largely the readout used as target (App. F.3); our CVFM-S and c-OT-CFM lower on some BM2 cells; under point-optimal scoring CVFM also passes CADENCE on BM2. Synthetic only
Lower individual error than tuned context-blind latent sequential models given dense full trajectories	Oracle block: lower MAE than NDP, CoDA, Latent ODE, ODE^{2}VAE, LaDID on all six; BM1 best (NDP) 5.84 raw, 3.12 on Stage-1 latents, vs. 2.05 (App. I.1)	Dense-trajectory *transport* still leads (OT-CFM on both metrics on BM2, BM4, BM5, on SW_2 on BM1, BM3; full-trajectory conditional transport wins ten of the twelve benchmark-by-metric comparisons, App. I.10); most of the BM1 margin is the representation (unrouted 3.25 on the context-indexed latents); with the true leaf a Latent ODE is lower on BM2 (0.019)
Subgroup structure recovered without supervision	Routing 0.937, 0.947, 0.923, 0.843, 0.850 (BM1, 2, 4, 5, 6) vs. chance 0.333/0.50 (Fig. 10); a held-out alignment selects the same permutation everywhere (App. J.7)	Labels only resolve the expert permutation at evaluation; routing is a K-class metric, not individual-level
On lineage-traced data, unsupervised routing discriminates the fates better than label-supervised classifiers on the same day-2 context, with unlabelled later-day populations as extra input	Tab. 16, per cell: AUROC 0.82 vs. 0.60–0.61 (clone bootstrap $p < 0.001$), accuracy 0.645 vs. 0.599 (margin over gradient boosting 50 clones of 986, exact McNemar $p < 10^{-4}$); scDiffEq re-run 0.586/0.78; Tabs. 14, 18	One dataset, two-class fate, all 1,488 day-2 cells of 986 clones cross-fitted over five by-clone folds; every comparator re-run; clone-clustered s.e. $\pm$0.015 (effective n = clones); no fate annotation selects training data; 92% of a supervised reference of 0.700 (App. J.4); first in all five folds (Tab. 17); only the held-out-day scDiffEq and Pearson comparisons use published values
The BM7 margin needs the persistent dynamics as well as the later-day populations	Tab. 18: static one-step map in place of the ODE 0.612 (AUROC 0.76), 28 clones below CADENCE (McNemar $p = 0.005$); disjoint gene sets for context and state 0.632 (AUROC 0.80)	One dataset; the no-ODE arm isolates the ODE, not every property of the dynamics model
Forecasting a withheld day	Tab. 20: 1.20–1.31$\times$ the oracle-tuned model-free bound in five of six columns; 23.5% better than published scDiffEq on pooled interpolation	No separation on pooled interpolation (bound 1.16$\times$ a measured floor); bounding rows are model-free
Routing degrades along a predicted null when context loses information	App. I.5: null $(1-\eta)a_0 + \eta/K$ tracked to 0.04 (a_0=0.931) and 0.02 subset-scored (a_0=0.975); at η=1 MAE $\times$3.3, SW_2 $\times$3.5, worse than the unrouted field (6.84 vs. 3.25)	Needs ground-truth regimes, so synthetic data only
The cross-sectional objective is non-degenerate on these data	Between-bin $\mathrm{MMD}^2 \geq 119\times$ the within-bin split-half floor (App. B.9, Tab. 47)	Sixteen of seventeen leaves measurable (the smaller BM7 fate excluded); verified on encoded latents, not a guarantee for arbitrary data
Follow-up observations improve accuracy	MAE 28–36% better on BM1–BM5, 5% on BM6, with three follow-ups (App. I.6)	BM1–BM6 only
Routing is not an artefact of context entering the encoder	Unconditional score keeps routing at 0.90 (BM4), 0.92 (BM1) (App. I.4)	Two benchmarks, three seeds
Accuracy does not come from context smoothing	MAE 3.1–12.7$\times$ below a context smoother and 1.3–3.5$\times$ below a time-aligned $(\mathbf{c}, \mathbf{z})$ nearest neighbour reading training futures (Tab. 32)	Synthetic benchmarks

Table 4: How the identifiability analysis connects to the design and to the evidence.

Analysis	Design choice it motivates	Test or diagnostic of that choice
Gaussian pinning reduces the spatial ambiguity to G (Prop. 2)	Score-based PF-ODE encoder	Routing 0.920–0.936 with the PF-ODE, 0.343–0.348 without (chance 0.33); unconditional score keeps 0.90 (BM4), 0.92 (BM1), so the dependence is on the Gaussianising transport, not on context entering the encoder (App. I.4); the effect exceeds what Prop. 2 predicts (App. B.2)
Context identifies the regime (COA$'$)	Router sees static context only	Intact units 0.97–0.98, corrupted 0.35–0.36; aggregate on the null to 0.02 (subset-scored) and 0.04 (full sweep) (Tab. 29)
Regular parameter image (MRA, Props. 5, 8)	Convex basis of K experts	Archetype recovery minimised at $K{=}\lvert\Lambda\rvert{=}3$ on BM1 (0.081; 0.291 at $K{=}2$, 0.131 at $K{=}5$) while MAE shares the minimiser but varies only $1.29\times$ over $K \in \{2, \dots, 5\}$: sharpness, not location, distinguishes the axes (Tab. 48); $K{=}1$ raises MAE $2.032 \to 3.201$; context-conditioned dynamics lower MAE 23–37% over an unconditioned field on the same latents
FOA and its sufficient condition (Prop. 4)	Per-subgroup cross-sectional MMD objective	Diagnostic: separation statistic ≥ 1.4 orders above a resampling floor, leaf-wise and unaugmented (Tab. 5); between-bin $\mathrm{MMD}^2 \geq 119\times$ the split-half floor on all sixteen measurable leaves (Tab. 47)
DFA: units in a leaf share a dynamical parameter (and, theory-matched, a vector field)	Context reaches the dynamics only through the router	Theory-matched $\mathbf{c}$-free h_{base} costs 8.3–9.3% MAE and ≤ 0.009 routing on BM1, BM3, BM4, BM6 (Tab. 43)

B Proofs of Theoretical Results

The results in this appendix characterise the model class under the stated assumptions; Sec. 5 explains their role in motivating and interpreting the method.

B.1 Proposition 1: Proof and a Conjugation Observation

Proof of Proposition 1. Let $q \geq 2$, $A \neq 0$ skew-symmetric and $h_w(z) = wAz$. Then $\Phi^w_{t,t_0} = \exp(w(t{-}t_0)A) \in SO(q)$, so the pushforward of $\rho_0 = \mathcal{N}(\mathbf{0}, \mathbf{I})$ is $\mathcal{N}(\mathbf{0}, \mathbf{I})$ for every w and every t: the whole family generates identical marginals. Yet since $A \neq 0$, $\ker A$ is a proper subspace and hence ρ_0-null, and for every $z_0 \notin \ker A$ the initial velocities wAz_0 differ for distinct w, so the individual trajectories differ across the family. □

Conjugation observation (frame-dependence). Let $\phi : \mathbb{R}^q \to \mathbb{R}^q$ be a smooth diffeomorphism. Define $\tilde{\mathbf{z}}^i_t = \phi(\mathbf{z}^i_t)$. By the chain rule, $\frac{d\tilde{\mathbf{z}}^i_t}{dt} = D\phi(\phi^{-1}(\tilde{\mathbf{z}}^i_t))h_\theta(\phi^{-1}(\tilde{\mathbf{z}}^i_t), t, \mathbf{c}_i; \mathbf{w}_i) = \tilde{h}(\tilde{\mathbf{z}}^i_t, t, \mathbf{c}_i; \mathbf{w}_i)$, so $\tilde{\mathbf{z}}^i_t$ satisfies the ODE with $\tilde{h}$ and initial condition $\phi(\mathbf{z}^i_0)$. Reconstruction is unchanged: $\mathrm{Dec}_\psi(\phi^{-1}(\phi(\mathbf{z}^i_{t_i}))) = \mathbf{x}^i_{t_i}$. Since $\mathrm{Diff}(\mathbb{R}^q)$ is infinite-dimensional, valid configurations are uncountably infinite.

B.2 Residual Spatial Symmetry: Statements and Proofs

This subsection states the four facts behind Proposition 2 precisely. Throughout, $\nu := E_{\#}\, p_{\mathrm{data}}$ with E the (lossless) compressor, which is the identity in every headline run but BM4's, which precedes it by a lossy PCA-10 pre-compressor (App. I.2), and $\gamma_q = \mathcal{N}(\mathbf{0}, \mathbf{I})$. *Standing hypothesis: ν has a smooth positive density.*

G1 (Residual symmetry at fixed E). If $P, \tilde{P} \in \mathrm{Diff}(\mathbb{R}^q)$ satisfy $P_{\#}\nu = \tilde{P}_{\#}\nu = \gamma_q$, then $\varphi := \tilde{P} \circ P^{-1} \in G := \{\varphi \in \mathrm{Diff}(\mathbb{R}^q) : \varphi_{\#}\gamma_q = \gamma_q\}$, a *proper* subgroup of $\mathrm{Diff}(\mathbb{R}^q)$.

Proof. $\varphi_{\#}\gamma_q = \varphi_{\#}(P_{\#}\nu) = (\varphi \circ P)_{\#}\nu = \tilde{P}_{\#}\nu = \gamma_q$. Properness: a translation $\tau_b(x) = x + b$ with $b \neq 0$ has $(\tau_b)_{\#}\gamma_q = \mathcal{N}(b, I) \neq \gamma_q$, so $\tau_b \in \mathrm{Diff}(\mathbb{R}^q) \setminus G$. □

Note. G is the residual *at fixed* E, and fixing E is a design choice.

G2 (Structure of G). (i) $O(q) \subseteq G$. (ii) $q = 1$: $G = \{\pm\mathrm{id}\}$. (iii) $q \geq 2$: $G \supsetneq O(q)$; the twists $T_f(r, \vartheta) = (r, \vartheta + f(r))$, $f \in C^\infty_c((0, \infty))$, on a fixed 2-plane lie in G and are non-isometric; $f \mapsto T_f$ is an injective homomorphism, so G admits injective continuous homomorphisms $\mathbb{R}^n \to G$ for every n; hence G *carries no finite-dimensional Lie-group structure compatible with the C^1 topology, and is non-compact in it.* (The topology qualifier is essential: as abstract groups $\mathbb{R} \cong \mathbb{R}^n$.) Non-compactness in the C^1 topology follows independently of the image being closed: the twists $\{T_{sf}\}_{s \in \mathbb{R}}$ form an unbounded one-parameter family in the C^1 metric, so G contains a C^1-unbounded subset.

Proof sketch. The C^1-compatibility qualifier is what licenses the smoothness theorem used next: once a finite-dimensional Lie structure on G is assumed compatible with the C^1 subspace topology, a homomorphism continuous into that topology is a continuous homomorphism of Lie groups, hence smooth. A continuous homomorphism between second-countable Lie groups is smooth, and an injective one has injective differential

at the identity, since a one-parameter subgroup with vanishing generator is trivial; an injective continuous homomorphism $\mathbb{R}^n \to G$ therefore forces $\dim G \geq n$, which yields the last claim. (i) Orthogonal maps preserve $\mathcal{N}(\mathbf{0}, \mathbf{I})$. (ii) In one dimension a smooth bijection is monotone; preserving the centered Gaussian law forces $|\varphi'| \equiv$ the density ratio, whose only monotone solutions are $\pm\mathrm{id}$. (iii) A twist on the (r, ϑ) plane pushes γ_2 (whose density depends on r only) to itself, since the angular coordinate is uniform and independent of r; it is non-isometric whenever $f \not\equiv 0$. Injectivity of $f \mapsto T_f$ is immediate, giving continuum many independent one-parameter subgroups. □

G3 (Exactness and the approximate regime). Fix the forward SDE, the noise schedule, the terminal horizon, the time-dependent score, and regularity ensuring uniqueness of the ODE flow (e.g. locally Lipschitz drift). Under the *exact terminal-law idealisation* ($P_{\#}\nu = \gamma_q$ exactly, an infinite-time or exact-Gaussian-terminal reading, since a finite-horizon VP diffusion is only approximately Gaussian at its terminal time), the PF-ODE of the exact score is unique (Song et al., 2021), so with a common reference $P_{\#}\nu = \tilde{P}_{\#}\nu = \gamma_q$ one has $\varphi = \mathrm{id}$. Under *approximate* score matching / finite horizon the terminal laws are $P_{\#}\nu = \eta$, $\tilde{P}_{\#}\nu = \tilde{\eta}$ with $\eta, \tilde{\eta} \approx \gamma_q$; then $\varphi_{\#}\eta = \tilde{\eta}$ and only an *approximate* symmetry holds, where $D(\varphi_{\#}\gamma_q, \gamma_q)$, for D a metric compatible with pushforwards (e.g. W_2) and φ Lipschitz, is bounded by the two terminal-law errors and the Lipschitz distortion of φ. The input side is approximate as well: the encoder treats the data as the diffusion's $\tau = 0.05$ marginal and integrates to $\tau = 0.999$ (Alg. 1), so even with the exact score $P_{\#}\nu$ equals γ_q only approximately at both ends.

G4 (Reparameterisation identity). For *any* bijection φ of $\mathbb{R}^q$, the conjugated model $(\varphi \circ \mathrm{Enc},\ \varphi \circ \Phi_t^w \circ \varphi^{-1},\ \mathrm{Dec} \circ \varphi^{-1})$ yields identical predictions:

$$\mathrm{Dec} \circ \varphi^{-1}\big(\varphi \circ \Phi_t^w \circ \varphi^{-1}(\varphi \circ \mathrm{Enc}(x))\big) = \mathrm{Dec}(\Phi_t^w(\mathrm{Enc}(x))) = \hat{x}.$$

Scope. G4 is a *prediction identity*: it needs only bijectivity *of* φ, since the conjugation cancels whatever $\mathrm{Dec} \circ \mathrm{Enc}$ does, and so holds even when E is lossy. Theorems 1–2, by contrast, require E lossless ($q = p$), as their statements record; the $q < p$ variants of App. I.2 are outside their scope. Admissibility as a *model-class* reparameterisation is separate: staying a Gaussianising PF-ODE forces $\varphi \in G$, and the exact-terminal-law idealisation (G3) forces $\varphi = \mathrm{id}$. Theorem 2 uses G4 for soundness; this is why the residual symmetry is observationally harmless for prediction.

G4 and the encoder ablation. G4 states that the residual symmetry is prediction-neutral, and App. I.4 shows that removing the PF-ODE collapses routing to chance. The two statements concern different objects: G4 quantifies over models already fitted to the data, whereas the training objective is not conjugation-invariant. Its Gaussian MMD with a median-heuristic bandwidth (Tab. 9) is invariant to similarity transforms of the latent space, so scale is not what Gaussian pinning contributes: the non-isometric elements of G, the twists of G2, deform pairwise distances unevenly and change the objective's landscape. Consistent with this, App. I.4's VAE arm, whose latents are already approximately standardised, still routes at chance, so the effect is not scale normalisation. The PF-ODE therefore acts on the optimisation problem rather than on the set of representable predictors; the five-encoder test of App. I.4 locates the property the router depends on in the *joint* Gaussianisation: a RealNVP flow to $\mathcal{N}(\mathbf{0}, \mathbf{I})$ nearly matches the PF-ODE, a marginal Gaussian copula recovers part of the gap, and whitening does not. The statements G1–G4 are made for an unconditional Gaussianising encoder P. The implemented score network is context-conditional, so in practice P is a family $P_{\mathbf{c}}$ and the residual symmetry a $\mathbf{c}$-indexed family of groups; we do not extend G1–G4 to that family. Empirically the conditioning is inessential for routing on the tested benchmarks (BM4, BM1): an unconditional-score encoder keeps routing accuracy at 0.90 on BM4 and 0.92 on BM1 (App. I.4), so the effect is attributable to the Gaussianising transport rather than to context entering the encoder.

B.3 Formal Regularity Conditions (Assumptions 4 & 5)

Assumption 4 (Parametric Restriction, PRW). *h_θ belongs to a parametric family $\mathcal{H} = \{h_\theta : \theta \in \Theta\}$, $\Theta \subset \mathbb{R}^m$ compact. θ indexes the shared function class and is* fixed across leaves*; $w \in \Theta_w \subset \mathbb{R}^d$ is the* per-leaf *dynamics parameter.* ***(a)*** *This parametric restriction and role split is what Theorem 1 uses.* ***(b)*** *The map $(\theta, w) \mapsto \Phi^{\theta,w}$ is injective on $\Theta \times \Theta_w$* modulo the parameter symmetries of the backbone *(hidden-unit permutations of h_{base} leave every flow unchanged and are quotiented out): distinct parameter classes yield distinct flows. Theorem 1 consumes only part (a); part (b) is used only through FOA$_{\mathrm{i}}$ for Theorem 2's vector-field clause.*

The identified leaf parameter in the deployed FiLM. The deployed FiLM is linear, $s(w) = \mathbf{1} + A_s w$ and $b(w) = A_b w$ (App. H.1), so the field depends on w only through $u = Mw$ with $M = [A_s; A_b] \in \mathbb{R}^{2(q+n_{\mathrm{aug}}) \times d_w}$. If d_w exceeded $2(q + n_{\mathrm{aug}})$, every $w + v$ with $v \in \ker M$ would give the same field, FOA would fail at every w, and the identified object would be u, with Lemma 2 and Prop. 8 read for the matrix $M\mathbf{W}$. The reported runs therefore use $d_w = 14 \leq 2(q + n_{\mathrm{aug}})$ on every benchmark (Tab. 8), and the trained M has full column rank 14 on every seed (smallest singular value at least 0.08 of the largest), so $\ker M = \{0\}$ and w itself is the identified parameter. Because $u = M\mathbf{W}\alpha$ ranges over the same set for any $d_w \geq 2(q + n_{\mathrm{aug}})$ at full rank, this

choice does not change the model class: a $d_w = 32$ configuration gives MAE within 1.5% on BM1 and BM5 (App. H.1).

Assumption 5 (Flow Observability, FOA)**.** *For each leaf λ, the map $w \mapsto \{(\Phi^{w,\mathbf{c}}_{t,t_0})_{\#}\rho^{\lambda}_{0,\mathbf{c}} : t \in \mathcal{T}_\lambda,\ \mathbf{c} \in L_\lambda\}$ is injective within $\mathcal{H}$ at the fixed θ of PRW, where $\Phi^{w,\mathbf{c}}$ is the flow of $h_\theta(\cdot,\cdot,\mathbf{c};w)$ and $\rho^{\lambda}_{0,\mathbf{c}}$ the context-conditional baseline: distinct leaf parameters produce observationally distinct families of time- and context-indexed pushforward marginals. When h_θ does not depend on $\mathbf{c}$ within a leaf the two readings coincide and the $\mathbf{c}$ index may be dropped, as it is in the notation below for legibility. The hypothesis is quantified per context: for every $\mathbf{c} \in L_\lambda$ the map $w \mapsto \{(\Phi^{w,\mathbf{c}}_{t,t_0})_{\#}\rho^{\lambda}_{0,\mathbf{c}} : t \in \mathcal{T}_\lambda\}$ is injective; the joint map over $\mathbf{c}$ is then injective a fortiori. The per-context form is what Step 3 of the proof of Theorem 1 consumes to recover the partition (Step 4 needs only the joint form), and it does not by itself imply the corresponding statement for leaf marginals (see the non-implication remark after Eq. 6).*

Assumption 6 (Leaf Separation, SEP)**.** *Distinct leaves carry distinct dynamics parameters: $\lambda \neq \lambda' \Rightarrow w(\lambda) \neq w(\lambda')$.*

How SEP is used. Under the per-context form of FOA each context's identified baseline and family determine $w(\lambda(\mathbf{c}))$, so SEP makes the leaf partition recoverable from the data law (Step 3 of the proof of Theorem 1). SEP is a condition on the model: two leaves with equal parameters whose context-conditional families also coincide are observationally indistinguishable, since regrouping them leaves the joint law $p(\mathbf{x}, t, \mathbf{c})$ unchanged, and MRA's compactness does not exclude this. Leaves that differ only in their baselines violate SEP and are merged by the recovered partition, which is harmless for forecasting, since merged leaves share one field. Where SEP fails the affected leaves are merged. Two of BM2's four simulated archetypes have parameters too close to separate at this sample size, a weak-FOA rather than a strict SEP failure, and are merged in the same way (App. F.3), which is why BM2 is reported with $K = 3$ rather than $K = 4$.

Inferring the reference stratification. Eq. 3 weights each reference unit by its routing $\alpha_{j,k}$, which needs that unit's context. Where reference observations carry none (BM7's later culture days, App. H.5), the stratification is inferred instead. At each reference time t let ν_t be the pooled observed latent law and replace the stratified reference by a coupling $\gamma = (\gamma_1, \dots, \gamma_K)$ of nonnegative measures with $\sum_k \gamma_k = \nu_t$ and *free* masses, each bounded below by $m_{\min}\,\nu_t(\mathbb{R}^q)$:

$$\mathcal{L}^{\gamma}_{\mathrm{cross}}(w) = \mathbb{E}_t \min_{\gamma:\ \sum_k \gamma_k = \nu_t,\ \gamma_k(\mathbb{R}^q) \geq m_{\min}\nu_t(\mathbb{R}^q)} \sum_{k=1}^{K} \mathrm{MMD}^2\big(\hat{\mu}^{(k)}_t,\ \gamma_k/\gamma_k(\mathbb{R}^q)\big). \tag{4}$$

Masses are left free above the floor because the leaves' shares of the population need not be those at enrolment: on BM7 the two fates proliferate at different rates. The floor keeps every normalisation $\gamma_k/\gamma_k(\mathbb{R}^q)$ defined; the reported runs use $m_{\min} = 0.05$, below both fates' later-day shares. *Implementation.* At each query time γ is an entropic optimal-transport plan (Sinkhorn, $\varepsilon = 0.05$, 50 iterations, squared Euclidean cost in the latent space) from the K predicted expert populations, with the free masses as source marginals, to the reference cells; the masses are updated by a projected gradient step on the simplex truncated at $m_{\min}$, and the coupling is re-estimated every step without gradient through the plan. *Mass guard.* With $m_{k,t} = \gamma_k(\mathbb{R}^q)/\nu_t(\mathbb{R}^q)$, the guard is the entropy deficit $\mathcal{L}_{\mathrm{mass}} = \mathbb{E}_t\big[\log K - H(m_t)\big] \geq 0$, zero at equal masses; it penalises concentration of the reference on few experts and does not force equal masses, since the MMD terms pull the masses towards the leaves' true shares.

Proposition 3 (Latent stratification)**.** *Assume the conditions of Theorem 1 with $K = |\Lambda|$, and in addition (**SEP**$_{\mathrm{s}}$): every per-leaf baseline ρ^{λ}_0 has connected support, and for every reference time t the per-leaf laws ρ^{λ}_t have pairwise disjoint supports. Let $m_{\min}$ lie below every leaf's share of ν_t, so that the stratified coupling is feasible. Assume finally that each expert's source is leaf-pure, i.e. its routing-weighted baseline is one leaf's ρ^{λ}_0, as holds when routing is exact or t_0 is itself a reference time. Then $\mathcal{L}^{\gamma}_{\mathrm{cross}}(w) = 0$ if and only if the stratified objective $\mathcal{L}_{\mathrm{cross}}(w) = 0$ up to a permutation of expert labels.*

Proof. If the stratified objective vanishes, $\gamma_k = \pi_k \rho^{\lambda_k}_t$ attains zero. Conversely, let $\mathcal{L}^{\gamma}_{\mathrm{cross}}(w) = 0$. Each $\hat{\mu}^{(k)}_t$ is the push-forward of a connected-support baseline under a flow, a homeomorphism, so its support is connected; it equals the normalised γ_k, whose support lies in $\operatorname{supp}\nu_t = \bigsqcup_\lambda S_\lambda$, a disjoint union of closed sets, and therefore lies in exactly one $S_{\sigma(k)}$. Since $\sum_k \gamma_k = \nu_t$ charges every S_λ, each receives at least one expert, and with $K = |\Lambda|$ exactly one, so σ is a bijection and $\gamma_k = \nu_t|_{S_{\sigma(k)}}$, whence $\hat{\mu}^{(k)}_t = \rho^{\sigma(k)}_t$ for every t. This is the vanishing of the stratified objective under the relabelling σ, and FOA then gives $w_k = w^*(\sigma(k))$. □

SEP$_{\mathrm{s}}$ is strong and holds on BM7 only approximately: the two fates are well separated by day 6 and overlap at day 4. Hence, the proposition motivates the objective on BM7 rather than certifying it. We note that POS, which the proposition inherits from Theorem 1, also fails there, and t_0 is not a reference time, leaving a heuristic version of the theory. The free masses make a mass-collapse guard necessary in practice, since the floor alone lets one expert claim almost all of ν_t. The guard is $\mathcal{L}_{\mathrm{mass}}$ above, the entropy deficit of the coupling's expert masses;

on BM7 it shares the weight $\lambda_{\text{router}} = 2.0$ of Tab. 8 with the router's per-unit sharpening term (App. C.2), which runs with $\beta = 0$ and does not itself balance usage.

Joint form (FOA$_{\text{j}}$). Theorem 2's vector-field clause additionally requires FOA in its *joint* form, taken over all leaves at once because θ is shared: the map $(\theta, \{w(\lambda)\}_{\lambda\in\Lambda}) \mapsto \{(\Phi^{\theta,w(\lambda),\mathbf{c}}_{t,t_0})_{\#}\rho^{\lambda}_{0,\mathbf{c}}\}_{\lambda\in\Lambda}$ is injective on $\Theta \times \Theta_w^{\Lambda}$. A single leaf may exercise only part of θ, so a per-leaf version would be too strong. The fixed-θ form (Assumption 5) does not imply it, and PRW, which is injectivity of $(\theta, w) \mapsto \Phi^{\theta,w}$ at the level of *flows*, does not either, since matching marginals does not match flows. FOA$_{\text{j}}$ is what upgrades C3's marginal equality to C4's field equality.

Roles disambiguated. PRW separates two parameter roles, the shared backbone θ and the per-leaf w. Theorem 1 identifies $w^*(\lambda)$ *at a fixed, known* θ. If θ is instead to be identified from data, assume the joint injectivity above and read the conclusion as identifying $(\theta, \{w^*(\lambda)\}_\lambda)$; App. J.1 measures the difference between learning θ and fixing it.

PRW and FOA are assumptions rather than consequences of the model class. The following proposition gives explicit sufficient conditions for FOA, and makes the failure regimes of Section 5 precise. Matching pushforward measures does not imply that the generating flows agree pointwise: it implies agreement only up to a measure-preserving rearrangement of ρ^{λ}_0, and G1–G2 above are precisely the statement that this rearrangement group is large. The correct sufficient condition is therefore a condition on the *transport* of the baseline measure, obtained from the Liouville equation.

Proposition 4 (Sufficient condition for FOA)**.** *Fix a leaf λ with t_0 a right-adherent point of $\mathcal{T}_\lambda$. Assume that for every $\mathbf{c} \in L_\lambda$, $\rho^{\lambda}_{0,\mathbf{c}}$ has a C^1 density on an open set $V_{\mathbf{c}} \supseteq \operatorname{supp}\rho^{\lambda}_{0,\mathbf{c}}$ and that each $h(\cdot; w)$ is C^1 and locally Lipschitz. Suppose that for all $w \neq w' \in \Theta_w$, writing $\Delta h_t := h(\cdot, t, \mathbf{c}; w) - h(\cdot, t, \mathbf{c}; w')$ for $t \in \mathcal{T}_\lambda$ and $\mathbf{c} \in L_\lambda$,*

$$\nabla \cdot \left(\rho^{\lambda}_{0,\mathbf{c}}\, \Delta h_{t_0}\right) \not\equiv 0 \quad \text{on } \operatorname{supp}\rho^{\lambda}_{0,\mathbf{c}} \quad \text{for every } \mathbf{c} \in L_\lambda.$$

Then FOA holds in the per-context form within $\mathcal{H}$ at the fixed θ, already from observations in an arbitrarily short window to the right of t_0.

Proof. Suppose $w \neq w'$ generate the same family in some context slice $\mathbf{c} \in L_\lambda$, so $\rho^{w}_{t,\mathbf{c}} = \rho^{w'}_{t,\mathbf{c}}$ for all $t \in T_\lambda$ with common initial value $\rho^{\lambda}_{0,\mathbf{c}}$; the argument below runs inside that slice. Each satisfies the Liouville equation $\partial_t \rho_t + \nabla \cdot (\rho_t h) = 0$. Subtracting and evaluating at $t \downarrow t_0$, the time-derivative terms cancel directly: along any sequence $t_n \downarrow t_0$ in T_λ, $\langle \rho^{w}_{t_n,\mathbf{c}} - \rho^{w'}_{t_n,\mathbf{c}}, \psi\rangle = 0$ for every $\psi \in C_c^\infty$, and likewise at t_0 (common initial value), so the difference quotients along $\{t_n\}$ vanish; since $t \mapsto \rho^{w}_{t,\mathbf{c}} - \rho^{w'}_{t,\mathbf{c}}$ is weakly C^1 (Liouville for h with $\rho^{\lambda}_{0,\mathbf{c}}$ of class C^1; BAS licenses the one-sided evaluation), the quotients converge to $\langle \partial_t(\rho^w - \rho^{w'})|_{t_0}, \psi\rangle$, which is therefore 0 for all ψ, leaving $\nabla \cdot (\rho^{\lambda}_{0,\mathbf{c}}\, \Delta h_{t_0}) \equiv 0$ in the one-sided limit $t \downarrow t_0$ on $\operatorname{supp}\rho^{\lambda}_{0,\mathbf{c}}$, contradicting the hypothesis. □

Remark 1 (Scope of the density premise)**.** *The premise is a condition on the baseline of the physical state coordinates. The deployed dynamics model is an augmented ODE carrying n_{aug} scratch coordinates on BM1–BM6 (Tab. 8), initialised at zero, so its full baseline $\rho^{\lambda}_0 \otimes \delta_{\mathbf{0}}$ has no density and Proposition 4 does not cover it. The proposition covers the unaugmented field; the augmented model is outside the analysed class (Tab. 37), and the unaugmented variant, within 5% of its MAE, is the model to which the separation statistic's strictest evaluation refers (App. J.8); Tab. 5's caption records how this bears on the evaluated statistic.*

Corollary 1 (Local rank form)**.** *FOA holds locally at w if for every $\mathbf{c} \in L_\lambda$ the composite map $v \mapsto \nabla \cdot \left(\rho^{\lambda}_{0,\mathbf{c}} \sum_{j=1}^{d} v_j\, \partial_{w_j} h(\cdot, t_0, \mathbf{c}; w)\right)$ from $\mathbb{R}^d$ to $L^2(\rho^{\lambda}_{0,\mathbf{c}})$ is injective, and h is C^1 in w with $\partial_w h$ of class C^1 in $\mathbf{z}$, locally uniformly. Both factors are needed: injectivity of $v \mapsto \nabla \cdot (\rho^{\lambda}_{0,\mathbf{c}} v)$ on the span of the sensitivity fields alone does not suffice, because the parameter-to-field map may already collapse. For $h(\mathbf{z}; w) = -(w_1 + w_2)\mathbf{z}$ both sensitivity fields equal $-\mathbf{z}$, the divergence map is injective on their one-dimensional span since $\nabla \cdot (\rho_0 \mathbf{z}) \not\equiv 0$, yet w and $w + s(1, -1)$ give the same field, so FOA fails at every scale. This is the correct analogue of a persistent-excitation condition (Ljung, 1999): no parameter direction may generate a flow that transports ρ^{λ}_0 into itself.*

Sketch. *For $w' = w + \varepsilon v$ with $v \in \mathbb{R}^d \setminus \{0\}$, $\|v\| = 1$, $\nabla \cdot (\rho^{\lambda}_{0,\mathbf{c}}\, \Delta h_{t_0}) = \varepsilon\, L(v) + o(\varepsilon)$ in $L^2(\rho^{\lambda}_{0,\mathbf{c}})$, where L is the linear map in the hypothesis; the regularity hypothesis makes the remainder's divergence $o(\varepsilon)$ uniformly over the unit sphere, and on $\mathbb{R}^d$ an injective linear map is bounded below, so $\varepsilon\, L(v)$ dominates the remainder for ε small and $v \neq 0$, and the hypothesis of Prop. 4 holds for every $w' \neq w$ in a neighbourhood of w.*

Remark 2 (Why a naive sensitivity condition does not suffice)**.** *Let $\rho^{\lambda}_0 = \mathcal{N}(\mathbf{0}, \mathbf{I})$, $q \geq 2$, A skew-symmetric and $h(z; w) = wAz$. Then $\Phi^{w}_{t,t_0} = \exp(wtA) \in SO(q)$, so the pushforward is $\mathcal{N}(\mathbf{0}, \mathbf{I})$ for every w and every t, and the whole family is observationally identical. The enrolment window, the open baseline support and the pointwise sensitivity $D_w \Phi^{w}_{t,t_0}(z) = tA\exp(wtA)z$ are all nondegenerate, so conditions of that kind cannot imply FOA. Proposition 4 excludes the example directly, since $\nabla \cdot (\rho^{\lambda}_0 Az) = -z^\top Az\, \rho^{\lambda}_0 = 0$.*

Remark 3 (Interaction between Gaussian pinning and FOA)**.** *The same symmetry issue is relevant after aggregate Gaussian pinning. Stage 1 pins the aggregate latent law to $\mathcal{N}(\mathbf{0}, \mathbf{I})$, and the Gaussian is the most symmetric reference measure available, so where a leaf's baseline inherits that symmetry, for instance when leaves overlap heavily in the latent space, the stabiliser of ρ_0^λ is correspondingly large, and it is that stabiliser, not the aggregate's, that governs temporal identifiability. Pinning the aggregate does not by itself Gaussianise any individual leaf's baseline. Proposition 4 is what separates the two effects, and its statistic is evaluable on a trained model: for each expert pair, evaluate $\|\nabla \cdot (\hat{\rho}_0^\lambda(h(\cdot, t_0, \mathbf{c}; w_k) - h(\cdot, t_0, \mathbf{c}; w_l)))\|_{L^2}$ by Monte Carlo over encoded latents, normalised by $\|h(\cdot, t_0, \mathbf{c}; w_k) - h(\cdot, t_0, \mathbf{c}; w_l)\|_{L^2}$.*

A separation statistic on the trained models. We evaluate Proposition 4's statistic between every pair of trained expert heads, $r_{kl} = \|\nabla \cdot (\hat{\rho}_0 (h(\cdot; w_k) - h(\cdot; w_l)))\|_{L^2(\hat{\rho})} / \|h(\cdot; w_k) - h(\cdot; w_l)\|_{L^2(\hat{\rho})}$, by Monte Carlo over 4,096 encoded latents at t_0 with automatic differentiation, with $\hat{\rho}_0$ the baseline pooled over each benchmark's test units rather than estimated leaf-wise, which the non-implication remark above identifies as the wrong object for the leaf-level statement; the leaf-wise recomputation on unaugmented coordinates is in App. J.8. The reference floor is a resampling floor: the same statistic recomputed between two heads fitted to bootstrap resamples of one leaf, which carries the density-estimation error of $\hat{\rho}_0$ and the head-refitting variability of one leaf. It is a bootstrap reference distribution, not a noise-only floor, so we also report it as a test: on every benchmark every trained expert pair exceeds all 200 same-leaf bootstrap pairs (empirical $p < 0.005$). A degenerate linear system built like Remark 2's example, $\rho_0 = \mathcal{N}(\mathbf{0}, \mathbf{I}_q)$ with $\Delta h(\mathbf{z}) = M\mathbf{z}$ for skew-symmetric M, has exact value zero, so a Hutchinson-estimated value on it measures only probe noise, which the exact-divergence statistic does not have; we report that probe floor as a diagnostic only. Table 5 reports the result: in the strictest evaluation (leaf-wise $\hat{\rho}_0$, unaugmented coordinates) every expert pair on every discrete-leaf benchmark exceeds the resampling floor by at least 1.4 orders of magnitude. Three limits apply. The underlying condition is sufficient, not necessary; it is checked between the K converged heads, which is a local check that does not certify injectivity over all of Θ_w; and the resampling floor is estimated from 20 bootstrap pairs per leaf, so the margins carry about ± 0.1 orders of sampling error.

Table 5: Expert-head separation diagnostic. r_{kl} is the first-order statistic of Prop. 4; margins are orders above the indicated floor under increasingly strict evaluation. This diagnoses, but does not verify, FOA.

			Probe-floor margin			Resamp. margin
BM	Pairs	r_{kl}	pooled aug.	leaf aug.	leaf unaug.	leaf unaug.
BM1	3	0.661–0.845	3.1	2.5	2.4	**2.1**
BM2	3	1.118–1.914	2.4	1.9	1.8	**1.5**
BM4	3	3.185–5.047	2.9	2.3	2.2	**1.9**
BM5	3	1.438–1.515	2.4	1.8	1.7	**1.4**
BM6	1	1.170	3.2	2.6	2.5	**2.2**
BM7	1	1.417	2.8	2.2	2.1	**1.8**
min			2.4	1.8	1.7	**1.4**

Notes. r_{kl} uses pooled $\hat{\rho}_0$. Bold values are the reported margins: leaf-wise $\hat{\rho}_0$, unaugmented coordinates, and the resampling floor. BM3 has no discrete expert pairs. BM7 is descriptive because its marginals are observed only at a few culture days rather than over the continuous window assumed by Prop. 4.

Remark 4 (Failure regimes)**.** *The condition fails exactly when Δh is divergence-free with respect to ρ_0^λ. The analysis therefore predicts degraded parameter recovery when two leaf parameters are close, since Δh then shrinks towards the degenerate case. Symmetries of the baseline are unidentifiable, which is consistent with the residual archetype-direction error in App. I.8; near a fixed point the sensitivity fields shrink and the separation modulus of condition (S2) of Lemma 5 degrades, which is the BM5 near-bifurcation behaviour; and leaves with more concentrated, less symmetric baselines are better identified.*

Under FOA, spatial richness converts pointwise flow divergence into distributional distinguishability. The regimes where FOA fails remain:

- Observations concentrate near a fixed point of the dynamics, where distinct vector fields yield indistinguishable pushforward measures.
- Enrolment times collapse to a single point, eliminating temporal variation.
- The baseline ρ_0^λ concentrates on a "parameter confusion zone" where distinct parameters produce identical flows.

The quantitative strengthening of FOA used in the finite-sample analysis (Lemma 5, condition S2 there) is likewise an assumption: it is *not* derived from the forward-integration dynamics of Section 4, and we claim it only as a modelling condition on the leaf risk's curvature.

B.4 Proof of Theorem 1 (Parameter Identifiability, Frame-Conditional)

Tab. 6 lists, for each proof step, the assumptions it consumes, what it identifies, and on which benchmarks those assumptions hold. Throughout, the *frame* is the pair (lossless compressor E, Gaussianising encoder P), with $\mathrm{Enc} = P \circ E$ and decoder $\mathrm{Dec} = E^{-1} \circ P^{-1}$. Recall the assumptions:

DFA, COA′ (+ SEP under reading (b)), BAS, POS, ENR (in the context form, $t_i \perp \mathbf{z}^i_{t_0} \mid \mathbf{c}_i$), PRW, FOA. MRA is not used; it enters only the basis-recovery statements.

Table 6: Assumptions consumed by each step of the proof of Theorem 1; the steps are numbered in dependency order.

Step	Consumes	Identifies	Holds on
0 Frame	fixed E, P (and θ or $\mathrm{FOA_j}$)	latents and their joint law	assumed; BM4 excluded (lossy E)
1 Marginals	ENR (context form), POS	per-context families $\rho^{\lambda}_{t,\mathbf{c}}$	BM1–BM6; POS fails on BM7
2 Baseline	BAS, POS, Step 1	$\rho^{\lambda}_{0,\mathbf{c}}$ and its leaf mixture	BM1–BM6
3 Partition	(a) COA′(a); (b) Steps 1, 2, per-context FOA, SEP	leaf partition up to labels	(a) BM1–BM4 by construction
4 Parameters	PRW(a), FOA, Steps 1–3	$w^*(\lambda)$	FOA assumed, not verified

Step 0 (Frame). With Enc *fixed*, $z^i_{t_i} = \mathrm{Enc}(x^i_{t_i})$ and the joint latent law are determined by the data law. This is *explicitly frame-conditional*: a different bijective encoder yields different latents, exactly as the conjugation observation of App. B.1 records, so encoder bijectivity alone does not make the latents unique; latents are unique only *given the frame*. Frames form a G-orbit (G1 in Appendix B.2); the orbit is quotiented globally by Theorem 2.

Throughout, T_λ is the support of the enrolment law on leaf λ, not the evaluation window: the data determine the marginal family only where units enrol. On BM4, enrolment is confined to $[40, 50]$ while forecasts run to 80 (App. F.5), so that benchmark's identified family covers the first quarter of its horizon and the remainder is extrapolation of the fitted flow.

Step 1 (Latent marginal families). Under ENR in its context form, $t_i \perp \mathbf{z}^i_{t_0} \mid \mathbf{c}_i$, conditioning the latent law on $(\mathbf{c}, t)$ identifies the per-context family $\{\rho^{\lambda}_{t,\mathbf{c}} : t \in T_\lambda,\ \mathbf{c} \in L_\lambda\}$ at population level, for $P_{\mathcal{C}}$-almost every context, and aggregating over $\mathbf{c} \in L_\lambda$ gives the leaf family $\{\rho^{\lambda}_t\}$. ENR is what licenses reading each conditional law as a marginal rather than a state-biased slice of it: without it $p(\mathbf{z} \mid \mathbf{c}, t)$ is biased by the enrolment mechanism, and both this step and Step 4 fail. Step 4 consumes the context form. The context and leaf forms are not nested (Remark below); the leaf form, which Proposition 10 uses, follows from the context form together with $t_i \perp \mathbf{c}_i \mid \lambda(i)$. Dependency order: Step 2 depends only on Step 1, and Step 3's partition argument consumes Step 2's baselines; under COA′(b) Step 4 is vacuous (the per-context reading already determines $w(\lambda(\mathbf{c}))$). (Finite-sample version: the temporal KDE of Lemma 4, with the uniform-convergence requirement stated in Appendix B.9.)

(POS). Steps 1 and 2 read the observable law at $(\mathbf{c}, t)$ and take a one-sided limit at t_0, so both consume a positivity condition, POS, which Theorem 1 lists among its hypotheses: for $P_{\mathcal{C}}$-a.e. $\mathbf{c} \in L_\lambda$ the enrolment law $P(t \mid \mathbf{c})$ must charge T_λ, with t_0 right-adherent to its support. Without it a context slice can be observed at a single time, the per-context family collapses to one measure, and Step 4 has nothing to apply FOA to; enrolment determined by context, $t = \tau(\mathbf{c})$ with τ injective, satisfies ENR and leaf-level BAS and still defeats the argument. POS holds by construction on BM1–BM6, whose enrolment times are drawn independently of state and context, and fails on BM7, where every clone enrols at culture day 2; this is one more reason Theorem 1 is not claimed there.

Step 2 (Baseline). By BAS, applied within each context slice, $\rho^{\lambda}_{0,\mathbf{c}}$ is the identified t_0-marginal (one-sided limit) of the Step-1 per-context family, and ρ^{λ}_0 is its $\mathbf{c}$-mixture. (This licenses Theorem 2's C3 line $\tilde{\rho}^{\pi(\lambda)}_0 = \varphi_{\#}\rho^{\lambda}_0$; without BAS that line would over-reach, since the data determine only $\{\rho^{\lambda}_t : t \in T_\lambda\}$.)

Step 3 (Partition). Under COA′(a): the leaf label is computable from context, so the partition is immediate. Under COA′(b): the partition is recovered through the per-context parameter map, not by clustering raw conditional families. Steps 1 and 2 identify, for every context $\mathbf{c}$, the family $\{\rho^{\lambda}_{t,\mathbf{c}}\}$ and its baseline $\rho^{\lambda}_{0,\mathbf{c}}$ (Step 2 uses only Step 1 and BAS), and the per-context form of FOA (Assumption 5) makes this identified pair determine $w(\lambda(\mathbf{c}))$ uniquely; this uses FOA's injectivity hypothesis on identified inputs, not Step 4's conclusion. Define $\mathbf{c} \sim \mathbf{c}'$ iff $w(\lambda(\mathbf{c})) = w(\lambda(\mathbf{c}'))$. DFA gives same-leaf $\Rightarrow \sim$ (one leaf, one w); SEP (Assumption 6: distinct leaves carry distinct parameters) gives distinct-leaf $\Rightarrow \not\sim$; hence $\sim$ recovers $\{L_\lambda\}$ exactly, modulo $P_{\mathcal{C}}$-null context sets, with labels arbitrary, which is the permutation ambiguity in the conclusion. (This is the step Theorem 2's C2 cites; the partition is established here, independently and prior to that use.) The argument is at population level. In finite samples, recovering the partition under COA′(b) needs repeated contexts, or continuity in context so that neighbouring contexts can be pooled; the finite-sample results of App. B.9 take a consistent leaf assignment as an input and do not derive it. *Remark (why not cluster raw families).* Clustering contexts by their raw conditional latent families can split a single leaf: if the within-leaf baselines $\rho^{\lambda}_{0,\mathbf{c}}$ vary with $\mathbf{c}$, contexts with different baselines carry different conditional families although they share $w(\lambda)$, even when the field is $\mathbf{c}$-independent. The identified parameter, not the raw family, is the partition's invariant; this matters on BM7, where the day-2 context varies within a fate. *Finite-sample rider:* recovering $\sim$ from data

requires either repeated contexts or continuity of the conditional family in $\mathbf{c}$, so nearby contexts can be pooled, the context-space analogue of Lemma 4's temporal KDE. Stated as an assumption wherever COA′(b) is used.

Step 4 (Within-leaf w). The identified family satisfies $\rho^{\lambda}_{t,\mathbf{c}} = (\Phi^{w^*(\lambda),\mathbf{c}}_{t,t_0})_{\#}\rho^{\lambda}_{0,\mathbf{c}}$ $(\mathbf{c} \in L_\lambda)$ with $\rho^{\lambda}_{0,\mathbf{c}}$ now known (Step 2), so FOA's injectivity hypothesis is applied to *identified* inputs. The relation is understood per context $\mathbf{c}$ within the leaf; the leaf marginal is the $\mathbf{c}$-mixture of these, and it is the per-context family that FOA constrains. $w^*(\lambda)$ is unique at the fixed θ of PRW (jointly $(\theta, w^*(\lambda))$ under joint injectivity). A misspecified expert basis ($\varepsilon > 0$) lies outside Theorem 1; Lemma 5 (Appendix B.10) bounds that case. Under FOA this step follows directly from the definition of injectivity.

Step 5 (Assembly). Steps 3 + 4 give $(g^*, \{w(\lambda)\})$ up to a single leaf-label permutation. □ *Remark (where the context form is used, and why it is free here).* Steps 1, 2 and 4 all read the observable law per context, because h_θ carries $\mathbf{c}$; the leaf-level statements are the $\mathbf{c}$-mixtures of the per-context ones. The context and leaf forms of ENR are not nested. The leaf form does not imply the context form: within one leaf, enrolment can be state-dependent in each context while the two biases cancel in the leaf mixture. Nor does the context form imply the leaf form: take one leaf, $\mathbf{c} \in \{0, 1\}$ equiprobable, $z_{t_0} \mid \mathbf{c} \sim \mathcal{N}(\mathbf{c}, 1)$ and enrolment densities $2(1 - t)$ given $\mathbf{c} = 0$ and $2t$ given $\mathbf{c} = 1$ on $[0, 1]$; then $t \perp z_{t_0} \mid \mathbf{c}$ holds while $P(\mathbf{c} = 1 \mid t) = t$, so $\mathbb{E}[z_{t_0} \mid t, \lambda] = t$. The leaf form follows once $t_i \perp \mathbf{c}_i \mid \lambda(i)$ is added. On BM1–BM6 enrolment times are drawn independently of state and context by construction (App. F.1), and on BM7 every clone enrols at culture day 2, so both forms and the added condition hold on all seven benchmarks and none of this costs anything empirically.

B.5 The Expert Count

This subsection supplies the machinery behind the case split stated after Assumption 3: Proposition 8 covers finite Λ ($K = |\Lambda|$ is exact, and minimal under affine independence), Lemma 1 and Propositions 5–6 the continuous case, and Proposition 7 the lower bound for exact representation. Lemmas 2 and 3 below are special cases of the approximate-anchor recovery framework of Appendix B.11, between whose two regimes (exact anchors / nonempty simplex interior) the practical runs sit.

Lemma 1 (Lipschitz images do not raise Hausdorff dimension)**.** *Let $A \subset \mathbb{R}^r$ and $F : A \to \mathbb{R}^d$ be Λ_L-Lipschitz. Then* $\dim_{\mathrm{H}} F(A) \le \dim_{\mathrm{H}} A$ $(\le r)$.

Proof. Fix $s > \dim_{\mathrm{H}} A$, so $\mathcal{H}^s(A) = 0$. Given $\eta, \delta > 0$, take a cover $\{U_i\}$ of A with $\operatorname{diam} U_i \le \delta$ and $\sum_i (\operatorname{diam} U_i)^s < \eta$. Lipschitzness gives $\operatorname{diam} F(U_i \cap A) \le \Lambda_L \operatorname{diam} U_i \le \Lambda_L \delta$, and $\{F(U_i \cap A)\}$ covers $F(A)$, so $\sum_i (\operatorname{diam} F(U_i \cap A))^s \le \Lambda_L^s \eta$. Let $\eta \to 0, \delta \to 0$: $\mathcal{H}^s(F(A)) = 0$ for all $s > \dim_{\mathrm{H}} A$. □

Proposition 5 (Finite ε-approximation)**.** *$\mathcal{W}$ compact ⇒ for all $\varepsilon > 0$ there exist $w_1, \dots, w_K \in \mathcal{W}$ with* $\sup_{w \in \mathcal{W}} \operatorname{dist}(w, \operatorname{conv}\{w_k\}) \le \varepsilon$*, with $K = N(\mathcal{W}, \varepsilon)$ the covering number.*

Proof. By total boundedness $\mathcal{W} \subseteq \bigcup_k B(w_k, \varepsilon)$ with $w_k \in \mathcal{W}$; for $w \in \mathcal{W}$, some $|w - w_k| \le \varepsilon$, and $w_k \in \operatorname{conv}\{w_j\}$. □

Proposition 6 (Bound on $K(\varepsilon)$)**.** *If $\mathcal{W} = F(\mathcal{C})$ with $\mathcal{C} \subset \mathbb{R}^r$ compact of radius $R_{\mathcal{C}}$ and F Λ_L-Lipschitz, then $K(\varepsilon) \le N(\mathcal{C}, \varepsilon/\Lambda_L) \le (1 + 2\Lambda_L R_{\mathcal{C}}/\varepsilon)^r$. The exponent r comes from the* domain*; the bound needs only compactness and Lipschitz F.*

Proposition 7 (Necessity, exact representation only)**.** *If $\mathcal{W} \subseteq \operatorname{conv}\{w_k\}$* exactly*, then $K \ge \dim \operatorname{aff}(\mathcal{W}) + 1$.*

Proof. $\operatorname{aff}(\mathcal{W}) \subseteq \operatorname{aff}\{w_k\}$ and $\dim \operatorname{aff}\{w_k\} \le K - 1$. □

Remark. Under ε-approximation no dimension-only lower bound holds without separation: the unit circle S^1 is witnessed by an antipodal pair ($K = 2$ at $\varepsilon = 1$) although $\dim \operatorname{aff}(\mathcal{W}) + 1 = 3$.

Proposition 8 (Discrete leaves: threshold exactly $|\Lambda|$)**.** *Let $\mathcal{W}$ be finite with $|\mathcal{W}| = L$. With $\varepsilon(K) := \min_{\{w_k\} \subseteq \mathcal{W}, |\{w_k\}| \le K} \sup_{w \in \mathcal{W}} \operatorname{dist}(w, \operatorname{conv}\{w_k\})$ for $K \ge 1$, minimising over subsets of size* at most *K, which is always attained and requires no separate convention for $K > L$: $\varepsilon(K) = 0$ iff $\{w_k\} \supseteq \operatorname{ext}\operatorname{conv}(\mathcal{W})$; and if the L leaf parameters are affinely independent (Lemma 2's condition), $\operatorname{ext}\operatorname{conv}(\mathcal{W}) = \mathcal{W}$, so $\varepsilon(K) > 0$ for $K < L$ and $= 0$ for $K \ge L$.*

Proof. $\mathcal{W} \subseteq \operatorname{conv}\{w_k\}$ implies $\operatorname{conv}\mathcal{W} = \operatorname{conv}\{w_k\}$, whose extreme points must lie in any generating set. Affine independence makes $\operatorname{conv}\mathcal{W}$ a simplex, so each w is a vertex; omitting one leaves a proper face, and $\varepsilon(L - 1) \ge \operatorname{dist}(\text{vertex}, \text{opposite facet}) > 0$. □

Remark (unconstrained dictionaries). $\varepsilon(K)$ minimises over dictionaries drawn from $\mathcal{W}$, whereas the learned $\mathbf{W} \in \mathbb{R}^{d \times K}$ is unconstrained. The threshold is unchanged: under affine independence $\operatorname{conv}\mathcal{W}$ has affine dimension $L - 1$, and a polytope that is the convex hull of K points and contains an $(L - 1)$-dimensional set

needs $K \geq L$ vertices wherever they lie. What the restriction does provide is interpretability, since each expert is then an archetype in $\mathcal{W}$ rather than an arbitrary enclosing vertex.

Remark 5 (K selection, heuristic)**.** *Estimation cost increasing in K (an assertion, not a theorem) together with Proposition 8 implies any optimum $K \leq |\Lambda|$, with point prediction $K = |\Lambda|$* among *$K \geq |\Lambda|$; ruling out $K < |\Lambda|$ is not settled by theory. The component ablation (Table 27 in Appendix I.3) locates the BM1 minimum at $K = 3 = |\Lambda|$, confirming the construction-given default rather than selecting it.*

Lemma 2 (Basis identifiability under anchor conditions)**.** *Under Theorem 1 and MRA ($\varepsilon = 0$), if additionally (i) for each $k \in [K]$ there exists λ_k with $\alpha(\lambda_k) = e_k$ (anchor leaves), and (ii) $\{w_1, \ldots, w_K\}$ are affinely independent, then $\mathbf{W} = [w_1, \ldots, w_K]$ is identifiable up to column permutation.*

Proof. At anchor leaves $\hat{w}(\lambda_k) = \mathbf{W}e_k = w_k$ directly recovers each column. Affine independence ensures distinctness; routing is recovered from the *affine* system, $\alpha(\lambda) = \left[\begin{smallmatrix} \mathbf{W} \\ \mathbf{1}^\top \end{smallmatrix}\right]^\dagger \left[\begin{smallmatrix} \hat{w}(\lambda) \\ 1 \end{smallmatrix}\right]$, equivalently least squares subject to $\mathbf{1}^\top \alpha = 1$. The unaugmented $\mathbf{W}^\dagger$ returns linear rather than barycentric coordinates and need not lie on the simplex. □

Lemma 2 and Lemma 3 below describe the two *exact* regimes of basis recovery, and neither is the operative one: at finite softmax temperature every routing weight is strictly positive, so no leaf is an exact anchor. Lemma 7 (Appendix B.11) supersedes them for practice: it quantifies recovery when anchors are only *approximate* ($\alpha(\lambda_k) \approx e_k$ at finite softmax temperature) and per-leaf estimates are finite-sample, with an end-to-end bound $O(K^{3/2}e^{-\Delta/\tau}) + O(\sqrt{Kr_N})$ whose zero-defect limit is the exact-anchor regime of Lemma 2.

B.6 Proof of Theorem 2 (Composition Identifiability in Quotient Space)

Theorem 2 (formal statement). Fix the compressor E (lossless, $q = p$, $\mathrm{Dec} = \Psi \circ P^{-1}$ with $\Psi = E^{-1}$) and let $\nu := E_{\#}\, p_{\mathrm{data}}$ have smooth positive density. Let $M = (P, h_\theta, g^*, w)$ and $\tilde{M} = (\tilde{P}, h_{\tilde{\theta}}, \tilde{g}^*, \tilde{w})$ be two models in the class, with $P, \tilde{P}$ unconditional Gaussianising PF-ODEs for ν and dynamics in $\mathcal{H}$, both exactly consistent with the cross-sectional data law, under DFA, COA′, SEP (if reading (b)), BAS, POS, ENR, PRW, FOA (and $\mathrm{FOA_j}$ for the vector-field clause), and $t \perp \mathbf{c} \mid \lambda$ wherever a leaf-level mixture is used. Then there exist $\varphi \in G$ and a leaf-label permutation π such that:

(i) $\widetilde{\mathrm{Enc}} = \varphi \circ \mathrm{Enc}$;
(ii) $\tilde{g}^* = \pi \circ g^*$;
(iii) for every leaf λ, every context $\mathbf{c} \in L_\lambda$ and $t \in T_\lambda$,

$$(\widetilde{\Phi}^{\tilde{w}(\pi(\lambda)),\mathbf{c}}_{t,t_0})_{\#}(\varphi_{\#}\rho^{\lambda}_{0,\mathbf{c}}) = \varphi_{\#}(\Phi^{w(\lambda),\mathbf{c}}_{t,t_0})_{\#}\rho^{\lambda}_{0,\mathbf{c}},$$

i.e. the flows are conjugate by φ as measure transports on the observed window and baseline support. If additionally $\varphi \in G_H$ (the conjugated field $D\varphi^{-1} \cdot h_{\tilde{\theta}} \circ \varphi$ again lies in $\mathcal{H}$), the conjugacy holds at the vector-field level: $h_{\tilde{\theta}}(\cdot, \cdot, \mathbf{c}; \tilde{w}(\pi(\lambda))) = D\varphi \cdot h_\theta(\cdot, \cdot, \mathbf{c}; w(\lambda)) \circ \varphi^{-1}$ for every $\mathbf{c} \in L_\lambda$.

The encoders are unconditional. A context-indexed encoder $P_{\mathbf{c}}$, as deployed on BM1–BM6, leaves a family $\{\varphi_{\mathbf{c}}\}$ rather than one φ, and two per-context rotations inside one leaf already defeat a single conjugacy of the leaf mixtures; that case is outside this statement (Tab. 37, first row). MRA plays no role here; it enters only the basis-recovery statements (Lemmas 2 and 3, Cor. 3). Conversely (G4 in Appendix B.2), any such conjugate is observationally equivalent. Hence $[M] \in \mathcal{M}/G$ is data-determined; under the extra hypothesis $\varphi \in G_H$, $[M] \in \mathcal{M}/G_H$, a *finer* quotient, hence a strictly stronger statement requiring that extra hypothesis.

Remark (how large G_H is). For the architectures used here G_H is small. The FiLM of Sec. 4 acts coordinate-wise on the output, $s(w) \odot h_{\mathrm{base}} + b(w)$, with one backbone shared by all leaves. Conjugating by $R \in O(q)$ turns the scale into $R^\top \mathrm{diag}(s(w))R$, which must be diagonal for every leaf's $s(w)$. If the scale vectors $\{s(w(\lambda))\}_\lambda$ jointly separate coordinates, i.e. no two coordinates carry equal scales in every leaf, which holds generically, this forces R to be a signed permutation, and $G_H \cap O(q)$ lies in the hyperoctahedral group of coordinate permutations and sign flips; without that condition larger orthogonal subgroups remain (with isotropic scales $s(w) = a(w)\mathbf{1}$ every R qualifies). A non-affine twist T_f of App. B.2 generically produces a conjugated field $DT_f \cdot h \circ T_f^{-1}$ whose coordinate-dependent Jacobian no weight setting of the same architecture reproduces, so $T_f \notin G_H$. The vector-field clause therefore identifies the dynamics up to G_H, whose orthogonal part is generically hyperoctahedral and which generically excludes the twists; we do not characterise its non-orthogonal part, and the measure-transport clause, which holds for all of G, is the operative statement.

Proof.

Soundness (⇐). By G4: for any bijection φ, $\mathrm{Dec} \circ \varphi^{-1}(\varphi \circ \Phi_t^w \circ \varphi^{-1}(\varphi \circ \mathrm{Enc}(x))) = \mathrm{Dec}(\Phi_t^w(\mathrm{Enc}(x)))$, so conjugates are prediction-identical. This direction uses bijectivity alone.

Completeness (⇒).

C1 (encoder residual). Both models are in the class, so $P_{\#}\nu = \tilde{P}_{\#}\nu = \gamma_q$. By G1, $\varphi := \tilde{P} \circ P^{-1} \in G$, and $\widetilde{\mathrm{Enc}} = \tilde{P} \circ E = (\tilde{P} \circ P^{-1}) \circ (P \circ E) = \varphi \circ \mathrm{Enc}$. This gives (i); latent frames relate by $\tilde{z} = \varphi(z)$ on the support of $E_{\#}p_{\mathrm{data}}$.

C2 (leaf map). By COA′ both models assign leaves from the same contexts; by Theorem 1 Step 3 (Appendix B.4) the leaf *partition* of $\mathcal{C}$ is determined by the data law, so the label maps agree up to a bijection of labels: $\tilde{g}^* = \pi \circ g^*$. This gives (ii).

C3 (matching marginal families in a common frame). Since E is lossless and $\mathrm{Dec} = \Psi \circ P^{-1}$ is the exact inverse of Enc, the data law determines, in frame 1, the per-context marginal families $\mu_t^{\lambda,\mathbf{c}} := (\Phi_{t,t_0}^{w(\lambda),\mathbf{c}})_{\#}\rho_{0,\mathbf{c}}^{\lambda}$ for $\mathbf{c} \in L_\lambda$, $t \in T_\lambda$; conclusion (iii) is per context and needs no leaf-level object (their leaf mixture is a time-invariant mixture of these only under the added condition $t \perp \mathbf{c} \mid \lambda$, which holds on all seven benchmarks, Tab. 37) (Theorem 1 Step 1 + Lemma 4). By **BAS** (Theorem 1 Step 2), $\rho_{0,\mathbf{c}}^{\lambda}$ is the identified t_0-marginal in each context slice, so invoking ρ_0^{λ} here is licensed. (Without BAS, or if $t_0 \notin T_\lambda$, phrase (iii) over the observed family only, not ρ_0^{λ}.) In frame 2 the same data determine $\tilde{\mu}_t^{\pi(\lambda),\mathbf{c}}$; by C1 the frames relate by φ, so $\tilde{\mu}_t^{\pi(\lambda),\mathbf{c}} = \varphi_{\#}\mu_t^{\lambda,\mathbf{c}}$ and $\tilde{\rho}_{0,\mathbf{c}}^{\pi(\lambda)} = \varphi_{\#}\rho_{0,\mathbf{c}}^{\lambda}$ per context, hence $\tilde{\rho}_0^{\pi(\lambda)} = \varphi_{\#}\rho_0^{\lambda}$. Exact consistency of model 2 gives $(\widetilde{\Phi}_{t,t_0}^{\tilde{w}(\pi(\lambda)),\mathbf{c}})_{\#}\tilde{\rho}_{0,\mathbf{c}}^{\pi(\lambda)} = \tilde{\mu}_t^{\pi(\lambda),\mathbf{c}}$ for all $t \in T_\lambda$ and $\mathbf{c} \in L_\lambda$. Substituting yields (iii).

C4 (stabiliser case: conjugacy of dynamics). Suppose $\varphi \in G_H$, i.e. the conjugated field $\check{h} := D\varphi^{-1} \cdot h_{\tilde{\theta}} \circ \varphi \in \mathcal{H}$ (the conjugation observation of App. B.1). Pulling model 2's flow back to frame 1, $\check{\Phi}_t := \varphi^{-1} \circ \widetilde{\Phi}_t^{\tilde{w}(\pi(\lambda))} \circ \varphi$ is the flow of $\check{h}$ and, by C3, generates from each $\rho_{0,\mathbf{c}}^{\lambda}$ the *same* per-context marginal family as $\Phi^{w(\lambda),\mathbf{c}}$ over T_λ. The conjugated field need not carry the same class index θ, so the fixed-θ form of FOA does not apply directly; we invoke FOA in its joint form ($\mathrm{FOA_j}$, App. B.3), injectivity of (θ, w) in the induced marginal family, which forces the two pairs to coincide and gives: $h_{\tilde{\theta}}(\cdot; \tilde{w}(\pi(\lambda))) = D\varphi \cdot h_\theta(\cdot; w(\lambda)) \circ \varphi^{-1}$. □

Four caveats.

- *Support/window.* C3 yields equality of the pushforward families only, on the observed windows $t \in T_\lambda$; it says nothing pointwise, and off-support and off-window behaviour is not data-constrained. Matching cross-sectional marginals alone does *not* give pointwise flow conjugacy; $\mathrm{FOA_j}$ with $\varphi \in G_H$ is what upgrades it to a vector-field statement, and only within $\mathcal{H}$ (C4).
- *G_H hypothesis.* $\varphi \in G_H$ is *not* implied by class membership and must be assumed (or verified) for the vector-field conclusion, together with $\mathrm{FOA_j}$ (for the vector-field clause); without it, only the transport-level (iii) and the $\mathcal{M}/G$ quotient hold. The hypothesis quantifies over leaves: $D\varphi^{-1} \cdot h_{\tilde{\theta}}(\cdot\,; \tilde{w}(\pi(\lambda))) \circ \varphi \in \mathcal{H}$ is required for every $\lambda \in \Lambda$ under the one shared backbone $\tilde{\theta}$, so that C4's argument applies leaf-by-leaf; a φ satisfying it for a single leaf supports only that leaf's conclusion.
- *Exact vs approximate.* Under the exact-terminal-law idealisation, G3 pins $\varphi = \mathrm{id}$ and the quotient is trivial; under approximate score matching the whole statement holds only approximately (G3).
- *When the quotient is non-trivial.* If both models use the same forward process, schedule and horizon and both attain the exact score, G3 collapses $\mathcal{M}/G$ to a point. The quotient carries content exactly when the forward process is free across models, or when the score is attained only approximately, which is the situation for every trained model.

B.7 On the Role of PRW

Remark (Why PRW is essential). Without restricting to a parametric family $\mathcal{H}$, Theorem 1 would be false: many nonparametric vector fields produce identical marginal distributions from cross-sectional samples (the inverse transport problem from marginals alone is known to be ill-posed without regularization). PRW is what makes the inverse problem well-posed, and it marks the boundary of what cross-sectional data determines: inside a parametric class that satisfies FOA the marginals pin the parameters; outside such a class they need not, and Prop. 1 shows that parametric restriction alone is not enough.

B.8 Temperature Annealing Convergence

Proposition 9 (Temperature Annealing). *Let $\boldsymbol{\alpha}_i(\tau) = \mathrm{Softmax}(g_\omega(\mathbf{c}_i)/\tau)$ for $\tau > 0$, and let $g_\omega(\mathbf{c}_i)$ have a unique maximiser $\kappa(i)$ (for every unit considered, or almost surely under the context law). As $\tau \to 0^+$, $\boldsymbol{\alpha}_i(\tau) \to \mathbf{e}_{\kappa(i)}$. If the routing network is such that $\arg\max_k [g_\omega(\mathbf{c}_i)]_k = \kappa(i)$, this limit recovers the hard DFA assignment; the limit itself is a property of the softmax and not of learning.*

Proof. For $v \in \mathbb{R}^K$ with unique maximiser k^*, $[\mathrm{Softmax}(v/\tau)]_k = (1+\sum_{j\neq k} e^{-(v_k-v_j)/\tau})^{-1} \to \mathbf{1}[k=k^*]$ as $\tau \to 0^+$, since $e^{-\Delta/\tau} \to 0$ for $\Delta > 0$. Holds for Lebesgue-a.e. v (ties have measure zero). □

Remark (connection to approximate-anchor recovery). Proposition 9 is the zero-defect limit of the finite-temperature analysis in Appendix B.11: as $\tau \downarrow 0$ the anchor defect $\bar{\delta} \leq 2(K-1)e^{-\Delta/\tau}$ of Lemma 7 vanishes and the end-to-end recovery bound collapses to the exact-anchor regime of Lemma 2. Practical runs terminate at a finite $\tau_{\mathrm{end}} = 0.1$, sitting between the two regimes of Corollary 2; the resulting defect is quantified there rather than assumed away.

Lemma 3 (Basis identifiability via minimum-volume simplex)**.** *Assume MRA ($\varepsilon = 0$), let $\{w_k\}$ be affinely independent and let $A := \{\alpha(\lambda) : \lambda \in \Lambda\} \subset \Delta^{K-1}$ be the barycentric coordinates of $\hat{w}(\Lambda)$. If $\mathrm{conv}(A)$ contains $\{\alpha \in \Delta^{K-1} : \|\alpha\|_2 \leq r\}$ for some $r > 1/\sqrt{K-1}$ (the uniform purity condition of Lin et al. (2015); on the simplex $\|\alpha\|_2^2 = \|\alpha - \mathbf{1}/K\|_2^2 + 1/K$, so this is a ball about the barycentre, and for $K = 2$ it is the whole simplex and the condition reduces to Lemma 2's anchor condition), then $\mathrm{conv}(\mathbf{W})$ is the unique minimum-volume simplex containing $\hat{w}(\Lambda)$, and $\mathbf{W}$ is identifiable up to column permutation without anchor leaves.*

Proof sketch. By construction $\hat{w}(\Lambda) \subseteq \mathrm{conv}(\mathbf{W})$. Non-empty relative interior does not identify $\mathrm{conv}(\mathbf{W})$, and neither does meeting every facet: the medial triangle of a triangle T meets all three edges and is itself a simplex of area $|T|/4$ containing it. The uniform purity condition excludes both, and under it the minimum-volume simplex enclosing the barycentric set is Δ^{K-1} itself (Lin et al., 2015, Thm. 1); mapping back through the affine map $\alpha \mapsto \mathbf{W}\alpha$, which preserves volume ratios, gives $\mathrm{conv}(\mathbf{W})$. Its vertices are the columns of $\mathbf{W}$ up to permutation. □

B.9 From the Cross-Sectional MMD Objective to FOA Identifiability

Section 5 states identifiability of the model class at population level under FOA, while training minimises a finite-sample MMD criterion (Eq. 3). This subsection relates the two: for an idealised version of that objective, with hard, sample-split leaf assignments and canonical-time pushforwards, the per-leaf minimiser is a consistent estimator of the FOA-identified parameters, under the leaf-marginal form of FOA (Assumption 5) rather than the per-context form of Theorem 1. Where the trained objective departs from that idealisation is stated after Proposition 10.

Main-text notation. Section 4's cross-sectional loss (Eq. 3) is defined using the SMoE routing weights $\alpha_{i,k}$ over the per-archetype predicted and reference marginals

$$\hat{\mu}_t^{(k)} \propto \sum_{i:t_i \leq t} \alpha_{i,k}\, \delta_{\hat{\mathbf{z}}_t^i}, \qquad \mu_t^{\mathrm{ref},(k)} \propto \sum_{j=1}^{N} \alpha_{j,k}\, \kappa_h(t_j - t)\, \delta_{\mathbf{z}_{t_j}^j}, \tag{5}$$

i.e. the routing-weighted empirical distribution of predicted states at time t and a temporally smoothed (κ_h), routing-weighted reference built from the observed latents.

We structure this consistency argument in two modular stages. Decoupling separates the estimation of the reference measure from the recovery of the routing parameters:

1. **Leaf-level argmin consistency** (Proposition 10): *Given* a consistently estimated leaf assignment, the per-leaf MMD argmin is a consistent estimator of the FOA-faithful parameters $w^\star(\lambda)$. This follows from standard M-estimation under a quantitative FOA condition.
2. **Dictionary recovery** (Corollary 2): Given consistent leaf-level estimates $\hat{w}_N(\lambda) \to w^\star(\lambda)$, the basis $\mathbf{W}^\star$ and routing $\boldsymbol{\alpha}^\star(\lambda)$ are identified up to column permutation. We delineate two distinct convergence regimes governed by the annealing schedule: an *anchor regime* ($\tau_N \to 0^+$ for one-hot routing) and a *simplex-interior regime* ($\tau_N \to \tau_0 > 0$ to preserve soft routing).

By decoupling these steps, we avoid the complex uniform convergence bounds that would otherwise be required if routing weights and reference measures were analyzed simultaneously. We discuss the standalone consistency of the leaf assignment $\hat{\lambda}_N$ (e.g., via the COA′-implied context-to-leaf map) at the end of this subsection.

Setup. Let N index units in the cross-sectional ensemble and N_λ denote the number of units in leaf λ. Denote by $h_N > 0$ the bandwidth of the temporal Gaussian kernel $\kappa_{h_N} : \mathbb{R} \to \mathbb{R}$; the smoothing acts on the scalar enrolment-time axis only, so $d_{\mathcal{T}} = 1$. Let τ_N denote the routing softmax temperature. Assume the temporal observation interval $\mathcal{T} \subset \mathbb{R}$ is compact and the parameter space Θ_w is compact, with $h_\theta \in \mathcal{H} = \{h_\theta(\cdot\,; w) : w \in \Theta_w\}$ (PRW). Let $k_\sigma : \mathbb{R}^q \times \mathbb{R}^q \to \mathbb{R}$ be a bounded translation-invariant *characteristic* kernel (Sriperumbudur et al., 2010) (Gaussian, fixed bandwidth σ, in the analysis); $\mathrm{MMD}_{k_\sigma}(\mu,\nu) = 0$ iff $\mu = \nu$. The analysis fixes k_σ; the implementation sets σ by the batch median heuristic (Tab. 9), which is a data-dependent

bandwidth and therefore a further departure of the trained objective from the idealised one, alongside the four gaps below. We write $\rho_t^\lambda := (\Phi_{t,t_0}^{w^\star(\lambda)})_\# \rho_0^\lambda$ for the FOA-faithful pushforward of leaf λ at time t and assume the joint density $(t, z) \mapsto p^\lambda(z \mid t)$ of ρ_t^λ is continuous on $\mathcal{T}^\circ \times \mathbb{R}^q$; this naturally follows from the continuity of the baseline density ρ_0^λ and the smoothness of the ODE flow under PRW.

For the bandwidth: the standard univariate-KDE consistency rate is $h_N \to 0$, $Nh_N \to \infty$. A common explicit choice is $h_N = N^{-a}$ for any $a \in (0, 1)$; under twice-differentiable temporal densities, the MSE-optimal univariate-KDE rate is $h_N \asymp N^{-1/5}$. Boundary effects are avoided by restricting integration to $\mathcal{T}^\circ$.

Quantitative FOA. The standard FOA (Assumption 5) provides *population-level* injectivity: distinct parameters yield distinct pushforward families. To operationalize this for statistical argmin-consistency, we define a stable, *Quantitative FOA* (QFOA) condition. Under our setup (Θ_w compact, $w \mapsto (\Phi_{t,t_0}^w)_\# \rho_0^\lambda$ continuous in MMD_{k_σ} uniformly in $t \in \mathcal{T}_\lambda$), FOA strictly implies QFOA when FOA is read in its leaf-marginal form; the per-context form of Assumption 5 alone does not imply it (non-implication remark below):

$$\forall\, \eta > 0, \quad \inf_{w \in \Theta_w :\, \|w - w^\star(\lambda)\| \geq \eta} \int_{\mathcal{T}_\lambda} \mathrm{MMD}_{k_\sigma}^2 \Big((\Phi_{t,t_0}^w)_\# \rho_0^\lambda,\ \rho_t^\lambda \Big)\, \mathrm{d}\nu_\lambda(t) \ > \ 0. \tag{6}$$

This holds because the integrand is continuous in w, vanishes only at $w = w^\star(\lambda)$ (due to leaf-level FOA and the characteristic kernel), and the infimum over the compact set $\{w : \|w - w^\star(\lambda)\| \geq \eta\} \subset \Theta_w$ is attained and strictly positive. Equation 6 serves as the well-separated-minimum condition required for M-estimation. The finite-sample analyses elsewhere (Lemma 5 in Appendix B.10) use the equivalent *modulus* form (S2 there): $L_\lambda(w) \geq \omega_\lambda(\|w - w^\star(\lambda)\|)$ for a strictly increasing modulus ω_λ with $\omega_\lambda(0) = 0$, for example $c_\lambda s^2$ (nondegenerate minimum) or $c_\lambda s$ (transversal). This quantitative FOA is an *assumption*; the compactness argument above supplies the positivity form, not a rate.

Non-implication remark (leaf level vs. per context). The leaf-level risk L_λ below integrates the *leaf marginals* ρ_t^λ, the $\mathbf{c}$-mixtures of the per-context families. The per-context form of Assumption 5 does *not* imply leaf-level injectivity: if $h_\theta(\cdot, t, \mathbf{c}; w)$ depends on $\mathbf{c}$ within the leaf, two parameters $w \neq w'$ can produce per-context families that differ on every slice while their $\mathbf{c}$-mixtures coincide, so $L_\lambda(w) = L_\lambda(w')$ and the leaf-level minimum is not unique. Every leaf-level statement in this appendix and in App. B.10 therefore assumes FOA in its leaf-marginal form. Where h_θ is $\mathbf{c}$-independent within a leaf and the baseline does not vary with $\mathbf{c}$, the per-context and leaf-marginal forms coincide. The deployed $h_\theta(\mathbf{z}, t, \mathbf{c}; w_i) = s(w_i) \odot h_{\mathrm{base}}(\mathbf{z}, t, \mathbf{c}) + b(w_i)$ passes $\mathbf{c}$ to h_{base}, so units in one leaf share a dynamical *parameter* but not a *vector field*, and the deployed model is therefore always in the non-coinciding case: Assumption 1 is to be read as parameter-sharing, and every leaf-level statement here is consumed in its leaf-marginal form. A $\mathbf{c}$-free h_{base}, which places the deployed model inside the coinciding case, costs 8–9% MAE (App. J.5).

Proposition 10 (Leaf-level MMD-to-FOA argmin consistency, idealised estimator)**.** *Fix a leaf λ. Suppose the per-unit leaf assignment is either known or estimated by $\hat{\lambda}_N$ with $\Pr(\hat{\lambda}_N(i) \neq \lambda(i)) \to 0$ as $N \to \infty$. Define the leaf-level temporally smoothed empirical reference*

$$\hat{\rho}_{t,N}^\lambda \ \propto \sum_{j : \hat{\lambda}_N(j) = \lambda} \kappa_{h_N}(t_j - t)\, \delta_{\mathbf{z}_{t_j}^j},$$

the population leaf-level risk

$$L_\lambda(w) \ := \ \int_{\mathcal{T}_\lambda} \mathrm{MMD}_{k_\sigma}^2 \Big((\Phi_{t,t_0}^w)_\# \rho_0^\lambda,\ \rho_t^\lambda \Big)\, \mathrm{d}\nu_\lambda(t),$$

and the empirical leaf-level risk

$$L_{\lambda,N}(w) \ := \ \int_{\mathcal{T}_\lambda} \mathrm{MMD}_{k_\sigma}^2 \Big((\Phi_{t,t_0}^w)_\# \rho_0^\lambda,\ \hat{\rho}_{t,N}^\lambda \Big)\, \mathrm{d}\nu_\lambda(t).$$

Assume:

1. Compactness and continuity: *Θ_w is compact and $w \mapsto (\Phi_{t,t_0}^w)_\# \rho_0^\lambda$ is continuous in MMD_{k_σ}, uniformly in $t \in \mathcal{T}_\lambda$. (Holds whenever the ODE flow is jointly continuous in (w, t, z) and the baseline density is continuous.)*
2. Kernel: *k_σ is the Gaussian kernel used in training, which is bounded, characteristic and metrises weak convergence on $\mathbb{R}^q$; boundedness and characteristicness alone would not give the last property.*
3. Sampling: *enrolment times $\{t_j\}$ are i.i.d. from a density f_T continuous and bounded below on $\mathcal{T}_\lambda$, and ρ_0^λ has a continuous Lebesgue density bounded below on an open set $V \subseteq \mathbb{R}^q$.*
4. Bandwidth: *$h_N \to 0$, $N_\lambda h_N \to \infty$.*

5. Quantitative FOA *(Eq. 6):* $w^\star(\lambda)$ *is a well-separated minimum of* L_λ *in its leaf-marginal form, per the non-implication remark above; this follows from leaf-marginal FOA with 1 and 2, as shown above, and is listed for reference.*
6. Reference consistency: $\int_{\mathcal{T}_\lambda} \mathrm{MMD}^2_{k_\sigma}(\hat{\rho}^\lambda_{t,N},\, \rho^\lambda_t)\,\mathrm{d}\nu_\lambda(t) \xrightarrow{\mathrm{P}} 0$ *(Lemma 4 below).*
7. Uniform convergence: $\sup_{w\in\Theta_w} |L_{\lambda,N}(w) - L_\lambda(w)| \xrightarrow{\mathrm{P}} 0$. *For this estimator it is not an additional assumption: because the source* $(\Phi^w_{t,t_0})_\#\rho^\lambda_0$ *is exact, it follows from 6 by the bound in the proof. It is listed because the trained objective, whose source is also empirical, would need it separately ("Uniform-convergence requirement" below).*
8. Leaf-form enrolment: $t_j \perp (\mathbf{z}^j_{t_0}, \mathbf{c}_j) \mid \lambda(j)$*, so that the observed leaf law at* t *is the leaf marginal (Remark after the proof of Theorem 1).*

Then the leaf-level estimator $\hat{w}_N(\lambda) \in \arg\min_{w\in\Theta_w} L_{\lambda,N}(w)$ *satisfies* $\hat{w}_N(\lambda) \xrightarrow{\mathrm{P}} w^\star(\lambda)$ *as* $N\to\infty$.

Proof. This follows from standard M-estimation arguments (van der Vaart, 1998, Theorem 5.7). By assumption (2), characteristic-kernel identifiability, and the leaf-marginal form of FOA (see the non-implication remark above), $L_\lambda(w) = 0 \iff w = w^\star(\lambda)$. Quantitative FOA (Eq. 6) ensures that $w^\star(\lambda)$ is a well-separated minimum. Reference consistency (assumption 6, Lemma 4) combined with the uniform continuity of $w \mapsto (\Phi^w_{t,t_0})_\#\rho^\lambda_0$ on the compact set Θ_w (assumption 1) yields uniform convergence of the empirical risk, $\sup_{w\in\Theta_w} |L_{\lambda,N}(w) - L_\lambda(w)| \xrightarrow{\mathrm{P}} 0$. Specifically, for any probability measures $P, Q, \hat{Q}$, $|\mathrm{MMD}^2(P,\hat{Q}) - \mathrm{MMD}^2(P,Q)| \le (\mathrm{MMD}(P,\hat{Q}) + \mathrm{MMD}(P,Q))\,\mathrm{MMD}(\hat{Q},Q) \le 4\sqrt{\|k_\sigma\|_\infty}\,\mathrm{MMD}(\hat{Q},Q)$, uniformly in P. With $P = (\Phi^w_{t,t_0})_\#\rho^\lambda_0$, integrating over t and applying Cauchy–Schwarz gives $\sup_w |L_{\lambda,N}(w) - L_\lambda(w)| \le 4\sqrt{\|k_\sigma\|_\infty}\,\big(\int \mathrm{MMD}^2(\hat{\rho}^\lambda_{t,N}, \rho^\lambda_t)\,\mathrm{d}\nu_\lambda\big)^{1/2} \xrightarrow{\mathrm{P}} 0$ by 6, with no equicontinuity argument needed. The argmin-consistency theorem then guarantees $\hat{w}_N(\lambda) \xrightarrow{\mathrm{P}} w^\star(\lambda)$. □

Four gaps between this analysis and the trained model. The reported runs do not sample-split: the router is trained jointly with w on the same data, so Proposition 10 characterises the population behaviour of the leaf-level estimator rather than the training procedure. The closest available check is routing accuracy along the training-set-size sweep of App. I.3, which rises from 0.42 at N=100 through 0.93 at N=1,000 to 0.97 at N=5,000 on BM1, consistent with $\Pr(\hat{\lambda}_N \neq \lambda) \to 0$.

A second gap concerns hard vs. soft routing. Proposition 10 and Lemma 3 treat the leaf assignment as a hard indicator, whereas the implemented router uses soft weights $\alpha_{j,k}$ at a finite $\tau_{\mathrm{end}} = 0.1 > 0$, so the reference measure $\mu^{\mathrm{ref},(k)}$ used in training is contaminated by other leaves at order $e^{-\Delta/\tau}$, and the softmax-bound defect bound $\bar{\delta} \le 2(K-1)e^{-\Delta/\tau}$ (App. B.11) does not vanish unless $\tau_N \to 0$. Consistency of the soft, jointly-trained estimator is open. A related gap concerns the routed parameter itself: $w_i = \sum_k \alpha_{ik} w_k$ is a convex combination, but $w \mapsto \Phi^w$ is not convex, so an interpolated parameter can in principle realise dynamics qualitatively unlike either archetype, for instance across a bifurcation of the parameterised family. We give no bound on the resulting trajectory error relative to hard assignment; the empirical mitigation is that the trained router is close to one-hot (routing entropy 0.222 of a maximum $\ln 3$ on BM1, Tab. 29), so genuinely interpolated parameters are rare rather than excluded.

A third gap concerns which pushforward the training objective actually uses. The theory above pushes ρ^λ_0 forward from the canonical time t_0; the training loss $\mathcal{L}_{\mathrm{cross}}$ instead integrates each unit from its own enrolment time t_i, so the predicted population at time t is the mixture $\int(\Phi_{t,s})_\#\rho_s\,\mathrm{d}\nu(s \mid s \le t)$ over enrolment times s rather than the canonical pushforward $(\Phi_{t,t_0})_\#\rho^\lambda_0$. The discrepancy is bounded by the drift difference. *Lemma (mixture vs. canonical pushforward).* Fix a leaf λ and suppose the deployed field $h(\cdot,t;\mathbf{c},w)$ is L-Lipschitz in the state uniformly over $t\in[t_0,T]$, $\mathbf{c}\in L_\lambda$ and w, and let $D(w) := \sup_{t,\mathbf{c},z} \|h(z,t,\mathbf{c};w) - h(z,t,\mathbf{c};w^\star(\lambda))\|$, where $w^\star(\lambda)$ is the true leaf parameter of Sec. 3. Under the generative model a leaf-λ unit enrolled at s carries state law $\rho_s = (\Phi^{w^\star}_{s,t_0})_\#\rho^\lambda_0$, and for every $t\in[t_0,T]$,

$$\mathrm{W}_1\Big(\int(\Phi^w_{t,s})_\#\rho_s\,\mathrm{d}\nu(s \mid s\le t),\ (\Phi^w_{t,t_0})_\#\rho^\lambda_0\Big) \ \le\ 2\,e^{L(t-t_0)}(t-t_0)\,D(w).$$

For $z\sim\rho^\lambda_0$ set $y = \Phi^{w^\star}_{s,t_0}(z)$. The mixture's state at t is $\Phi^w_{t,s}(y)$ and the canonical pushforward's is $\Phi^w_{t,t_0}(z)$. Splicing at s and applying Grönwall on $[s,t]$ (two fields from the common state y) and on $[t_0,t]$ (two fields from the common state z, the flow property supplying the $w^\star$-segment) bounds the pointwise gap by $2e^{L(t-t_0)}(t-t_0)D(w)$ uniformly in s, and the coupling z transfers the bound to W_1. The right-hand side vanishes at $w=w^\star$, so $w^\star$ minimises both the mixture objective the training loop evaluates and the canonical objective of Prop. 10, and the consistency analysis applies up to the displayed additive margin. Uniqueness of the mixture objective's minimiser is a further assumption: the bound degrades as $D(w)$ grows, so it does not exclude a second zero far from $w^\star$. The splice also uses the flow property, which the unaugmented field satisfies; we state the lemma for that field.

Lemma 4 (Integrated-risk consistency of the temporal-KDE reference)**.** *Under the sampling and bandwidth conditions of Proposition 10, with k_σ the Gaussian kernel and ρ_t^λ having a jointly continuous density $(t, z) \mapsto p^\lambda(z \mid t)$,*

$$\int_{\mathcal{T}_\lambda} \mathrm{MMD}^2_{k_\sigma}\left(\hat{\rho}^\lambda_{t,N},\, \rho^\lambda_t\right)\, \mathrm{d}\nu_\lambda(t) \xrightarrow{\mathrm{P}} 0.$$

Proof sketch. Pointwise: write $\int f \,\mathrm{d}\hat{\rho}^\lambda_{t,N} = A_N/B_N$ with $A_N = N_\lambda^{-1}\sum_j f(Z_j)\kappa_{h_N}(t_j - t)$ and $B_N = N_\lambda^{-1}\sum_j \kappa_{h_N}(t_j - t)$. The ratio is treated directly; it is not a ratio of expectations. $\mathbb{E}[A_N]$ equals $\int_{\mathcal{T}_\lambda}\int f(z)\, p^\lambda(z \mid s)\, f_T(s)\, \kappa_{h_N}(s-t)\,\mathrm{d}z\,\mathrm{d}s$, so $A_N \xrightarrow{\mathrm{P}} f_T(t)\int f \,\mathrm{d}\rho_t^\lambda$ as $h_N \to 0$ by the joint continuity of p^λ; likewise $B_N \xrightarrow{\mathrm{P}} f_T(t) > 0$. The variance of each is $\mathcal{O}(1/(N_\lambda h_N)) \to 0$. By the continuous mapping theorem, division being continuous at $B = f_T(t) > 0$, $A_N/B_N \xrightarrow{\mathrm{P}} \int f \,\mathrm{d}\rho_t^\lambda$, i.e. weak convergence in probability $\hat{\rho}^\lambda_{t,N} \Rightarrow \rho_t^\lambda$ for each $t \in \mathcal{T}_\lambda^\circ$. Because MMD with a bounded, continuous, integrable translation-invariant kernel such as the Gaussian metrizes weak convergence (Sriperumbudur et al., 2010) (bounded and characteristic alone is not enough: an almost-periodic kernel whose spectral support is dense in $\mathbb{R}$ is characteristic yet does not), and the kernel is bounded ($\mathrm{MMD}^2_{k_\sigma} \le 4\|k_\sigma\|_\infty$), we obtain pointwise convergence $\mathrm{MMD}^2_{k_\sigma}(\hat{\rho}^\lambda_{t,N}, \rho^\lambda_t) \xrightarrow{\mathrm{P}} 0$. Dominated convergence on the compact set $\mathcal{T}_\lambda$ then upgrades this pointwise convergence to integrated convergence. □

Corollary 2 (Dictionary and routing recovery)**.** *Suppose Proposition 10 applies for every leaf $\lambda \in \Lambda$, ensuring $\hat{w}_N(\lambda) \xrightarrow{\mathrm{P}} w^\star(\lambda)$ uniformly over Λ under a uniform lower bound on per-leaf occupancy N_λ (automatic for finite Λ; a continuum of zero-mass leaves such as BM3's is outside this statement, since exact-leaf occupancy is zero there and a finite-sample result would need smoothing or regression over context). Then $\mathbf{W}^\star$ and $\boldsymbol{\alpha}^\star(\lambda)$ are recovered up to column permutation, with the specific regime determined by the temperature schedule:*

- Anchor regime *(Lemma 2, $\tau_N \to 0^+$). If anchor leaves $\{\lambda_k\}_{k=1}^K$ exist such that $\boldsymbol{\alpha}^\star(\lambda_k) = \mathbf{e}_k$ and $\{w_k^\star\}$ are affinely independent, then $\mathbf{W}^{(N)} = [\hat{w}_N(\lambda_1), \ldots, \hat{w}_N(\lambda_K)] \xrightarrow{\mathrm{P}} \mathbf{W}^\star$ up to column permutation. Furthermore, softmax annealing $\tau_N \to 0^+$ drives the routing $\boldsymbol{\alpha}_i^{(N)} \xrightarrow{\mathrm{P}} \mathbf{e}_{\Pi(\kappa(i))}$ to one-hot leaf indicators (Proposition 9).*
- Simplex-interior regime *(Lemma 3, $\tau_N \to \tau_0 > 0$). If the barycentric coordinates of $\hat{w}(\Lambda)$ satisfy the uniform purity condition of Lemma 3 (Craig, 1994; Lin et al., 2015), then $\mathbf{W}^\star$ is recovered (up to permutation) as the vertex set of that simplex applied to the consistent estimates $\{\hat{w}_N(\lambda)\}$. This regime inherits Lemma 3's uniform purity hypothesis; nonempty relative interior, or meeting every facet, is not enough. The routing converges to the soft barycentric coordinates $\boldsymbol{\alpha}_i^{(N)} \xrightarrow{\mathrm{P}} \left[\begin{smallmatrix}\mathbf{W}^\star \\ \mathbf{1}^\top\end{smallmatrix}\right]^\dagger \left[\begin{smallmatrix}\hat{w}(\lambda(i)) \\ 1\end{smallmatrix}\right]$. In this regime, the temperature must be held bounded away from zero ($\tau_N \to \tau_0 > 0$) to appropriately preserve the interior coordinates.*

In either regime, the resulting estimator $(\mathbf{W}^{(N)}, \{\boldsymbol{\alpha}_i^{(N)}\})$ satisfies

$$\mathrm{dist}\left(\mathbf{W}^{(N)},\, \mathcal{W}_\epsilon^\star\right) \xrightarrow{\mathrm{P}} 0, \qquad \max_i \|\boldsymbol{\alpha}_i^{(N)} - \boldsymbol{\alpha}^\star(\lambda(i))\| \xrightarrow{\mathrm{P}} 0,$$

where $\mathcal{W}_\epsilon^\star$ denotes the set of dictionaries achieving the MRA ϵ-residual (Eq. 2). Convergence to a single $\mathbf{W}^\star$ holds when $\mathcal{W}_\epsilon^\star$ is a singleton (e.g., under exact representability $\epsilon = 0$, or under affine independence with anchors, or with the uniform purity condition of Lemma 3).

Proof sketch. Anchor regime. Under Proposition 10, $\hat{w}_N(\lambda_k) \xrightarrow{\mathrm{P}} w^\star(\lambda_k) = w_k^\star$ for each anchor leaf, meaning the columns directly provide $\mathbf{W}^{(N)} \to \mathbf{W}^\star$ up to permutation; the affine independence of $\{w_k^\star\}$ ensures this permutation is unique. The router converges to one-hot under $\tau_N \to 0$ by Proposition 9 applied to the population-optimal logits. *Simplex-interior regime.* The set $\{\hat{w}_N(\lambda) : \lambda \in \Lambda\}$ converges in probability (uniformly in λ) to $\{w^\star(\lambda) : \lambda \in \Lambda\} = \hat{w}(\Lambda)$, which lies in $\mathrm{conv}(\mathbf{W}^\star)$. By Lemma 3, under its uniform purity condition, the columns of $\mathbf{W}^\star$ are recovered as the unique minimum-volume simplex vertices. Since the affine pseudo-inverse $\left[\begin{smallmatrix}\mathbf{W}^\star \\ \mathbf{1}^\top\end{smallmatrix}\right]^\dagger$ is continuous, $\boldsymbol{\alpha}_i^{(N)} = \left[\begin{smallmatrix}\mathbf{W}^{(N)} \\ \mathbf{1}^\top\end{smallmatrix}\right]^\dagger \left[\begin{smallmatrix}\hat{w}_N(\lambda(i)) \\ 1\end{smallmatrix}\right] \xrightarrow{\mathrm{P}} \left[\begin{smallmatrix}\mathbf{W}^\star \\ \mathbf{1}^\top\end{smallmatrix}\right]^\dagger \left[\begin{smallmatrix}\hat{w}(\lambda(i)) \\ 1\end{smallmatrix}\right] = \boldsymbol{\alpha}^\star(\lambda(i))$. □

On the consistent estimation of the leaf assignment $\hat{\lambda}_N$. Proposition 10 treats $\hat{\lambda}_N$ as a structurally independent input, providing a clean separation from the empirical reference estimation. Under COA′ (Assumption 2), the leaf assignment is a deterministic measurable function $\lambda(i) = g^*(\mathbf{c}_i)$ of the static context. Therefore, any standard technique that consistently approximates g^*, such as a context classifier fitted on a held-out fold, satisfies $\Pr(\hat{\lambda}_N(i) \neq \lambda(i)) \to 0$, because the assignment depends on $\mathbf{c}_i$ alone and not on the temporal snapshots that build the reference measure. In strictly unsupervised settings, simple sample-splitting (estimating $\hat{\lambda}_N$ on one fold and evaluating $L_{\lambda,N}$ on another) ensures the independence required by the M-estimation framework.

Methodological considerations in finite-sample regimes. A fourth gap concerns the source side. The empirical risk $L_{\lambda,N}$ keeps the source measure ρ_0^λ exact and randomises only the reference, whereas Eq. 3 compares two empirical measures: its predicted marginal is itself a weighted sum of Dirac masses over the active units. Consistency of that estimator needs a joint uniform law of large numbers over the source and reference empirical processes, which the bound behind hypothesis (7) does not supply, since it uses an exact source. The Grönwall bound of the third gap has the same limitation in a different form: it controls the value of the objective, not the location of its minimiser, and bounding the shift of the minimiser would need a curvature condition at $w^\star(\lambda)$ that we do not verify.

The proofs above establish consistency of the idealised estimator: hard, sample-split leaf assignments and canonical-time pushforwards, in the joint limit $N \to \infty$, $h_N \to 0$. The four gaps above separate that estimator from the trained one. As with any theoretical bridge to real-world implementations, a few standard considerations apply to finite-sample regimes. *(i) Bandwidth schedule:* The proof uses a vanishing temporal bandwidth $h_N \to 0$; every run uses a single fixed $h = 0.1$ (Tab. 9), which introduces the usual finite-sample bias–variance tradeoff and does not shrink with N. So Lemma 4's asymptotic consistency is not claimed for the deployed kernel, whose reference is a fixed-bandwidth smoothed version of the leaf marginal. *(ii) Optimization dynamics:* As is customary in deep learning theory, Proposition 10 establishes statistical consistency for the global optimum of the objective space, abstracting away the non-convex optimization dynamics of gradient descent. *(iii) Finite-K parameterization:* The MRA residual ϵ constitutes a formal approximation bound, meaning the convergence target $\mathcal{W}_\epsilon^\star$ is broadly defined as an ϵ-neighborhood set rather than a single point unless exact representability holds.

Degeneracy of the cross-sectional objective. If the observed marginals were time-invariant over T_λ, every ρ_t-preserving field, the zero field included, would attain the same value of L_λ, and distribution matching alone would not pin the vector field. This is the degenerate case Proposition 4 excludes, since a stationary marginal means $\nabla \cdot (\rho_t^\lambda h) \equiv 0$. On the benchmarks the per-leaf baselines are concentrated at t_0 and disperse in phase across the observation window, so the marginals are strongly time-varying and the stationary-marginal degeneracy is absent. Time variation does not by itself make the minimiser unique: $h_w(\mathbf{z}, t) = v + wA(\mathbf{z} - vt)$ with A skew-symmetric carries $\mathcal{N}(\mathbf{0}, \mathbf{I})$ to $\mathcal{N}(vt, \mathbf{I})$ for every w, so uniqueness still rests on FOA in its quantitative form (Eq. 6). No barrier, penalty or auxiliary displacement term is used to avoid the degenerate solution in any reported run. Binning encoded latents by enrolment time into five windows and comparing per-leaf marginals, the between-bin MMD^2 exceeds a within-bin split-half floor by a factor of at least 119 wherever that floor is measurable, and on the remaining leaves the floor is statistically indistinguishable from zero. The marginals therefore carry substantial temporal signal on every benchmark. Per-benchmark, per-leaf values, with bootstrap intervals and the leaves on which the floor is not measurable, are in Tab. 47; the $119\times$ figure is the minimum over the leaves on which the ratio is defined.

Uniform-convergence requirement. The finite-sample statements that build on this analysis, namely Lemma 5 (finite-sample part) and the estimators above, consume uniform convergence of the empirical leaf risk,

$$\sup_{w \in \Theta_w} \left| L_{\lambda,N}(w) - L_\lambda(w) \right| \xrightarrow{\mathrm{P}} 0.$$

It needs the integrated reference consistency of Lemma 4, not only its pointwise form. For the idealised risk, whose source is exact, the bound in the proof of Proposition 10 then gives it from reference consistency alone. For the trained objective, whose predicted marginal is itself an empirical measure over the active units, transferring the control additionally needs a stochastic-equicontinuity (or metric-entropy) argument over the flow class, which we do not construct; that is the fourth gap above. It is the condition under which the finite-sample bounds hold.

B.10 Identification Under MRA Misspecification

This section closes the Step-4 gap of Theorem 1 for $\varepsilon > 0$: when the router's expert set cannot represent the true leaf parameter exactly, how far does the identified parameter move? Throughout, the per-leaf population risk is (Eq. 3 / Appendix B.9):

$$L_\lambda(w) := \int_{T_\lambda} \mathrm{MMD}^2_{k_\sigma}\left((\Phi^w_{t,t_0})_\# \rho_0^\lambda,\ \rho_t^\lambda \right) d\nu_\lambda(t), \qquad L_\lambda(w^*(\lambda)) = 0.$$

By MRA there exists $\bar{w} \in R$ (the router's reachable parameter set) with $\|\bar{w} - w^*(\lambda)\| \le \varepsilon$; this $\varepsilon = \varepsilon_{\mathrm{rep}} = \mathrm{dist}(w^*(\lambda), R)$, which is 0 for discrete leaves since $w^*(\lambda) = \mathbf{W} e_k \in R$.

(S1) Flow upper-Lipschitz in MMD (assumption). Like (S2), this is a modelling condition rather than a consequence of the model class; it follows from a bounded Lipschitz kernel together with local Lipschitz dependence of the flow on w over the compact window. There exists L_Φ with $\mathrm{MMD}_{k_\sigma}\left((\Phi^w_{t,t_0})_\# \rho_0^\lambda, (\Phi^{w'}_{t,t_0})_\# \rho_0^\lambda \right) \le L_\Phi \|w - w'\|$, uniformly in $t \in T_\lambda$ (T_λ compact).

(S2) Quantitative FOA (assumption). There exists ω_λ strictly increasing with $\omega_\lambda(0) = 0$ such that $L_\lambda(w) \geq \omega_\lambda(\|w - w^*(\lambda)\|)$, globally over R (or with $\hat{w}$ localised near w^*). Forms: $c_\lambda s^2$ (nondegenerate minimum) or $c_\lambda s$ (transversal). As discussed in Appendix B.9, this is an assumption, not a consequence of the model class. (S2) is a leaf-level condition; it does not follow from the per-context form of Assumption 5 (non-implication remark in App. B.9).

Lemma 5 (Stability under MRA misspecification)**.** (Population.) *If* $\hat{w} \in \arg\min_{w \in R} L_\lambda$*, then*

$$L_\lambda(\hat{w}) \leq L_\lambda(\bar{w}) \leq L_\Phi^2 \varepsilon^2 \quad \textit{and} \quad \|\hat{w} - w^*(\lambda)\| \leq \omega_\lambda^{-1}\big(L_\Phi^2 \varepsilon^2\big);$$

i.e. the parameter error is $O(\varepsilon)$ *under the quadratic modulus and* $O(\varepsilon^2)$ *under the linear modulus.*

Proof. $\hat{w}$ minimises over $R \ni \bar{w}$, so $L_\lambda(\hat{w}) \leq L_\lambda(\bar{w})$. (S1) gives $\mathrm{MMD}((\Phi^{\bar{w}})_\# \rho_0^\lambda, \rho_t^\lambda) \leq L_\Phi \varepsilon$ for all t, so $L_\lambda(\bar{w}) \leq L_\Phi^2 \varepsilon^2$ (ν_λ a probability). Apply (S2) and invert. □

Lemma 6 (Finite sample)**.** *With* $\sup_w |L_{\lambda,N} - L_\lambda| \leq r_N \to_{\mathrm{P}} 0$ *(the uniform-convergence requirement of Appendix B.9) and* $\hat{w}_N$ *an* o_N*-approximate minimiser:*

$$\|\hat{w}_N - w^*(\lambda)\| \leq \omega_\lambda^{-1}\big(L_\Phi^2 \varepsilon^2 + 2r_N + o_N\big).$$

Proof. $L_\lambda(\hat{w}_N) \leq L_{\lambda,N}(\hat{w}_N) + r_N \leq L_{\lambda,N}(\bar{w}) + o_N + r_N \leq L_\lambda(\bar{w}) + 2r_N + o_N \leq L_\Phi^2 \varepsilon^2 + 2r_N + o_N$; then (S2). □

Remark (well-specification is only in $\mathcal{H}$). Lemma 5 handles *R-misspecification*, where the router cannot represent $w^*(\lambda)$ exactly, and assumes the flow class $\mathcal{H}$ contains the truth ($L_\lambda(w^*(\lambda)) = 0$). Misspecification of $\mathcal{H}$ itself is out of scope.

Bounded propagation under routing corruption (why degradation is bounded). The corruption sweep of Appendix I.5 replaces a fraction η of units' contexts by random ones, perturbing the routing weights. Two structural facts bound the resulting damage:

- *Simplex containment:* every realised parameter $w_i = \mathbf{W}\boldsymbol{\alpha}_i$ lies in $\mathrm{conv}\{w_1, \ldots, w_K\}$, for *any* routing weights $\boldsymbol{\alpha}_i \in \Delta^{K-1}$, including corrupted ones. Hence each per-unit parameter remains within the convex hull of the true expert basis, no matter how misleading the context.
- *Bounded trajectory discrepancy over finite horizons:* if two parameters satisfy $\|h(\cdot; w) - h(\cdot; w')\| \leq L_h \|w - w'\|$ along the trajectories of interest (locally Lipschitz vector field on a compact latent set), then by Grönwall's inequality the resulting state and observation discrepancies over a horizon H are bounded by $L_h \,\mathrm{diam}(\mathrm{conv}\{w_k\})\, H\, e^{L_z H}$, where L_z is the Lipschitz constant of h in the state. The bound is independent of the corruption mechanism but exponential in the horizon, so it is loose on expanding dynamics such as BM5's chaotic subgroup.

Together these bound the trajectory discrepancy caused by any routing error over a finite horizon, whatever the corruption mechanism; the realised degradation is measured directly by the sweep of App. I.5.

B.11 Approximate-Anchor Basis Recovery and End-to-End Rate

This subsection gives the approximate-anchor analysis: the softmax defect constant, the feasible set of the recovered parameters declared explicitly, and the two effects combined into an end-to-end recovery rate. We keep the exact-anchor (Lemma 2) and simplex-interior (Lemma 3) lemmas as the two exact regimes; the analysis here covers the finite-temperature regime between them.

Anchor quality: softmax bound. Let $\boldsymbol{\alpha} = \mathrm{Softmax}(\mathbf{g}/\tau) \in \Delta^{K-1}$ with logits $\mathbf{g} \in \mathbb{R}^K$, mode k, and gap $\Delta := g_k - \max_{j \neq k} g_j > 0$. Then $\|\boldsymbol{\alpha} - \mathbf{e}_k\|_1 \leq 2(K-1)e^{-\Delta/\tau}$, in five steps:

1. *Off-mode mass:* for $j \neq k$, $\alpha_j = e^{g_j/\tau} / \sum_l e^{g_l/\tau} \leq e^{g_j/\tau} / e^{g_k/\tau} = e^{-(g_k - g_j)/\tau} \leq e^{-\Delta/\tau}$.
2. *Mode deficit:* $1 - \alpha_k = \sum_{j \neq k} \alpha_j \leq (K-1)e^{-\Delta/\tau}$.
3. ℓ^1 *distance (the factor of 2):* since $\boldsymbol{\alpha} \in \Delta^{K-1}$, $\|\boldsymbol{\alpha} - \mathbf{e}_k\|_1 = (1 - \alpha_k) + \sum_{j \neq k} \alpha_j = 2(1 - \alpha_k) \leq 2(K-1)e^{-\Delta/\tau}$.
4. *Matrix bound:* the columns of $A - I$ are $\boldsymbol{\alpha}(\lambda_k) - \mathbf{e}_k$, and $\|\cdot\|_2 \leq \|\cdot\|_1$, so with $\delta := \|A - I\|_2$, $\delta \leq \|A - I\|_F = \big(\sum_k \|\boldsymbol{\alpha}(\lambda_k) - \mathbf{e}_k\|_2^2\big)^{1/2} \leq \sqrt{K}\, \max_k \|\boldsymbol{\alpha}(\lambda_k) - \mathbf{e}_k\|_1 \leq 2\sqrt{K}(K-1)e^{-\Delta/\tau}$.
5. *Validity:* $\tau < \Delta / \ln(2\sqrt{K}(K-1))$ (for $K \geq 2$) is *sufficient* for $\delta < 1$, since the right-hand side bounds δ from above; $\delta \to 0$ as $\tau \downarrow 0$.

Basis condition. $\{w_k\}$ distinct and affinely independent (Lemma 2's condition).

Setup, with the feasible-set fork. Let K anchor leaves be given, $A := [\boldsymbol{\alpha}(\lambda_1), \dots, \boldsymbol{\alpha}(\lambda_K)]$, Hungarian-relabelled to the identity; $w^*(\lambda_k) = \mathbf{W}\mathbf{e}_k$ (discrete leaves, basis condition above). The residual $E := [\hat{w}(\lambda_k)] - \mathbf{W}A$ decomposes *column-wise*:

$$E_k = \hat{w}(\lambda_k) - \mathbf{W}\boldsymbol{\alpha}(\lambda_k) = \underbrace{[\hat{w}(\lambda_k) - w^*(\lambda_k)]}_{\text{estimation error}} + \underbrace{\mathbf{W}[\mathbf{e}_k - \boldsymbol{\alpha}(\lambda_k)]}_{\text{routing error}}.$$

Let $\delta_k := \|\boldsymbol{\alpha}(\lambda_k) - \mathbf{e}_k\|_2 \le \|\boldsymbol{\alpha}(\lambda_k) - \mathbf{e}_k\|_1$ and $\bar{\delta} := \max_k \delta_k$. The routing term is deterministic: $\|\mathbf{W}(\mathbf{e}_k - \boldsymbol{\alpha}(\lambda_k))\| \le \|\mathbf{W}\|_2 \delta_k \le \|\mathbf{W}\|_2 \bar{\delta}$. The estimation term depends on the feasible set of $\hat{w}$, *which must be declared*:

- **(a) Unconstrained** ($\hat{w} \in \arg\min_R L_\lambda$; per-leaf parameters fit freely). Then $\varepsilon_{\text{rep}} = 0$ ($w^* = \mathbf{W}\mathbf{e}_k \in R$), so the estimation term is $\omega^{-1}(2r_N + o_N) = O(\sqrt{r_N})$ (finite-sample only, via Lemma 6), and $b \le \sqrt{K}\left[\omega^{-1}(2r_N + o_N) + \|\mathbf{W}\|\bar{\delta}\right]$.
- **(b) Softmax-constrained** ($\hat{w} \in \arg\min_{R_\tau} L_\lambda$; parameters realised only through the finite-τ gate). Then $\mathbf{W}\boldsymbol{\alpha}(\lambda_k) \in R_\tau$ gives $L_\lambda(\hat{w}) \le L_\lambda(\mathbf{W}\boldsymbol{\alpha}(\lambda_k)) \le L_\Phi^2 \|\mathbf{W}\|^2 \delta_k^2$ (by S1), and S2, centred at w^* so that the minimiser may lie on the far side, gives $\|\hat{w} - w^*\| \le \omega^{-1}(L_\Phi^2 \|\mathbf{W}\|^2 \bar{\delta}^2)$. The two δ-terms add in the worst case: $b \le \sqrt{K}\left[\omega^{-1}(L_\Phi^2 \|\mathbf{W}\|^2 \bar{\delta}^2 + 2r_N + o_N) + \|\mathbf{W}\|\bar{\delta}\right]$.

Use the uniform modulus $\omega := \min_k \omega_{\lambda_k}$ across the K anchor leaves. Under the quadratic modulus $\omega(s) = cs^2$, using $\sqrt{x+y} \le \sqrt{x} + \sqrt{y}$, both branches give the *same end-to-end order*; (b) exceeds (a) by the factor $(L_\Phi/\sqrt{c} + 1)$ vs. 1 on the $\bar{\delta}$-term:

$$\begin{aligned} &\text{(a)} \quad b \le \sqrt{K}\left[\sqrt{(2r_N + o_N)/c} + \|\mathbf{W}\|\bar{\delta}\right], \\ &\text{(b)} \quad b \le \sqrt{K}\left[(L_\Phi/\sqrt{c} + 1)\|\mathbf{W}\|\bar{\delta} + \sqrt{(2r_N + o_N)/c}\right]. \end{aligned} \tag{7}$$

(Which branch is operative depends on the recovery procedure, a training-pipeline fact: free per-leaf fit versus gate-constrained realisation. Both are stated with the feasible set declared, and (b) is the default.)

Lemma 7 (Approximate-anchor recovery)**.** *Under the softmax bound and the basis condition, $\|A^{-1}\|_2 \le 1/(1-\delta)$, and $\widehat{\mathbf{W}} := [\hat{w}(\lambda_k)]A^{-1}$ satisfies $\|\widehat{\mathbf{W}} - \mathbf{W}\|_2 \le b/(1-\delta)$, where $b := \|E\|_2$ as bounded above. If A is estimated by $\widehat{A}$ with $\|\widehat{A} - A\|_2 \le \delta_A$ and $\delta + \delta_A < 1$, then $\|\widehat{\mathbf{W}} - \mathbf{W}\|_2 \le (b + \|\mathbf{W}\|_2 \delta_A)/(1 - \delta - \delta_A)$.*

Proof. Neumann series: $\|A^{-1}\| \le 1/(1-\delta)$. Moreover $[\hat{w}(\lambda_k)]A^{-1} = (\mathbf{W}A + E)A^{-1} = \mathbf{W} + EA^{-1}$, so $\|\widehat{\mathbf{W}} - \mathbf{W}\| \le \|E\|\,\|A^{-1}\| \le b/(1-\delta)$. With estimated $\widehat{A}$: $\widehat{\mathbf{W}} - \mathbf{W} = E\widehat{A}^{-1} + \mathbf{W}(A - \widehat{A})\widehat{A}^{-1}$; combine the two terms. □ □

Corollary 3 (End-to-end recovery, discrete leaves)**.** *Suppose the router's logit gap on the anchor contexts is bounded below, $\Delta > 0$; this is a property of the trained router, not guaranteed by the model class. With $\bar{\delta} \le 2(K-1)e^{-\Delta/\tau}$ and $\delta \le 2\sqrt{K}(K-1)e^{-\Delta/\tau}$ (softmax bound):*

$$\|\widehat{\mathbf{W}} - \mathbf{W}^*\|_2 \le b/(1-\delta) = O\big(K^{3/2}e^{-\Delta/\tau}\big) + O\big(\sqrt{Kr_N}\big) \;\longrightarrow\; 0 \quad \text{as } \tau \downarrow 0,\ N \to \infty.$$

The bound is operative only where the Neumann series converges, i.e. $\delta < 1$, for which $\tau < \Delta/\ln\big(2\sqrt{K}(K-1)\big)$ suffices (step 5 above). This end-to-end rate reflects two effects: the annealing defect, which carries $K^{3/2}$ once the $\sqrt{K}$ of $\|E\|_2 \le \sqrt{K}\max_k \|E_k\|$ multiplies $\bar{\delta}$'s $O(K)$; and the finite-sample term, which the same $\sqrt{K}$ multiplies to give $O(\sqrt{Kr_N})$ from the quadratic modulus. The order is identical across feasible-set branches (a)/(b). Recovery holds up to permutation without *exact one-hot anchors and* without *a nonempty simplex interior, the two regimes Corollary 2 covered, between which the runs at $\tau_{\text{end}} = 0.1$ sit.*

C Implementation Details

C.1 Full Training Pseudocode

Notes on Algorithm 1. (i) The router takes *only* the static context $\mathbf{c}_i$; the auxiliary-latent variant $g_\omega(\mathbf{c}_i, \mathbf{z}^i_{\text{aux}})$ appears only in the component ablations (App. I.3), in no reported benchmark run. (ii) The loss shown is $\mathcal{L}_{\text{cross}}$, which every reported run optimises ($n_{\text{followups}} = 0$, Tab. 8); $\widehat{\text{MMD}}^2$ is the unbiased U-statistic with $M = 128$, $n_q = 8$ (Tab. 9). The unbiased estimator needs at least two active units, and the active set $\{i \in \mathcal{B} : t_i \le t\}$ shrinks as $t \downarrow t_0$, so query times are drawn from $[t_0 + \delta_t,\, t_{\text{end}}]$ with δ_t one evaluation-grid step, and an expert with fewer than two active units at a drawn t contributes zero rather than an undefined term; the guard fires on under 0.3% of (query time, expert) pairs. Where follow-ups exist, $\mathcal{L}_{\text{cross}}$ is replaced by the semi-longitudinal latent MSE of Eq. 10, exercised only in App. I.6. (iii) Ground-truth leaf labels never enter training; Hungarian matching (App. C.4) resolves the label permutation at evaluation only.

Algorithm 1 CADENCE Training and Inference

Require: Dataset $\{(\mathbf{x}^i, t_i, \mathbf{c}_i)\}_{i=1}^N$; experts K; temperature schedule $\{\tau_e\}$; query times per step n_q; reference size M; temporal bandwidth h

Stage 1: Score-based spatial encoding (pre-train, then freeze):

Train score network $s_\psi(\mathbf{x}, t, \mathbf{c})$ via denoising score matching on VP-SDE:

$$\mathcal{L}_{\text{DSM}} = \mathbb{E}_{t,\mathbf{x}_0,\epsilon}\left[\left\| s_\psi(\mathbf{x}_t, t, \mathbf{c}) - \nabla_{\mathbf{x}_t} \log p_{t|0}(\mathbf{x}_t|\mathbf{x}_0)\right\|^2\right]$$

Define encoder $\text{Enc}_\psi : \mathbf{x} \mapsto \mathbf{z}$ via probability-flow ODE integration over the diffusion time τ from 0.05 (data) to 0.999 (prior) in 100 steps; decoding runs the integration in reverse (Tab. 8)

Cache latent enrolments: $\mathbf{z}^i \leftarrow \text{Enc}_\psi(\mathbf{x}^i; \mathbf{c}_i)$ *(one-time pass)*

Freeze ψ.

Stage 2: Temporal dynamics (train on cached latents):

for epoch $e = 1, 2, \ldots, E$ **do**

  for each mini-batch $\mathcal{B} \subset \{1, \ldots, N\}$ **do**

    Compute soft routing from static context: $\boldsymbol{\alpha}_i = \text{Softmax}(g_\omega(\mathbf{c}_i) \,/\, \tau_e)$

    Compute personalised dynamics weights: $\mathbf{w}_i = \mathbf{W}\boldsymbol{\alpha}_i \in \mathbb{R}^{d_w}$

    Sample n_q query times $t^{(1)}, \ldots, t^{(n_q)} \sim \mathcal{U}(t_0, t_{\text{end}})$

    for each query time t and each expert $k = 1, \ldots, K$ **do**

      Forward-integrate the active units $\{i \in \mathcal{B} : t_i \le t\}$: $\hat{\mathbf{z}}_t^i = \mathbf{z}^i + \int_{t_i}^t h_\theta(\mathbf{z}_\tau^i, \tau, \mathbf{c}_i; \mathbf{w}_i)\, d\tau$

      Predicted marginal: $\hat{\mu}_t^{(k)} \propto \sum_{i:t_i \le t} \alpha_{i,k}\, \delta_{\hat{\mathbf{z}}_t^i}$

      Reference marginal from M sampled units: $\mu_t^{\text{ref},(k)} \propto \sum_j \alpha_{j,k}\, \kappa_h(t_j - t)\, \delta_{\mathbf{z}_{t_j}^j}$

      Context-free reference (BM7, later days): replace $\alpha_{j,k}$ by the Sinkhorn coupling weights of Eq. 4, re-estimated each step, with masses $m_{k,t} \ge m_{\min}$; $\mathcal{L}_{\text{mass}}$ is zero where no reference is context-free

    end for

    Compute loss with σ set by the batch median heuristic:

$$\mathcal{L} = \underbrace{\tfrac{1}{n_q}\sum_{b=1}^{n_q}\sum_{k=1}^{K} \widehat{\text{MMD}}^2_{k_\sigma}\big(\hat{\mu}_{t^{(b)}}^{(k)},\, \mu_{t^{(b)}}^{\text{ref},(k)}\big)}_{\mathcal{L}_{\text{cross}}\ \text{(Eq. 3)}} + \lambda_{\text{router}}\left(\mathcal{L}_{\text{router}} + \mathcal{L}_{\text{mass}}\right) + \lambda_{\text{feat}}\,\mathcal{L}_{\text{feat}}$$

    Update $(\theta, \omega, \mathbf{W})$ via Adam

  end for

  Anneal: $\tau_e \leftarrow \tau_{\text{start}} + (\tau_{\text{end}} - \tau_{\text{start}}) \cdot \min(e/(f \cdot E),\ 1)$ $\{f{=}0.8$; cf. Prop. 9$\}$

end for

Inference (new individual with observation $\mathbf{x}^*$ at time t^*, context $\mathbf{c}^*$):

$\mathbf{z}^* \leftarrow \text{Enc}_\psi(\mathbf{x}^*; \mathbf{c}^*)$ *(encode)*

$\mathbf{w}^* \leftarrow \mathbf{W}\,\text{Softmax}(g_\omega(\mathbf{c}^*)/\tau_{\text{end}})$ *(route)*

$\hat{\mathbf{z}}_t^* = \mathbf{z}^* + \int_{t^*}^t h_\theta(\mathbf{z}_\tau^*, \tau, \mathbf{c}^*; \mathbf{w}^*)\, d\tau$ *(integrate)*

$\hat{\mathbf{x}}_t^* \leftarrow \text{Dec}_\psi(\hat{\mathbf{z}}_t^*; \mathbf{c}^*)$ *(decode)*

C.2 Auxiliary Loss Terms

The Stage-2 objective is $\mathcal{L} = \mathcal{L}_{\text{dyn}} + \lambda_{\text{router}}\mathcal{L}_{\text{router}} + \lambda_{\text{feat}}\mathcal{L}_{\text{feat}}$, with $\mathcal{L}_{\text{dyn}} = \mathcal{L}_{\text{cross}}$ (Eq. 3) in every reported run. **Feature alignment.** Where an auxiliary measurement y_i is recorded with the state at the same enrolment time t_i, a linear head π_ϕ predicts $\hat{y}_i = \pi_\phi(\mathbf{z}_{t_i}^i)$ and

$$\mathcal{L}_{\text{feat}} = \frac{1}{|\mathcal{B}_{\text{obs}}|}\sum_{i \in \mathcal{B}_{\text{obs}}} \big(\hat{y}_i - y_i\big)^2, \qquad \mathcal{B}_{\text{obs}} = \{i \in \mathcal{B} : y_i \text{ observed}\},$$

zero when $\mathcal{B}_{\text{obs}} = \emptyset$. Both $\hat{y}_i$ and y_i are at t_i, so the term uses only the snapshot the model already receives, with no follow-up or temporally paired data. It is active on BM2 alone, where a noisy serology readout $y_i \approx R(t_i)$ is a second, independent reading of the recovered fraction R, which the observation also contains (noise $\sigma = 0.05$, App. F.3); elsewhere the mask is empty and the term contributes nothing despite the nonzero weight of Tab. 8.

Router regularisation. With $H(\cdot)$ the Shannon entropy on the simplex,

$$\mathcal{L}_{\text{router}} = \frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}} H(\boldsymbol{\alpha}_i) \;-\; \beta\, H\Big(\tfrac{1}{|\mathcal{B}|}\textstyle\sum_{i\in\mathcal{B}}\boldsymbol{\alpha}_i\Big);$$

the first term sharpens individual assignments, the second rewards balanced usage against expert collapse, and no ground-truth label enters either. It is used only on BM7 ($\lambda_{\text{router}} = 2.0$, zero elsewhere) with $\beta = 0$ in every reported run, so only the sharpening term is active; a run with $\beta = 0.5$ changes per-cell fate accuracy by one cell and AUROC by under 0.01. **Coupling-mass guard.** On BM7 the same weight multiplies the guard $\mathcal{L}_{\text{mass}}$ of Eq. 4 (App. B.3), which acts on the expert masses of the later-day reference coupling, not on the routing weights: the objective there is $\mathcal{L}^{\gamma}_{\text{cross}} + \lambda_{\text{router}}(\mathcal{L}_{\text{router}} + \mathcal{L}_{\text{mass}})$; elsewhere no reference is context-free and $\mathcal{L}_{\text{mass}}$ does not exist. Without it (sharpening alone at $\lambda_{\text{router}} = 2.0$) one expert absorbs 81% of the later-day mass and per-cell accuracy falls to 0.570 (AUROC 0.66), about the label-free OT fate map; App. H.5 varies its weight alone. **Objective actually optimised:** $\mathcal{L}_{\text{cross}}$ alone on BM1 and BM3–BM6 (empty mask, $\lambda_{\text{router}} = 0$); BM2 adds the feature term and BM7 the router entropy and mass guard under one weight. No barrier, warm-up, augmentation or label-based term is used in any reported run.

C.3 Structural Inductive Biases

Three structural priors are implemented in the code and active in no reported run (they act on a VAE-encoder variant that produces no result here): (i) *monotonicity*, a ζ constraint $\dot{z}_j \geq 0$ for inherently monotone quantities such as BM2's recovered fraction $R(t)$, which would stop the model conflating ensemble heterogeneity with disease progression; (ii) *simplex conservation*, projecting the ODE output onto the tangent space of the simplex for compartmental models ($S + I + R = 1$); (iii) *kinetic regularisation*, $\lambda_{\text{kin}}\|h_\theta\|^2$, discouraging high-frequency spurious dynamics.

C.4 Expert-Index Alignment for Evaluation, and an Unused Warm-Start Option

Under the anchor condition of Lemma 2 or the uniform purity condition of Lemma 3, $\mathbf{W}$ and the routing are identifiable up to column permutation (Cor. 3 covers finite temperatures between them), and the network's indexing of the K experts is arbitrary, so evaluation aligns expert indices with ground-truth subgroups before scoring. In all reported benchmarks the router is trained through the dynamics objective alone; labels enter only here. We run the trained router over the *validation* set to build a confusion matrix $C \in \mathbb{R}^{K\times K}$ ($C_{i,j}$: held-out units of true subgroup i whose $\arg\max_k \alpha$ is expert j) and solve the linear sum assignment problem

$$P^* = \arg\max_{P\in\mathcal{S}_K} \sum_{i=1}^{K} C_{i,P(i)} \tag{8}$$

over the permutations $\mathcal{S}_K$ by the Hungarian algorithm. P^* is applied unchanged to the test split and to every unsupervised comparator, and before scoring the recovery experiments of Apps. I.8 and I.9. Fitting it on the test confusion matrix, the common practice, would make the null not $1/K$ but the expected aligned accuracy of a label-independent router, $\mathbb{E}[\max_P \sum_i C_{i,P(i)}]/N$: for $K{=}2$, $0.5 + \mathbb{E}|X - N/2|/N$ with $X \sim \text{Bin}(N, 1/2)$, i.e. 0.510 on BM7's $1{,}488$ cells and 0.513 on its 986 clones, and 0.355–0.357 at $N = 750$–900 for the $K{=}3$ benchmarks. The selected permutation agrees under both fits on every benchmark and method, so every reported accuracy is unchanged; fitting on validation keeps the null at $1/K$ with no test label in an alignment. The same applies to the orientation of an unsupervised AUROC; supervised classifiers need no alignment. **Optional use of partial labels.** With partial subgroup labels the router may be pretrained by a brief warm-up with a permutation-invariant cross-entropy (exhaustive search over the $K!$ permutations per batch, then a fixed global Hungarian alignment refreshed periodically) and frozen for Stage 2. *No reported benchmark uses this option.*

D Extensions

D.1 Semi-longitudinal data

If unit i is observed at $n_i > 1$ times $t_i^{(1)} < \cdots < t_i^{(n_i)}$, the *conditional flow map* variant is

$$\mathbf{z}^i_{t+\Delta t} \mid \mathbf{z}^i_t = \mathbf{z}^i_t + \int_t^{t+\Delta t} h_\theta(\mathbf{z}^i_\tau, \tau, \mathbf{c}_i; \mathbf{w}_i)\, d\tau, \tag{9}$$

with $\mathbf{z}^i_t$ encoded from $\mathbf{x}^i_t$; the longitudinal observations supervise the trajectory directly across time steps and substantially reduce identifiability concerns. A multi-modal observation factors the decoder as $d_\psi(\mathbf{z}) = (d^{(1)}_\psi(\mathbf{z}), \ldots, d^{(M)}_\psi(\mathbf{z}))$, summing the reconstruction loss over modalities with modality-specific variances σ^2_m.

D.2 Continuous parameter variation beyond MRA

If the finite-K approximation of MRA is insufficient (error ϵ too large for a given K), increase K, replace the SMoE factorisation by a hypernetwork $\mathbf{w}_i = \text{HyperNet}_\omega(\mathbf{c}_i)$ projected onto a bounded set and trained with weight decay for stability, or use a continuous basis expansion.

D.3 Continuous Context-to-Parameter Router

BM3's reported run (Tab. 2) replaces the SMoE router by a direct context-to-parameter network (`CAPRouter`), since BM3's leaf family is continuous. The map $\mathbf{c} \mapsto \mathbf{w}$ is an MLP with two SiLU hidden layers of width $d_{h,\text{router}} = 32$ (Tab. 8) and a linear output layer of width $d_w = 14$ with a $\tanh$ output, so $\mathbf{w}_i \in (-1, 1)^{d_w}$; this bounds the continuous parameter image in place of MRA's compactness argument. Weights use PyTorch's default `nn.Linear` initialisation (Kaiming-uniform), and unlike the SMoE router the output layer is not zero-initialised. It trains with the Stage-2 optimiser and weight decay (10^{-5}, Tab. 9), with no router-specific weight decay. As in the discrete case $\mathbf{w}_i$ enters the shared backbone by FiLM, $h_\theta(\mathbf{z}, t, \mathbf{c}; \mathbf{w}_i) = s(\mathbf{w}_i) \odot h_{\text{base}}(\mathbf{z}, t, \mathbf{c}) + b(\mathbf{w}_i)$; nothing else differs.

D.4 Designing Static Context on New Datasets

Figure 1 groups static context into *baseline features* (demographic or design covariates at enrolment), *risk scores* (composite clinical or behavioural scores), *individual statistics* (per-unit summaries of prior measurements, e.g. a historical mean) and *system signatures* (global descriptors of the ensemble or environment). Any time-invariant covariate plausibly informative about the unit's dynamical regime is admissible; what the framework requires is COA′ (Assumption 2), a regime assignment that is a deterministic function of the context. A covariate correlated with the regime but not determining it degrades gracefully, while an uninformative one degrades forecasts below the unrouted model (App. I.5), so context should be validated before use. In practice: (i) prefer covariates with a causal or mechanistic link to the dynamics (e.g. a genotype or treatment arm known to alter progression) over incidental correlates; (ii) include enough dimensions to separate the regimes, since conditioning strength, not mere presence, drives recovery (Apps. I.8, I.5); (iii) where a subset of units lacks informative context, those units route at chance while the rest keep clean accuracy and overall degradation stays bounded (App. I.5); (iv) validate informativeness a posteriori through routing accuracy against any available ground truth, reading low accuracy as a sign that the context carries no regime information.

E Extended Related Work

Trajectory inference from cross-sectional data. Pseudo-time methods order single-cell profiles along a developmental tree (Trapnell et al., 2014); recovering continuous dynamics from snapshots was formalised by Hashimoto et al. (2016) and cast as optimal transport by Schiebinger et al. (2019), with continuous normalising flow variants in TrajectoryNet (Tong et al., 2020) and MIOFlow (Huguet et al., 2022). Where lineage information exists, LineageOT (Forrow & Schiebinger, 2021) couples snapshots through the lineage tree, moscot (Klein et al., 2025) scales the family to atlas-sized problems, and CellOT (Bunne et al., 2023) learns conditional maps for perturbation response. Uniqueness conditions for OT-regularised path inversion are due to Lavenant et al. (2021); related objectives include Action Matching (Neklyudov et al., 2023) and Schrödinger bridges (De Bortoli et al., 2021). All of these produce population-level couplings rather than per-unit forecasts.

Clinically, linear mixed-effects models (Laird & Ware, 1982) split longitudinal data into a population mean and Gaussian individual effects. CADENCE inherits that decomposition (a shared regime parameter $w(\lambda)$, a per-unit latent state z_0) but infers the random effect from a *single* snapshot, and its structural assumptions concern identification of $w(\lambda)$. Event-based progression models (Fonteijn et al., 2012) order disease stages non-temporally and are outside our scope.

Neural ODEs and continuous latent models. Neural ODEs (Chen et al., 2018) and their augmented variant (Dupont et al., 2019) parameterise hidden-state dynamics directly; Latent ODEs (Rubanova et al., 2019) encode observations into an initial state evolved in continuous time, the paradigm our Stage 2 inherits. Extensions handle irregular arrivals (De Brouwer et al., 2019), informative missingness (Kidger et al., 2020), long trajectories (Iakovlev et al., 2023) and meta-learned ODE priors (Mauel et al., 2026). All need a high per-unit sampling rate to constrain the initial condition; in CADENCE, DFA and FOA play the role of follow-up visits.

Context-conditioned dynamics. LEADS (Yin et al., 2021) separates local from global dynamics; CoDA (Kirchmeyer et al., 2022) conditions ODE parameters on a context vector inferred from reference trajectories; related schemes use Taylor-expanded modulation (Nzoyem et al., 2024), meta-learning (Wang et al., 2022), latent context variables (Norcliffe et al., 2021), modulation networks (Auzina et al., 2023), nuisance-invariant latents (Lagemann et al., 2023), static-dynamic factorisation (Chapfuwa et al., 2022) and shared-plus-adjustment parametrisations (Vermani et al., 2024). All infer the context from observations over time; in CADENCE it is an *observed baseline covariate*, so under COA′ routing is learned at ensemble level without per-unit longitudinal data. Conditional transport (CVFM (Generale et al., 2026), COT-FM (Kerrigan et al., 2024), cMonge (Driessen et al., 2026)) learns $P(x_t \mid c, t)$ in observation space without a per-unit latent state, so its individual predictions are draws from the conditional law (Sec. 8); BrainODE (Park et al., 2026) is an application-level relative in clinical imaging.

Mixture of experts and basis identifiability. MoE architectures (Jacobs et al., 1991; Shazeer et al., 2017) route across specialised subnetworks; Soft MoE (Puigcerver et al., 2024) replaces hard dispatch with differentiable slot-based routing. Our experts are dynamical parameter vectors rather than networks, with dense rather than top-k gating whose annealing recovers hard DFA assignments (Prop. 9). Generating parameters from a shared basis relates to HyperNetworks (Ha et al., 2016) and LoRA (Hu et al., 2022); under anchor or interior conditions the basis is identifiable up to column permutation (Lemmas 2–3).

Score-based encoders. Song et al. (2021) established the PF-ODE as a deterministic bijection between data and a Gaussian prior, making VP-SDEs usable as encoders (our Stage 1); downstream uses include LSGM (Vahdat et al., 2021) and DiffAE (Preechakul et al., 2022), with Flow Matching (Lipman et al., 2023) and Stochastic Interpolants (Albergo et al., 2025) as simulation-free alternatives. Proposition 2 shows that pushing the data bijectively to $\mathcal{N}(\mathbf{0}, \mathbf{I})$ restricts the diffeomorphic ambiguity to the $\mathcal{N}(\mathbf{0}, \mathbf{I})$-preserving subgroup G.

Identifiability. Latent variable models are unidentifiable without structure (Locatello et al., 2019), a gap closed by conditioning the prior on auxiliary variables (Hyvärinen et al., 2019; Khemakhem et al., 2020), which is structurally what conditioning the router on static context does here. Temporal extensions include Klindt et al. (2020), Yao et al. (2021) and Hälvä et al. (2021); Yao et al. (2024) recast parameter estimation as latent-variable identification; classical system identification (Ljung, 1999) has been extended to switching systems (Balsells-Rodas et al., 2024), transition functions (Hızlı et al., 2024) and causal disentanglement (Baumgartner et al., 2026). These assume full time-series access; our analysis addresses the gap under PRW and FOA: Theorem 1 identifies the per-leaf parameters within a fixed frame, Lemma 5 bounds the error under MRA misspecification, and Theorem 2 characterises the remaining ambiguity as a quotient.

Variational and neural-process methods. The two-stage architecture reads as a structured variational objective, like Kalman VAEs (Fraccaro et al., 2017). Neural Processes (Garnelo et al., 2018b;a) amortise inference over functions; CADENCE shares that philosophy but commits to explicit ODE dynamics (PRW), the inductive bias needed for identifiable extrapolation.

Control theory and geometry. COA′ is analogous to the observability rank condition of Hermann & Krener (1977) and Isidori (1985): static context separates regimes as outputs separate internal states. DFA rests on the Rokhlin disintegration theorem, whose regular conditional distributions for measurable partitions underpin pooling within leaves. MRA posits a compact parameter image; App. B.5 gives $K = |\Lambda|$ for discrete leaves and an ϵ-net bound for continuous regimes.

Applications. Disease progression models (Oxtoby et al., 2018) increasingly use deep latent dynamics for neurodegeneration (Wang et al., 2025). In genomics, RNA velocity (La Manno et al., 2018) supplies derivative estimates that cross-sectional models must otherwise infer. Foundation models for differential equations (Mauel et al., 2026) suggest that context-conditioned routing may amortise across domains.

F Extended Benchmark Description

Real cross-sectional ensembles provide no ground-truth individual trajectories, so whether a model has recovered the correct dynamics cannot be assessed on them directly. We therefore use seven benchmarks of increasing realism: five synthetic benchmarks with fully controlled data-generating processes (BM1–BM5), one semi-synthetic benchmark driven by a biologically grounded stochastic gene-circuit simulator (BM6), and one primary real-world benchmark of lineage-traced single-cell data (BM7). On the synthetic benchmarks the true vector field, subgroup assignments, canonical baseline states and complete individual trajectories are known, so we can measure (i) routing accuracy against ground-truth subgroups and (ii) forecasting error at time points never seen in training. Each benchmark isolates a distinct modelling assumption or failure mode, across several scientific domains (Tab. 7). The SERGIO simulator is distributed under GPL-3.0; the official LARRY repository provides no explicit software license. The LARRY files are the in-vitro haematopoiesis matrices and metadata of `AllonKleinLab/paper-data`.

The seven advance along three axes: observation dimensionality, dynamical complexity and the severity of the identifiability challenge. BM1 verifies core ODE fitting in isolation; BM2 adds an auxiliary serology readout recorded at enrolment; BM3 evaluates continuous rather than discrete routing; BM4 demands stable high-dimensional encoding, a precursor to multi-omics ensembles; BM5 tests the method near the limit of the flow-observability condition; BM6 and BM7 test state-of-the-art single-cell simulations and primary living-cell measurements.

F.1 Problem Setup and Data Generation Protocol

Trajectory simulation. Each simulated benchmark (BM1–BM6) defines a ground-truth system with K subgroups (BM7 is real data, App. H.5). Unit i is assigned to a leaf $\lambda_i \in \Lambda$ deterministically via $g^*(\mathbf{c}_i^{\text{static}})$ on BM1–BM4, and on BM5–BM6 by drawing the leaf first and sampling the context conditionally, formalising

Table 7: Benchmark suite. BM1–BM6 are simulated with known ground truth; BM7 is primary living-cell data. p: observation dimension; K: subgroups; r_s/r_d: static/generated dynamic context dimensions (r_d is not supplied to any model). Splits are 70/15/15 by unit (by clone on BM7), with generator seeds 0–4; Stage 1 is fitted on training data only, including fold-specific gene selection and PCA-50 context fitting on BM7. On BM5–BM7, COA′ holds only in the Bayes-classifier sense because context supports overlap; routing accuracy is 84.3% (BM5), 85.0% (BM6), and 64.5% per cell / 69.0% per clone (BM7) (App. I.5). BM3 has a continuous leaf family and uses the continuous router of App. D.3, so K is n/a.

ID	System	p	K	r_s/r_d	COA′ reading	Window	N (train / val / test)	Key challenge
BM1	Lotka–Volterra	2	3	4/2	(a)	sys. time [20, 80]	5,000 (3,500 / 750 / 750)	SMoE routing; exact distinguishability
BM2	SIR epidemic	3	3	4/2	(a)	days [0, 80]	5,000 (3,500 / 750 / 750)	auxiliary serology readout at enrolment
BM3	Van der Pol	2	n/a	1/0	(a)	phys. time [0, 20]	4,000 (2,800 / 600 / 600)	continuous MRA approximation
BM4	Gene expression	500	3	20/5	(a)	time [40, 80]	6,000 (4,200 / 900 / 900)	high-dim. PF-ODE encoding; linear subspace recovery
BM5	Duffing	2	3	2/0	Bayes relaxation of (a), overlapping context support	phys. time [0, 50]	5,000 (3,500 / 750 / 750)	identifiability limit; chaotic dynamics
BM6	GRN (SERGIO)	100	2	2/0	Bayes relaxation of (a), overlapping context support	pseudotime [0, 1]	4,000 (2,800 / 600 / 600)	stochastic SDE biology; emergent bifurcations
BM7	LARRY lineage	200	2	50/0	Bayes relaxation of (a), fixed projection of the day-2 observation; partially informative	culture day (native units)	1,488 day-2 cells, 5-fold cross-fit by clone, each held out once	primary in-vitro data; unsupervised fate discovery

COA′ (Assumption 2). Within-subgroup variation enters through a random canonical baseline state $z_0^i \in \mathbb{R}^q$ at a fixed canonical time t_0, integrated forward by a high-order Runge–Kutta solver (Dormand–Prince, relative tolerance 10^{-8}) to the full latent trajectory $\{z_t^i\}_{t\in[t_0,t_{\text{end}}]}$; on high-dimensional benchmarks (e.g. BM4) observations pass through a fixed random mapping d_ψ (linear or nonlinear by benchmark) simulating the unknown measurement process. **Heterogeneous enrollment.** The model never sees a full trajectory: each enrollment time t_i^{enroll} is drawn independently from $\mathcal{T}$ on $[t_0, t_{\text{end}}]$ (Uniform(t_0, t_{end}) on BM1–BM3, BM5 and BM6, whose windows start at t_0; Uniform$(40, 50)$ on BM4), mimicking units observed at arbitrary stages, and the model receives one snapshot $\mathbf{x}^i = \text{obs}(\mathbf{z}^i_{t_i^{\text{enroll}}}) + \varepsilon$ with its enrollment time and static context. Enc_ψ encodes it into $\mathbf{z}^i_{t_i^{\text{enroll}}}$ and h_θ must carry the dynamics over the unit-specific interval $[t_i^{\text{enroll}}, t_i^{\text{end}}]$. **Static and dynamic context.** *Static context* $\mathbf{c}_i^{\text{static}} \in \mathbb{R}^{r_s}$ holds time-invariant attributes (e.g. genetics, birth cohort) and is the sole router input; *dynamic context* $\mathbf{c}_i^t \in \mathbb{R}^{r_d}$ holds evolving attributes (e.g. comorbidities, lifestyle), treated at enrollment as auxiliary conditioning and never routed, preserving COA′'s time-invariance, as in longitudinal ensemble studies with static baselines and evolving measurements. **What the model actually receives:** in every reported run, for CADENCE and every baseline, only $(\mathbf{x}^i_{t_i}, t_i, \mathbf{c}_i^{\text{static}})$. Dynamic context is defined for completeness of the data-generating process and **is supplied to no model in any result in this paper**; in particular BM4's $\mathbf{c}_i^t = \mathbf{z}_t^i + \varepsilon_t$ is never used, nor are BM1's subgroup-specific OU signals, which would give a second route to regime identity beside the static context that the corruption sweep of App. I.5 severs. **Training and evaluation.** Stage 1 trains the PF-ODE Enc_ψ by denoising score matching to map the observations bijectively to $\mathcal{N}(\mathbf{0}, \mathbf{I})$; it is then frozen and observations are encoded offline. Stage 2 trains the router and h_θ in latent space, by $\mathcal{L}_{\text{cross}}$ in the cross-sectional default, enforcing distributional consistency across time without matched temporal pairs. Because the subgroup dynamics are shared, the heterogeneous enrolment snapshots jointly constrain the field across the window, and forecasting beyond t_i^{enroll} is a strict *out-of-distribution temporal generalisation* test; the full trajectory is used only at evaluation. Metrics: (1) *forecasting error*, MAE between the decoded prediction $\text{Dec}_\psi(\hat{\mathbf{z}}_t^i)$ and the true observation $\mathbf{x}_t^i$ in decoded $\mathbf{x}$-space over $[t_i^{\text{enroll}}, t_{\text{end}}]$; (2) *routing accuracy*, the fraction of units whose $\arg\max_k \alpha_{ik}$ is their ground-truth subgroup; (3) *population distance*, SW_2 between predicted and true state distributions at future times, capturing mode collapse or drift.

F.2 BM1: Lotka–Volterra Predator–Prey (Nonlinear Oscillatory Dynamics)

Scientific motivation. The Lotka–Volterra (LV) system is a canonical nonlinear oscillator whose qualitative structure is representative of a broad class of two-species interaction models arising in ecology (predator–prey cycles), biochemistry (autocatalytic reactions), and population competition dynamics. We define $K = 3$ dynamical archetypes distinguished by their growth and interaction parameters: a *balanced oscillation* regime with stable periodic orbits, a *high-predation* regime with elevated prey consumption rate and reduced predator mortality, and an *oscillatory collapse* regime characterised by rapid predator die-off following initial growth. Because the phase portrait of each subgroup is geometrically distinct, routing errors and ODE fitting inaccuracies are immediately transparent in phase space, making BM1 ideal for isolating the core dynamical learning problem under conditions where distinguishability (Theorem 1) holds by construction.

Quantities and roles.

- **Global time** $t_i \in [20, 80]$**:** System observation time; canonical reference $t_0 = 20$.

- **Static context $\mathbf{c}_i^{\text{static}} \in \mathbb{R}^4$:** Structured real-valued covariates (e.g., resource availability or environmental quality proxies) gating to $K = 3$ dynamical archetypes via threshold rules on two informative dimensions, set at the terciles of the context distribution so that the three leaves are equally prevalent (250 test units each).
- **Dynamic context $\mathbf{c}_i^t \in \mathbb{R}^2$:** Two time-varying auxiliary signals modeled as independent Ornstein–Uhlenbeck processes with subgroup-specific equilibrium levels, capturing slowly drifting environmental conditions.
- **Observation $\mathbf{x}_{t_i}^i \in \mathbb{R}^2$:** Raw ODE state (u, v) with additive Gaussian noise ($\sigma = 0.05$).
- **Challenge:** Verifying SMoE routing and Neural ODE integration in a 2D oscillatory system where the three subgroup phase portraits are geometrically distinct.

F.3 BM2: SIR Epidemic ensemble (Epidemiology)

Scientific motivation. Cross-sectional serological surveys represent a ubiquitous form of epidemiological data. BM2 directly instantiates this setting, requiring the recovery of subgroup-specific transmission/recovery dynamics from ensemble snapshots. It is also the one benchmark where an auxiliary measurement is available at enrolment, so it tests how a single scalar readout sharpens the recovered dynamics.

Quantities and roles.

- **Global time $t_i \in [0, 80]$:** Days since epidemic onset ($t_0 = 0$).
- **Static context $\mathbf{c}_i^{\text{static}} \in \mathbb{R}^4$:** Age, comorbidity index, vaccination status, and BMI z-score, aggregated into a continuous risk score that routes to vulnerability archetypes at empirical thresholds. Two of the four simulated archetypes are statistically indistinguishable in their dynamics and are merged, giving $K = 3$ effective subgroups.
- **Dynamic context $\mathbf{c}_i^t \in \mathbb{R}^2$:** Binary medication activation indicator ($\mathbf{1}[I(t) > 0.05]$ modulated by subgroup-specific uptake probability) and a noisy symptom-severity signal ($I(t) + \varepsilon$), both computed from the realized epidemic trajectory.
- **Observation $\mathbf{x}_{t_i}^i \in \mathbb{R}^3$:** All three epidemic compartments (S, I, R) are observed, with additive Gaussian noise ($\sigma = 0.05$). The conservation constraint $S + I + R = 1$ holds in the data-generating simulation, so the third compartment is determined and the observation is lossless with respect to the three-variable state.
- **Supervision:** A scalar serology readout $y_i \approx R(t_i^{\text{enroll}})$ supervises the recovery dimension via $\mathcal{L}_{\text{feat}}$; it is the only BM2-specific term in the objective. Removing it ($\lambda_{\text{feat}} = 0$, three seeds) raises MAE from 0.025 to 0.036 ± 0.004 and SW_2 from 0.050 to 0.077 ± 0.010, behind the strongest cross-sectional baseline (OT-CFM, $0.021/0.055$, Tab. 2), so the readout contributes to the BM2 margin as well as to absolute accuracy. The baselines receive the scalar as an input and CADENCE as a training target, so access is matched in data but not in use. Giving OT-CFM and CVFM the same auxiliary regression head, trained against the scalar as CADENCE's is, gives MAE/SW_2 of $0.021/0.053$ and $0.030/0.052$ (five seeds): CADENCE's SW_2 of 0.050 stays lowest but is no longer separated (uncorrected $p = 0.09$), so its BM2 SW_2 lead is largely a use-of-readout effect.
- **Challenge:** Recovering heterogeneous (β_k, γ_k) transmission–recovery pairs from cross-sectional snapshots with partial supervision from a scalar serology signal.

F.4 BM3: Van der Pol Oscillator (Neuroscience / Engineering)

Scientific motivation. Unlike BM1 and BM2 which evaluate discrete subgroup assignments (DFA), BM3 is designed to stress-test the *continuous* Mixture of Experts formulation. The damping parameter μ varies continuously as a function of individual context. There are no discrete subgroups; the SMoE with K experts must approximate a genuinely continuous one-dimensional family of dynamics.

Quantities and roles.

- **Global time $t_i \in [0, 20]$:** Physical time; canonical reference $t_0 = 0$.
- **Static context $\mathbf{c}_i^{\text{static}} \in \mathbb{R}^1$:** Scalar damping parameter $\mu_i \sim \text{Uniform}(0.5, 3.0)$, serving simultaneously as subgroup identity and as the router's sole input.
- **Dynamic context:** None ($r_d = 0$).
- **Observation $\mathbf{x}_{t_i}^i \in \mathbb{R}^2$:** State (x, y) with additive Gaussian noise ($\sigma = 0.05$).
- **Challenge:** Disentangling *which* limit cycle a realization occupies (captured by the router) from *where* on the limit cycle it currently resides (captured by the latent state $\mathbf{z}_{t_i}^i$). The reported BM3 run uses the continuous context-to-parameter router (App. D.3); App. I.9 evaluates how well $K \in \{2, 4, 8\}$ discrete experts tile this continuous one-dimensional family.

F.5 BM4: Gene Expression Trajectories (Bioinformatics)

Scientific motivation. Large genomic ensembles (e.g., UK Biobank, TCGA) typically capture high-dimensional molecular readouts at single (or very few) cross-sectional time points. BM4 simulates this high-dimensional regime ($p \gg q$). The latent dynamics are governed by three groups of linear ODEs with structurally different attractors (stable fixed point, oscillatory spiral, slow drift), while observations lie in a $p = 500$ dimensional ambient space via a fixed random nonlinear projection. The spatial mapping Enc_ψ must recover the latent subspace and the correct group dynamics entirely without paired temporal supervision.

Quantities and roles.

- **Global time $t_i \in [40, 80]$:** Observation time; enrolment restricted to the first quarter of the window $[40, 50]$, ensuring a long OOD forecasting interval.
- **Static context $\mathbf{c}_i^{\text{static}} \in \mathbb{R}^{20}$:** 5 informative dimensions with group-specific means (signal) and 15 pure-noise dimensions, mimicking a sparse genomic feature profile routing to $K = 3$ molecular subtypes.
- **Dynamic context $\mathbf{c}_i^t \in \mathbb{R}^5$:** Noisy readout of the latent trajectory, $\mathbf{c}_i^t = \mathbf{z}_t^i + \varepsilon_t$ with $\varepsilon_t \sim \mathcal{N}(\mathbf{0}, 0.01\,\mathbf{I})$, providing a directly informative but noisy auxiliary signal.
- **Observation $\mathbf{x}_{t_i}^i \in \mathbb{R}^{500}$:** Latent state projected through a frozen random nonlinear map with additive Gaussian noise ($\sigma = 0.05$).
- **Challenge:** Architectural stress test for the two-stage protocol: Stage 1 must identify the latent subspace from $p = 500$ dimensional observations.

F.6 BM5: Duffing Oscillator Near Bifurcation (Nonlinear Dynamics)

Scientific motivation. BM5 serves as a robustness evaluation. It places subgroup parameters intentionally near a dynamical bifurcation boundary, probing the limits of the Flow Observability Assumption (Assumption 5). In the chaotic regime, trajectories exhibit sensitive dependence on initial conditions (positive Lyapunov exponent), making long-term backcasting fundamentally ill-posed.

Quantities and roles.

- **Global time $t_i \in [0, 50]$:** Physical time; canonical reference $t_0 = 0$. The chaotic subgroup ($k = 2$) uses a finer integration grid (5000 points) to capture sensitive trajectory divergence.
- **Static context $\mathbf{c}_i^{\text{static}} \in \mathbb{R}^2$:** Two-dimensional coordinates determining membership in one of $K = 3$ forcing regimes: fixed-point ($\delta = 0.5$, $\gamma = 0$), limit-cycle ($\delta = 0.3$, $\gamma = 0.35$), or chaotic ($\delta = 0.05$, $\gamma = 0.40$). The limit-cycle and chaotic subgroups share overlapping context distributions, making them non-separable from context alone.
- **Dynamic context:** None ($r_d = 0$).
- **Observation $\mathbf{x}_{t_i}^i \in \mathbb{R}^2$:** State (x, y) with additive Gaussian noise ($\sigma = 0.05$).

F.7 BM6: Gene Regulatory Network (Single-Cell Genomics)

Scientific motivation. Bridging the gap between analytic ODEs and real transcriptomic data, BM6 employs SERGIO (Dibaeinia & Sinha, 2020), a biologically grounded single-cell simulator. Expression dynamics are governed by the stochastic Chemical Langevin Equation acting over a 100-gene Regulatory Network (GRN). This benchmark tests CADENCE's ability to handle highly nonlinear, high-dimensional regulatory cascades, intrinsic transcriptional bursting, and emergent bifurcations that arise naturally from GRN topology.

Quantities and roles.

- **Global time $t_i \in [0, 1]$:** Normalized differentiation pseudotime spanning the complete sequence from progenitor onset ($t_0 = 0.0$) to terminal-fate convergence ($t_{\text{end}} = 1.0$). Enrolment is uniform on $[0, 1]$, so t_0 is right-adherent and BAS holds; Fig. 3 plots $t \geq 0.3$ only, because post-enrolment RMSE at earlier times averages over the fewer than 30% of units enrolled by then.
- **Static context $\mathbf{c}_i^{\text{static}} \in \mathbb{R}^2$:** Fate-correlated synthetic covariates (mimicking polygenic risk scores). The shared standard deviation gives the two branch-conditional context laws an overlap coefficient $\int \min(p_U, p_T) = 0.27$ under equal branch priors (300 test units each), so the Bayes accuracy from context alone is $1 - 0.27/2 = 0.865$ (T vs. U branch), so that context alone separates the fates only imperfectly; the router sees context only, and its 85% accuracy lies just below that bound.
- **Observation $\mathbf{x}_{t_i}^i \in \mathbb{R}^{100}$:** Single-cell mRNA concentrations at enrolment, corrupted by additive technical noise ($\sigma = 0.05$).
- **Ground-truth trajectory generation:** Standard SERGIO pipelines output disjoint cross-sectional ensemble snapshots per differentiation bin. To provide a rigorous evaluation target, we modified the simulator to generate *true continuous longitudinal SDE trajectories*. By passing the exact stochastic

end-state of each individual cell as the initial condition for the subsequent differentiation stage, this approach seamlessly propagates genuine transcriptional bursting correlations and continuous biological memory across the entire developmental cascade.

- **Challenge:** CADENCE must learn population-level dynamics from cross-sectional SDE snapshots in which individual trajectories exhibit stochastic transcriptional bursting and emergent bifurcations not reducible to any analytic ODE. The 0.27 overlap between fate-correlated context features stress-tests the SMoE router under ambiguous inputs, leaving the router to resolve fate from static context alone, under the accuracy constraint that overlap induces, while the per-unit encoded state carries the rest (App. I.7).

F.8 BM7: LARRY Lineage-Traced Haematopoiesis (Primary Data)

BM7 is the primary real-world evaluation. LARRY (Weinreb et al., 2020) tracks primary haematopoietic progenitor cells across differentiation by heritable DNA barcodes; the dynamics encompass stochastic gene regulation, epigenetic remodelling and the full continuum of progenitor states, defying simple parametric ODE descriptions. **The cell is the unit**, observed once at culture day 2: one snapshot per unit, the setting of this paper, which LARRY supplies directly. A cell's observation is its own day-2 expression over the 200 highly variable genes (top 200 from log-normalised scRNA-seq counts) and its static context the projection of that same observation onto the top 50 principal components, in a basis fitted on the training day-2 cells over the published PRESCIENT/scDiffEq gene set (the *context basis*); there is no dynamic context ($r_d = 0$). $1{,}488$ day-2 cells carry a two-fate clonal label. Time is the culture day in native units (the solver step of Tab. 9 is 0.033 of a culture day, 120 steps over the four-day window).

The barcode enters in three places and nowhere else. **(i) Labels.** A cell's fate is its clone's terminal fate (Neutrophil- vs. Monocyte-dominant), the label of the published single-cell fate-prediction results, so fate accuracy and AUROC are the quantities those papers report. Labels never enter training; they are used only at evaluation, to resolve the expert-index permutation and to score fate prediction (App. C.4). The two-fate restriction defines which day-2 cells are *evaluated*; it selects no training data. **(ii) Splits.** Evaluation is five-fold cross-fitting by clone, stratified by fate, so no clone contributes cells to two folds and every clone is held out exactly once (per-fold counts in App. H.5); within each training fold 15% of the clones are held out for validation, together with 15% of the unbarcoded cells. **(iii) Later-day state.** A cell cannot be re-observed, so the clone is the only object with a measurable future: the per-unit MAE of App. I.7 is scored against clone-level day-4 and day-6 aggregates, the single BM7 metric not per cell; no clone-mean trajectory is formed in training. Every fate number is also reported per clone, from the same trained model by averaging a clone's day-2 cells before routing; with 1.51 day-2 cells per clone the clone unit is close to the cell unit.

Enrolment, context and temporal signal. Every cell enrols at culture day 2. Training uses every cell of the three culture days, barcoded or not and of any fate, except the held-out clones' cells: about $27{,}950$ day-2 units and pooled day-4 and day-6 populations of about $47{,}700$ and $52{,}600$ cells per fold, each later day forming one pooled, context-free reference population whose assignment to experts is inferred (App. H.5), so no reference cell is linked to a unit, no unit is paired with its own future, and the objective forms no per-unit trajectory. The gene set and context basis are fitted on each fold's training day-2 cells, and the Stage-1 score model on its training cells from all three days, since it must also encode the later-day reference cells; no held-out cell enters any of them. Test cells enter the model at day 2 only; their clones' later-day cells are evaluation targets, never training input. As on BM1–BM6, where Stage 1 never sees an evaluation unit, the held-out errors are not transductive; training Stage 1 on all units would lower held-out MAE by 3–8% across BM1, BM3, BM4 and BM6. Because all cells enrol at once, the context is a fixed per-unit attribute carrying no information beyond the enrolment observation and none from later days; it is only partially informative about fate, which is why BM7's routing accuracy is lower than on the synthetic benchmarks (Tab. 16). BM7's temporal signal comes from population snapshots at a few discrete culture days rather than a continuous observation window: the identifiability analysis assumes the latter, and POS fails when every unit enrols on one day, so Theorem 1 is not claimed on BM7, which tests the method outside the analysed setting. **Challenge:** no analytic ground-truth ODE exists and no cell is re-observed, so only population-level and fate metrics are available per cell; unsupervised fate discovery under genuine biological noise makes BM7 the most demanding test of the decoupled training.

G Detailed Baseline Methods and Evaluation Protocol

Tuning and selection. Every re-run baseline uses the authors' official code, tuned by the 3×3 grid of App. H.1 by validation MAE separately on every benchmark and run over five seeds; for CVFM and COT-FM the sweep selects their authors' published configuration. A second sweep that also gives every baseline network depths $\{2, 3, 5\}$ and twice the training epochs, a budget matching the depth arm of App. I.3, changes no cross-sectional MAE cell by more than 4% and no leader. BM2's scalar serology readout enters each method where it already accepts a covariate: concatenated to the static context for the conditional samplers (CVFM, COT-FM, cMonge) and to the observed state for the transport and sequence models (OT-CFM, TIGON, scNODE, NDP, CoDA, Latent ODE, ODE^{2}VAE, LaDID), so every method receives the same scalar once; CADENCE consumes it through $\mathcal{L}_{\mathrm{feat}}$, the only BM2-specific term in its objective. The selection data differ by method: the baselines'

validation-MAE selection needs the validation units' later states, information for model selection that the cross-sectional setting does not supply; the criterion is MAE, so the SW_2 comparison is against baselines tuned for the other metric and the claim of lowest SW_2 on all six carries that qualifier. CADENCE's encoder and loss weights were fixed on BM1's validation split before the other benchmarks were run and not changed; K is fixed by data construction; the only sweep scored on test is the sensitivity analysis of App. I.3, which selects nothing; and BM7's router-entropy weight, the only CADENCE hyperparameter selected rather than set, is chosen by the label-free held-out-day Sinkhorn divergence on a validation fold of training clones (App. H.5), so no fate label enters any CADENCE hyperparameter. **Evaluation.** For the transport baselines of groups (1)–(2) each per-unit prediction is one deterministic transport (an ODE rollout of the learned field or, for cMonge, one Monge-map application; Sec. 7), started from the unit's own enrolment state in the cross-sectional columns; in the oracle block OT-CFM and TIGON integrate the full-trajectory-conditioned population from the ground-truth initial state. No evaluation script averages draws: each cell of Tab. 2 is one transport, and SW_2 uses 200 random projections drawn once per benchmark and shared by every method. For group (1) and cMonge that transport is deterministic, so the cell has no inference-time variance; for CVFM and COT-FM it is one realisation of the fitted conditional law, which is why Tab. 40 can re-score them at the median of 128 draws and move them. These rows therefore score each conditional model's sampled transport, not its point-optimal prediction (the conditional median for absolute error, the mean for squared error), so the MAE gaps to groups (1)–(2) in Tabs. 2 and 35 are upper bounds on their point-prediction error, and the leaf-conditional-mean control of Tab. 31 is the point-prediction-fair comparison.

G.1 Cross-Sectional Population Dynamics

These infer global, continuous-time population flows from static, unpaired snapshots; they learn the aggregate evolution of an ensemble but generally lack rich individual-level context conditioning (e.g. genetic or demographic profiles). **scDiffEq** (Vinyard et al., 2025) reconstructs dynamics from single-cell snapshots with neural SDEs, decomposing deterministic drift and state-dependent diffusion, trained by Sinkhorn divergence to match ensemble marginals. **TIGON** (Sha et al., 2024) (trajectory inference with growth via OT and neural networks) poses a dynamic unbalanced OT problem in the Wasserstein–Fisher–Rao distance, capturing trajectories together with ensemble growth and death (source/sink dynamics). **OT-CFM** (Tong et al., 2023) (optimal-transport conditional flow matching) trains continuous normalising flows simulation-free by regressing against vector fields built from exact or minibatch OT maps, avoiding the stiff integration of standard neural ODEs. **scNODE** (Zhang et al., 2024) couples a VAE with a neural ODE in a two-stage pipeline: the VAE learns a nonlinear low-dimensional representation of ensemble gene expression across all observed timepoints, a first-order drift network parameterises developmental dynamics in that latent space, and a *dynamic regularisation* term penalises the Wasserstein discrepancy between VAE-encoded latent distributions and the ODE-evolved trajectory at each timepoint, for robustness to distribution shift at unobserved intervals (interpolation and extrapolation).

G.2 Contextual and Heterogeneous Dynamics

To test context-dependent dynamical parameters ($\mathbf{c}_i$; DFA, COA′, MRA) we compare against models built for heterogeneous environments. **CoDA** (Kirchmeyer et al., 2022) (context-informed dynamics adaptation) conditions a shared dynamics model on environment-specific context vectors through a hypernetwork, targeting distribution shift across physical systems with a low-rank, locality-constrained adaptation space. **Neural ODE Processes (NDPs)** (Norcliffe et al., 2021) extend neural processes to time, modelling a distribution over neural ODEs in which a context set stochastically controls both the initial state and the derivative function, a strong uncertainty-aware, context-driven baseline. Run cross-sectionally on BM1 (one snapshot per unit, static context supplied directly), CoDA and NDP reach MAE 69.83 and 37.14 against CADENCE's 2.05: both infer context from an environment's or unit's own trajectory (CoDA by adaptation, NDP via a context set), so a single observation leaves that inference underdetermined and both reduce to one shared flow. Tab. 2 reports them in the oracle block, with the dense trajectories their design requires.

G.3 Latent Continuous-Time Models

These irregular time-series models need sequential data and are trained with full-trajectory access, marking what dense longitudinal supervision provides. **Latent ODEs** (Rubanova et al., 2019), reported as "Latent ODE" in Tabs. 2–12, pair an ODE-RNN encoder with a neural ODE decoder for irregularly sampled observations (the "Latent-ODE" and "ODE-RNN" rows of App. I.1 are different objects: dynamics variants on frozen CADENCE latents). **ODE^{2}VAE** (Yildiz et al., 2019) is a deep generative second-order ODE model with Bayesian neural networks, splitting the latent space into position and momentum to capture higher-order kinematics and dynamical uncertainty by variational inference. **LaDID** (Lagemann et al., 2023) (latent dynamics via invariant decomposition) uses spatio-temporal transformers and VAEs to separate realisation-specific information (initial conditions, constants) from a universal, realisation-invariant latent dynamical system, an attention-based alternative to neural ODEs.

G.4 BM7 (LARRY) Domain-Specific Baselines

The BM7 comparisons (Fig. 4) add published methods for lineage-traced single-cell fate mapping. **PRESCIENT** (Yeo et al., 2021) learns a potential-driven SDE over expression space matched to population snapshots by Sinkhorn divergence, used for fate and trajectory prediction from early-timepoint cells. Every BM7 fate comparator was run by us: scDiffEq, PRESCIENT and TIGON from their authors' released code at published settings, on our clone folds, day-2 PCA-50 context and the same $1{,}488$ cross-fitted day-2 cells, through CADENCE's scoring code (App. H.5); their published values appear only in the reproduction checks of Tab. 14, which our re-runs match to within 0.010 on each protocol, and the two published protocols are kept apart throughout (App. H.5; Vinyard et al. (2025), Fig. 2d, is the multi-fate one). Fate-accuracy and fate-AUROC margins are computed between re-runs only; the held-out-day comparison with scDiffEq (Tab. 20) and the Pearson comparison of Tab. 14 use published values.

H Detailed Experimental Results

This section provides a comprehensive overview of additional experimental details, extended quantitative results, and qualitative visualizations that supplement the findings in the main text. We begin in Appendix H.1 by detailing the benchmark-specific hyperparameters to ensure full reproducibility. Next, in Appendix H.2, we present the complete quantitative performance tables (including standard deviations and extended prediction horizons), expanding upon the aggregated metrics shown in the main paper.

To provide deeper insight into the temporal dynamics of the models, Appendix H.3 compares the RMSE over time across the oracle-block baselines. This temporal analysis is further supported by qualitative visualizations, which illustrate the predicted trajectories against the ground truth, accompanied by per-realization RMSE curves. Finally, we explore the model's capacity to handle discrete subgroups via router accuracy heatmaps (Appendix H.4) and conclude with a detailed, granular breakdown of model performance on the LARRY benchmark (Appendix H.5).

H.1 Per-Benchmark Configuration and Hyperparameters

Tab. 8 lists every setting that varies by benchmark and Tab. 9 the shared defaults. Stage 1 learns the spatial encoder by denoising score matching and Stage 2 the latent dynamics. All experiments ran on one NVIDIA RTX A6000 GPU, 6–12 h per model for both stages combined; Stage 1 is a one-time pass per benchmark, and where the encoder is held fixed its latents are frozen once and reused across Stage-2 variants (App. C). Headline runs use 5 seeds; ablation sweeps (Apps. I.1, I.2, I.4, I.5, I.6, I.7, I.8 and I.9) use 3; the component ablation of App. I.3 is single-seed apart from its depth arm.

Table 8: Benchmark settings across all test cases.

	Stage 1 (score model)			Router			Loss weights	
BM	x_{dim}	context into encoder	d_{hidden} / n_{layers}	type	K	pre-compressor, n_{aug}	λ_{feat}	$\lambda_{\mathrm{router}}$ (β)
BM1	2	4	256 / 4	SMoE	3	none, 5	inert	0
BM2	3	4	256 / 4	SMoE	3	none, 5	0.1	0
BM3	2	1	256 / 4	continuous	n/a	none, 5	inert	0
BM4	10[a]	20	512 / 6	SMoE	3	PCA-10, 5	inert	0
BM5	2	2	256 / 4	SMoE	3	none, 5	inert	0
BM6	100	2	512 / 6	SMoE	2	none, 5	inert	0
BM7	200	0[b]	512 / 6	SMoE	2	none, 0 (plain ODE)	inert	2.0 (0)

Configuration notes. The BM4 entry of Tab. 2 is the PCA-10 configuration of App. I.2, with the same pre-compressor given to every cross-sectional and oracle baseline and every metric computed in decoded $\mathbf{x}$-space; its cells (0.0222, 0.0626) are the PCA column of Tab. 24, not the uncompressed column's (0.0608, 0.2413). The uncompressed linear encoder is weaker than PCA and the VAE, and the attention encoder is the weakest of the four; the uncompressed numbers, for CADENCE and every cross-sectional baseline, are in App. I.2. The VAE (q=20) is marginally better still and is not used, because it would interpose a second trained encoder between the data and the model the analysis describes. Score-network capacity is larger on the higher-dimensional BM4, BM6 and BM7. The router is trained through the dynamics objective alone on BM1–BM6; on BM7 an unsupervised entropy regulariser (App. C.2)

Table 9: Settings shared across CADENCE benchmarks and reported runs.

Setting	Value
Stage 1 (VP-SDE)	$d_{\mathrm{emb}} = 64$; $\beta_{\min}/\beta_{\max} = 0.1/20$; 500 epochs; batch 256; lr 10^{-3}; 100 PF-ODE encode/decode steps; $\tau = 0.05 \to 0.999$
Stage 2	$d_w = 14$; ODE/router widths $64/32$; 3 ODE layers; 300 epochs; batch 256; lr 10^{-3}; weight decay 10^{-5}
Routing temperature	$1.0 \to 0.1$ over the first 80% of training (BM3 uses the continuous router and no temperature)
ODE solver	Dopri5 with $N_{\mathrm{int}} = 120$ fixed steps per benchmark window (App. J.7)
MMD	Median-heuristic bandwidth; unbiased U-statistic; $M = 128$ reference points and $n_q = 8$ query times per step
Temporal kernel	$h = 0.1$ in native time units
SW_2	200 random projections (seed 42)

encourages distinct expert specialisation, together with the coupling-mass guard of Eq. 4 under the same weight; a variant that also rewards balanced expert usage performed equivalently and is not used. The FiLM maps are linear, $s(\mathbf{w}) = \mathbf{1} + A_s\mathbf{w}$ and $b(\mathbf{w}) = A_b\mathbf{w}$ with $A_s, A_b \in \mathbb{R}^{(q+n_{\text{aug}})\times d_w}$ initialised at zero, so training starts from the unmodulated backbone, and the expert basis $\mathbf{W}$ has i.i.d. $\mathcal{N}(0, 1/d_w)$ entries, so $\mathbf{w}_i \neq \mathbf{0}$ and A_s, A_b receive gradient from the first step. With $d_w = 14 \leq 2(q + n_{\text{aug}})$ the stacked $[A_s; A_b]$ has full column rank on every trained seed, which makes $\mathbf{w}$ identifiable rather than only its image (App. B.3); an otherwise identical configuration with $d_w = 32$, whose function class is the same, changes MAE by $+0.6\%$ on BM1 and -1.2% on BM5 (three seeds). On BM1–BM6 the context enters h_{base} by concatenation with $(\mathbf{z}, t)$ at its input layer; on BM7 h_{base} takes $(\mathbf{z}, t)$ only.

Baseline configurations. For all baselines (OT-CFM, TIGON, scNODE, NDP, Latent ODE, ODE^{2}VAE, CoDA, LaDID) we use the authors' official code, retain the architectural hyperparameters other than width and the training schedules of the original publications, and adapt only input/output dimensions. Every re-run baseline is tuned by a 3×3 grid over learning rate $\in \{10^{-4}, 3 \cdot 10^{-4}, 10^{-3}\}$ and hidden width $\in \{64, 128, 256\}$, selected by validation MAE on each benchmark separately, then run over five seeds (Tab. 10). The scNODE oracle arm refits its VAE stage on each benchmark's trajectories, as its original protocol does. The scNODE values in Tabs. 11–12 are our re-runs on these benchmarks, which the original papers did not evaluate.

Table 10: Selected learning rate (10^{-4}) / hidden width per baseline and benchmark, by validation MAE over the 3×3 grid. CVFM selects 3/256, COT-FM 10/128 on every benchmark (their authors' configuration); CVFM-S uses CVFM's.

	BM1	BM2	BM3	BM4	BM5	BM6
Cross-sectional						
OT-CFM	3/128	3/128	10/128	3/256	3/128	3/256
TIGON	1/128	3/64	1/128	3/256	1/128	1/256
scNODE	3/256	3/128	3/256	1/256	3/128	1/256
cMonge	3/128	10/64	3/128	3/256	3/128	3/256
Full-trajectory oracles						
OT-CFM	3/128	3/128	3/128	3/256	3/128	3/256
TIGON	1/128	1/128	1/128	1/256	1/128	1/256
scNODE	3/128	3/128	3/256	1/256	3/128	1/256
NDP	3/128	3/128	3/128	3/256	3/128	3/256
CoDA	10/128	3/128	10/64	3/256	3/128	3/256
Lat. ODE	3/128	3/64	3/128	3/256	3/128	3/256
ODE^{2}VAE	3/256	3/128	3/256	1/256	3/256	1/256
LaDID	3/128	3/128	3/128	3/256	3/128	3/256

The persistence audit. Persistence MAE follows from Tab. 31: CADENCE's MAE divided by the "MAE vs. persistence" fraction gives 15.65 (BM1), 0.281 (BM2), 1.830 (BM3), 0.859 (BM4), 0.873 (BM5) and 2.679 (BM6) at H=20; the fractions are printed to one decimal. BM1 has three subgroups whose trajectories diverge, and TIGON and scNODE are context-blind estimators built for single-cell transcriptomics (with an unbalanced-OT growth term and a VAE stage respectively) that integrate every unit through one shared population field. Run in raw observation space, with those stages mis-specified on a diverging three-subgroup problem, such a field places every unit near the cohort's average trajectory, which is further from each unit than its own last observation. A shared field is not the cause in itself: our unrouted shared field on Stage-1 latents reaches 3.25 on BM1 (App. I.1). Three controls locate the difference. (i) Per-coordinate z-scoring of the raw states, fitted on the training split, moves the raw-space baselines little on BM1 (OT-CFM $14.82 \to 13.94$, TIGON $18.64 \to 17.21$, scNODE $17.93 \to 16.85$; three seeds), because BM1's regimes differ in level and shape, not only in scale. (ii) A single shared field on the latents of the *unconditional* encoder reaches 11.62, so most of the drop to 3.25 comes from the context-indexed encoder, which maps each context's law to $\mathcal{N}(\mathbf{0}, \mathbf{I})$ separately and so passes context into a nominally shared field; the latent re-fits of Tab. 23 give the baselines the same benefit. (iii) Fitted to one BM1 subgroup at a time with their official preprocessing, TIGON reaches 3.9–4.6 and scNODE 4.2–5.1 per subgroup, so their pooled BM1 cells reflect the three-regime mixture rather than a configuration failure. OT-CFM and cMonge, which have no growth or VAE stage, reach 14.82 and 13.41 after tuning, below persistence. The same argument predicts no such cells on BM2–BM6, whose subgroups do not diverge comparably.

H.2 Full Quantitative Results

Tables 11 and 12 present the comprehensive MAE and SW_2 evaluations, detailing the mean and standard deviation over the 5 training seeds. These results capture the models' integration accuracy at temporal prediction horizons of $H = 20$ and $H = 40$, respectively, expanding upon the condensed summary provided in the main text. The comparative predictive fidelity remains consistent over these extended rollouts: CADENCE remains competitive in capturing the state-space evolution of the different benchmark systems. Relative dispersion over the cross-sectional baselines, both metrics and both horizons, is 0.03–0.23 on BM1–BM5 and 0.06–0.23 on BM6; what distinguishes BM6 is that its MAE cells sit systematically higher within it (0.14–0.23, against 0.05–0.23 on BM1–BM5), where trajectories pass close to the emergent bifurcation and a minority of hard initial conditions dominate the per-seed mean.

Seed-level variability. On BM2 MAE, OT-CFM's lead over CADENCE (0.021 ± 0.002 vs. 0.025 ± 0.002) has uncorrected $p = 0.013$, and CADENCE's BM2 SW_2 lead (0.050 ± 0.002 vs. 0.055 ± 0.003) has uncorrected $p = 0.017$ in the other direction; on this benchmark the two metrics disagree. On BM1 the SW_2 margin over

CVFM ($p = 0.058$) and on BM3 the margin over CVFM ($p = 0.105$) and COT-FM sit within seed variation, whereas CADENCE's MAE margins are significant on every benchmark it leads. On BM6, CADENCE's MAE is below that of the strongest oracle, OT-CFM with full trajectories (0.434 ± 0.093 vs. 0.466 ± 0.101), but not significantly ($p = 0.62$); its BM6 SW_2 lead over the same arm (0.139 ± 0.008 vs. 0.163 ± 0.013) has uncorrected $p = 0.011$.

Multiplicity. Seven tests over six contested cells (BM3's SW_2 is tested against both CVFM and COT-FM) are run above and the tests are not independent, so the raw p-values overstate what the family supports. Under a Holm correction across those seven, *none* reaches 0.05: the smallest, BM6's SW_2 margin at $p = 0.011$, faces a threshold of $0.05/7 = 0.0071$. No claim rests on a seed-level Welch test. The per-unit Wilcoxon tests of App. I.7 are a different family and a far larger one: over the fourteen benchmark-by-null comparisons the largest p is 6.7×10^{-5}, against a Holm threshold of $0.05/14 = 0.0036$, so all fourteen survive, the largest p lying about fifty-fold below the strictest Holm threshold.On SW_2, CVFM tracks CADENCE to within 6.0–18.2% on BM1–BM3 and BM5 and COT-FM to within 6.0–36.5%; the gap widens on BM4 ($1.3\times$, $1.4\times$) and BM6 (48%, 61%).

SW_2-selected baselines. Because the grid selects on validation MAE, we re-selected the three strongest cross-sectional baselines on validation SW_2 over the same grid (five seeds). Their SW_2 at $H{=}20$ improves by 3–6% and their MAE worsens by 2–6%: CVFM 0.072, 0.052, 0.051, 0.0768, 0.155, 0.196 on BM1–BM6, COT-FM 0.090, 0.059, 0.051, 0.0829, 0.178, 0.213, and OT-CFM 0.301, 0.052, 0.087, 0.0871, 0.419, 0.276, against CADENCE's 0.070, 0.050, 0.050, 0.0626, 0.137, 0.139. CADENCE keeps the lowest mean SW_2 on all six, and the BM1 and BM3 margins remain within seed variation. CADENCE's BM2 MAE is marginally lower at $H{=}40$ (0.024) than at $H{=}20$ (0.025): the field is time-dependent and the SIR states relax towards equilibrium, so the grid points added at the longer horizon carry small errors and lower the mean. The sequence-trained oracles are between $1.3\times$ and $21\times$ worse than CADENCE on SW_2 despite dense trajectories: the $21\times$ end is LaDID on BM2, where every sequence-trained oracle is at least $9\times$ worse; the runner-up on that benchmark is $13\times$. The seed-level tests here and the paired Wilcoxon tests over per-unit errors (App. I.7) answer different questions: the former whether each benchmark's aggregate margin is robust across independently retrained models, the latter whether individual forecasts beat the persistence and leaf-mean nulls unit by unit. Neither subsumes the other.

Table 11: MAE and Sliced Wasserstein Distance (SW_2) at $H{=}20$. All values report **mean $\pm$ standard deviation across $S{=}5$ independent random seeds**. Best cross-sectional mean per row in **bold**; the seed-level Welch tests and their Holm correction are in App. H.2. Best oracle value *underlined italic*. BM7 (LARRY) uses domain-specific metrics and is reported in Fig. 4 and App. H.5. Matched access on BM2 and BM4 and the baseline tuning sweep are described in the caption of Tab. 2. Three oracles appear only here (Latent ODE, ODE^{2}VAE, LaDID); ODE^{2}VAE has the best oracle MAE on BM3 (0.131).

		Cross-Sectional Data							Full Trajectory Data (Oracle)							
BM	**Metric**	**CADENCE**	OT-CFM	TIGON	scNODE	CVFM	COT-FM	cMonge	OT-CFM	TIGON	scNODE	NDP	CoDA	Latent ODE	ODE^{2}VAE	LaDID
1	MAE	**2.05** $\pm$**0.31**	14.82 $\pm$1.21	18.64 $\pm$1.52	17.93 $\pm$1.44	11.97 $\pm$0.85	13.97 $\pm$0.95	13.41 $\pm$1.08	*3.11 $\pm$0.24*	*9.86 $\pm$0.78*	*14.86 $\pm$1.19*	*5.84 $\pm$0.47*	*7.92 $\pm$1.03*	*8.46 $\pm$1.58*	*9.13 $\pm$1.04*	*10.75 $\pm$1.47*
	SW_2	**0.070** $\pm$**0.003**	0.318 $\pm$0.021	0.561 $\pm$0.038	0.604 $\pm$0.041	0.075 $\pm$0.004	0.095 $\pm$0.005	0.412 $\pm$0.029	*0.042 $\pm$0.002*	*0.094 $\pm$0.007*	*0.487 $\pm$0.031*	*0.092 $\pm$0.006*	*0.331 $\pm$0.077*	*0.324 $\pm$0.047*	*0.341 $\pm$0.031*	*0.357 $\pm$0.038*
2	MAE	0.025 $\pm$0.002	**0.021** $\pm$**0.002**	0.108 $\pm$0.006	0.131 $\pm$0.008	0.031 $\pm$0.003	0.036 $\pm$0.003	0.162 $\pm$0.011	*0.005 $\pm$0.002*	*0.008 $\pm$0.001*	*0.128 $\pm$0.009*	*0.155 $\pm$0.010*	*0.109 $\pm$0.008*	*0.169 $\pm$0.011*	*0.184 $\pm$0.014*	*0.271 $\pm$0.016*
	SW_2	**0.050** $\pm$**0.002**	0.055 $\pm$0.003	0.119 $\pm$0.006	0.842 $\pm$0.031	0.054 $\pm$0.003	0.062 $\pm$0.003	0.704 $\pm$0.029	*0.037 $\pm$0.002*	*0.043 $\pm$0.003*	*0.556 $\pm$0.018*	*0.462 $\pm$0.015*	*0.508 $\pm$0.022*	*0.512 $\pm$0.020*	*0.643 $\pm$0.031*	*1.073 $\pm$0.042*
3	MAE	**0.097** $\pm$**0.012**	1.402 $\pm$0.079	1.427 $\pm$0.078	1.118 $\pm$0.064	0.424 $\pm$0.028	0.524 $\pm$0.032	0.581 $\pm$0.036	*0.271 $\pm$0.017*	*0.463 $\pm$0.027*	*1.061 $\pm$0.058*	*0.964 $\pm$0.052*	*1.284 $\pm$0.081*	*1.183 $\pm$0.078*	*0.131 $\pm$0.011*	*0.988 $\pm$0.057*
	SW_2	**0.050** $\pm$**0.002**	0.091 $\pm$0.005	0.093 $\pm$0.005	0.612 $\pm$0.019	0.053 $\pm$0.003	0.053 $\pm$0.003	0.431 $\pm$0.016	*0.034 $\pm$0.002*	*0.041 $\pm$0.002*	*0.594 $\pm$0.019*	*0.196 $\pm$0.009*	*0.378 $\pm$0.019*	*0.301 $\pm$0.014*	*0.071 $\pm$0.004*	*0.681 $\pm$0.027*
4	MAE	**0.0222** $\pm$**0.0011**	0.0441 $\pm$0.0026	0.0607 $\pm$0.0035	0.0913 $\pm$0.0051	0.0508 $\pm$0.0029	0.0624 $\pm$0.0034	0.0941 $\pm$0.0058	*0.0139 $\pm$0.0013*	*0.0197 $\pm$0.0015*	*0.0864 $\pm$0.0049*	*0.0571 $\pm$0.0034*	*0.0913 $\pm$0.0081*	*0.0774 $\pm$0.0048*	*0.1185 $\pm$0.0074*	*0.0781 $\pm$0.0051*
	SW_2	**0.0626** $\pm$**0.0034**	0.0912 $\pm$0.0051	0.1287 $\pm$0.0069	0.2241 $\pm$0.0107	0.0801 $\pm$0.0044	0.0869 $\pm$0.0047	0.2604 $\pm$0.0131	*0.0186 $\pm$0.0018*	*0.0281 $\pm$0.0021*	*0.2213 $\pm$0.0107*	*0.1628 $\pm$0.0074*	*0.8264 $\pm$0.0412*	*0.2041 $\pm$0.0091*	*0.4573 $\pm$0.0214*	*0.2662 $\pm$0.0128*
5	MAE	**0.103** $\pm$**0.012**	0.204 $\pm$0.013	0.538 $\pm$0.029	0.261 $\pm$0.016	0.260 $\pm$0.016	0.358 $\pm$0.020	0.604 $\pm$0.033	*0.068 $\pm$0.005*	*0.228 $\pm$0.013*	*0.312 $\pm$0.019*	*0.301 $\pm$0.016*	*0.254 $\pm$0.015*	*0.512 $\pm$0.048*	*0.207 $\pm$0.013*	*0.548 $\pm$0.029*
	SW_2	**0.137** $\pm$**0.003**	0.441 $\pm$0.026	1.284 $\pm$0.061	0.712 $\pm$0.034	0.162 $\pm$0.008	0.187 $\pm$0.009	0.831 $\pm$0.042	*0.033 $\pm$0.002*	*0.079 $\pm$0.004*	*0.601 $\pm$0.023*	*0.294 $\pm$0.013*	*0.462 $\pm$0.023*	*0.724 $\pm$0.061*	*0.418 $\pm$0.020*	*0.897 $\pm$0.038*
6	MAE	**0.434** $\pm$**0.093**	0.688 $\pm$0.141	1.412 $\pm$0.211	1.244 $\pm$0.203	0.722 $\pm$0.161	0.822 $\pm$0.177	1.283 $\pm$0.254	*0.466 $\pm$0.101*	*1.318 $\pm$0.196*	*0.874 $\pm$0.158*	*0.571 $\pm$0.123*	*0.802 $\pm$0.241*	*0.938 $\pm$0.211*	*0.811 $\pm$0.227*	*1.394 $\pm$0.241*
	SW_2	**0.139** $\pm$**0.008**	0.291 $\pm$0.021	0.702 $\pm$0.058	0.708 $\pm$0.061	0.206 $\pm$0.019	0.224 $\pm$0.021	0.471 $\pm$0.037	*0.163 $\pm$0.013*	*0.624 $\pm$0.056*	*0.631 $\pm$0.071*	*0.371 $\pm$0.052*	*0.501 $\pm$0.041*	*0.715 $\pm$0.072*	*0.766 $\pm$0.081*	*0.418 $\pm$0.045*

Table 12: MAE and Sliced Wasserstein Distance (SW_2) at $H{=}40$. All values report **mean $\pm$ standard deviation across $S{=}5$ independent random seeds**. Best cross-sectional mean per row in **bold**; the seed-level Welch tests and their Holm correction are in App. H.2. Best oracle value *underlined italic*. BM7 (LARRY) uses domain-specific metrics and is reported in Fig. 4 and App. H.5. Matched access on BM2 and BM4 and the baseline tuning sweep are described in the caption of Tab. 2.

		Cross-Sectional Data							Full Trajectory Data (Oracle)							
BM	**Metric**	**CADENCE**	OT-CFM	TIGON	scNODE	CVFM	COT-FM	cMonge	OT-CFM	TIGON	scNODE	NDP	CoDA	Latent ODE	ODE^{2}VAE	LaDID
1	MAE	**2.91** $\pm$**0.42**	13.25 $\pm$0.89	18.28 $\pm$1.12	17.50 $\pm$1.11	20.37 $\pm$1.45	23.38 $\pm$1.55	18.57 $\pm$1.25	*6.21 $\pm$0.41*	*17.32 $\pm$1.17*	*14.89 $\pm$0.95*	*6.53 $\pm$0.49*	*9.08 $\pm$1.24*	*9.87 $\pm$1.26*	*11.01 $\pm$1.01*	*11.10 $\pm$1.29*
	SW_2	**0.078** $\pm$**0.003**	0.359 $\pm$0.014	0.558 $\pm$0.022	0.609 $\pm$0.021	0.092 $\pm$0.005	0.109 $\pm$0.006	0.471 $\pm$0.024	*0.075 $\pm$0.005*	*0.163 $\pm$0.010*	*0.542 $\pm$0.014*	*0.103 $\pm$0.007*	*0.378 $\pm$0.074*	*0.382 $\pm$0.047*	*0.378 $\pm$0.030*	*0.410 $\pm$0.038*
2	MAE	**0.024** $\pm$**0.002**	0.026 $\pm$0.006	0.108 $\pm$0.005	0.132 $\pm$0.008	0.034 $\pm$0.003	0.045 $\pm$0.003	0.164 $\pm$0.010	*0.009 $\pm$0.004*	*0.017 $\pm$0.002*	*0.130 $\pm$0.008*	*0.159 $\pm$0.010*	*0.106 $\pm$0.007*	*0.194 $\pm$0.011*	*0.138 $\pm$0.008*	*0.282 $\pm$0.016*
	SW_2	**0.047** $\pm$**0.003**	0.052 $\pm$0.008	0.128 $\pm$0.007	0.877 $\pm$0.027	0.066 $\pm$0.004	0.065 $\pm$0.004	0.753 $\pm$0.033	*0.055 $\pm$0.006*	*0.083 $\pm$0.005*	*0.561 $\pm$0.017*	*0.476 $\pm$0.015*	*0.497 $\pm$0.021*	*0.567 $\pm$0.024*	*0.504 $\pm$0.023*	*1.107 $\pm$0.048*
3	MAE	**0.102** $\pm$**0.015**	1.414 $\pm$0.077	1.439 $\pm$0.076	1.140 $\pm$0.059	0.927 $\pm$0.055	0.928 $\pm$0.055	0.926 $\pm$0.054	*0.509 $\pm$0.030*	*0.821 $\pm$0.046*	*1.025 $\pm$0.051*	*1.023 $\pm$0.053*	*1.485 $\pm$0.088*	*1.318 $\pm$0.069*	*0.178 $\pm$0.013*	*1.021 $\pm$0.057*
	SW_2	**0.053** $\pm$**0.002**	0.072 $\pm$0.004	0.072 $\pm$0.004	0.609 $\pm$0.021	0.066 $\pm$0.003	0.061 $\pm$0.003	0.457 $\pm$0.017	*0.041 $\pm$0.002*	*0.051 $\pm$0.003*	*0.594 $\pm$0.019*	*0.195 $\pm$0.011*	*3.249 $\pm$0.139*	*0.339 $\pm$0.016*	*0.087 $\pm$0.005*	*0.687 $\pm$0.030*
4	MAE	**0.0235** $\pm$**0.0016**	0.0496 $\pm$0.0025	0.0723 $\pm$0.0037	0.0950 $\pm$0.0050	0.0786 $\pm$0.0041	0.0690 $\pm$0.0033	0.1018 $\pm$0.0055	*0.0245 $\pm$0.0046*	*0.0368 $\pm$0.0020*	*0.0954 $\pm$0.0050*	*0.0571 $\pm$0.0028*	*0.1177 $\pm$0.0084*	*0.0959 $\pm$0.0042*	*0.1119 $\pm$0.0061*	*0.0877 $\pm$0.0048*
	SW_2	**0.0628** $\pm$**0.0038**	0.1136 $\pm$0.0080	0.1470 $\pm$0.0069	0.2267 $\pm$0.0080	0.0933 $\pm$0.0045	0.0923 $\pm$0.0044	0.3149 $\pm$0.0143	*0.0225 $\pm$0.0039*	*0.0532 $\pm$0.0026*	*0.2473 $\pm$0.0087*	*0.1619 $\pm$0.0061*	*1.2211 $\pm$0.0514*	*0.2327 $\pm$0.0078*	*0.4330 $\pm$0.0182*	*0.2917 $\pm$0.0121*
5	MAE	**0.174** $\pm$**0.008**	0.206 $\pm$0.012	0.548 $\pm$0.026	0.266 $\pm$0.014	0.401 $\pm$0.022	0.399 $\pm$0.022	0.693 $\pm$0.034	*0.142 $\pm$0.008*	*0.357 $\pm$0.019*	*0.315 $\pm$0.017*	*0.315 $\pm$0.023*	*0.296 $\pm$0.019*	*0.580 $\pm$0.054*	*0.276 $\pm$0.016*	*0.561 $\pm$0.029*
	SW_2	**0.154** $\pm$**0.010**	0.445 $\pm$0.016	1.312 $\pm$0.047	0.712 $\pm$0.024	0.197 $\pm$0.009	0.201 $\pm$0.009	0.853 $\pm$0.032	*0.043 $\pm$0.002*	*0.105 $\pm$0.005*	*0.606 $\pm$0.020*	*0.302 $\pm$0.013*	*0.881 $\pm$0.041*	*0.939 $\pm$0.066*	*0.498 $\pm$0.022*	*0.894 $\pm$0.034*
6	MAE	**0.497** $\pm$**0.098**	0.723 $\pm$0.163	1.455 $\pm$0.204	1.283 $\pm$0.209	0.797 $\pm$0.163	0.887 $\pm$0.192	1.334 $\pm$0.251	*0.533 $\pm$0.112*	*1.397 $\pm$0.218*	*0.922 $\pm$0.158*	*0.629 $\pm$0.237*	*0.879 $\pm$0.255*	*1.055 $\pm$0.214*	*0.896 $\pm$0.252*	*1.431 $\pm$0.240*
	SW_2	**0.148** $\pm$**0.009**	0.303 $\pm$0.019	0.740 $\pm$0.061	0.754 $\pm$0.058	0.223 $\pm$0.027	0.249 $\pm$0.029	0.530 $\pm$0.040	*0.181 $\pm$0.018*	*0.713 $\pm$0.060*	*0.684 $\pm$0.063*	*0.430 $\pm$0.062*	*0.547 $\pm$0.047*	*0.797 $\pm$0.082*	*0.857 $\pm$0.090*	*0.478 $\pm$0.051*

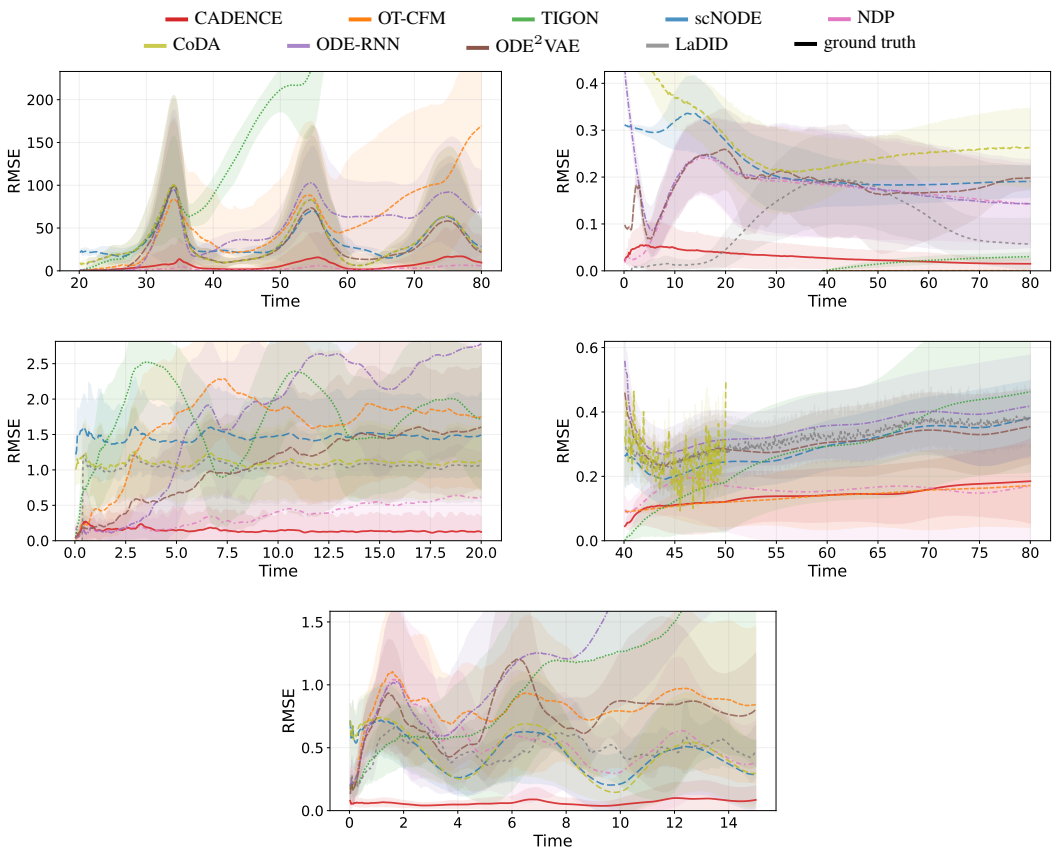

Figure 5: Per-realization RMSE over time (mean $\pm$ std, post-enrollment) for all synthetic benchmarks.

H.3 Trajectory Predictions and RMSE

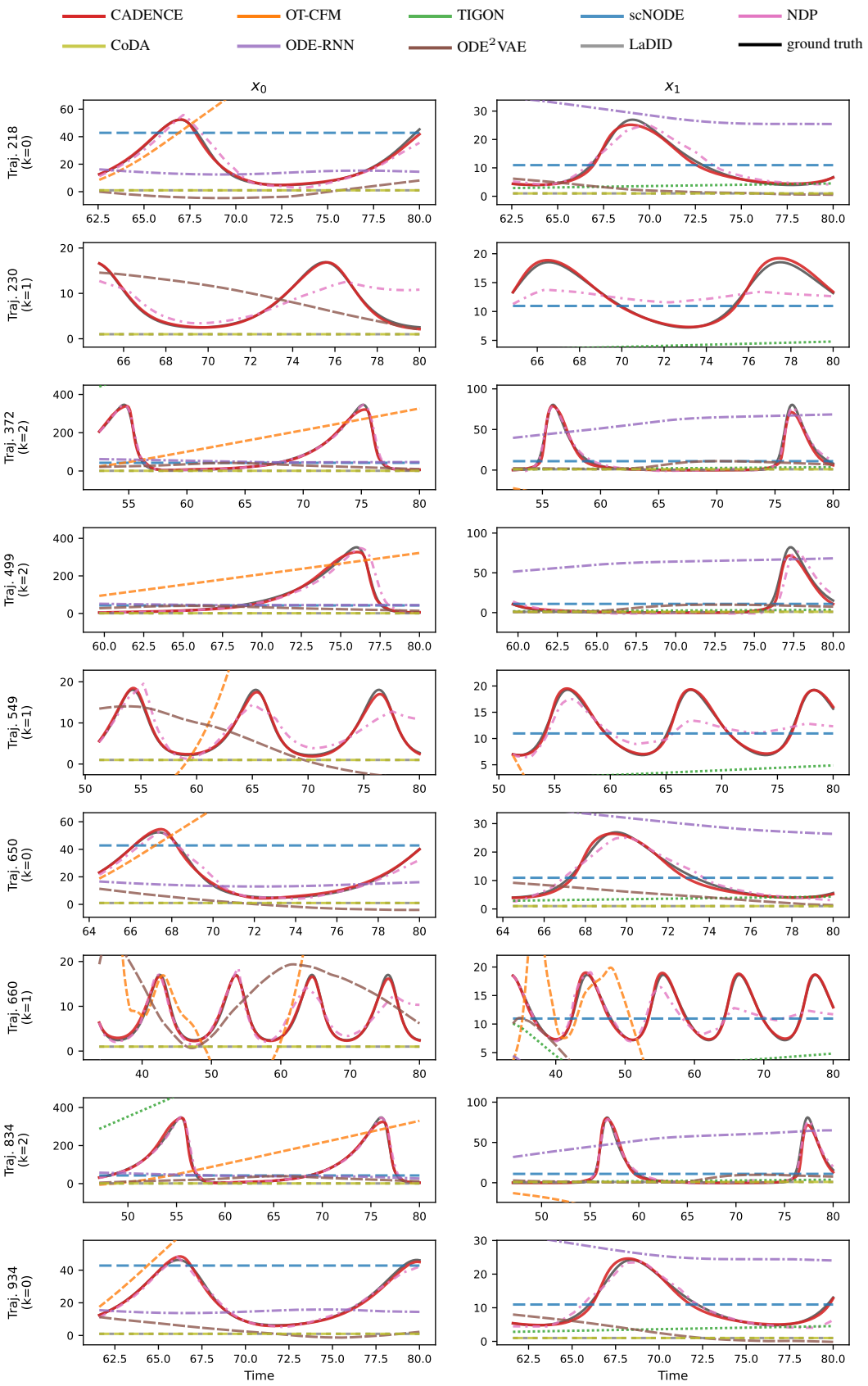

Figure 6: BM1: predicted trajectories vs. ground truth (axes: prey u / predator v; k indexes the subgroup/expert).

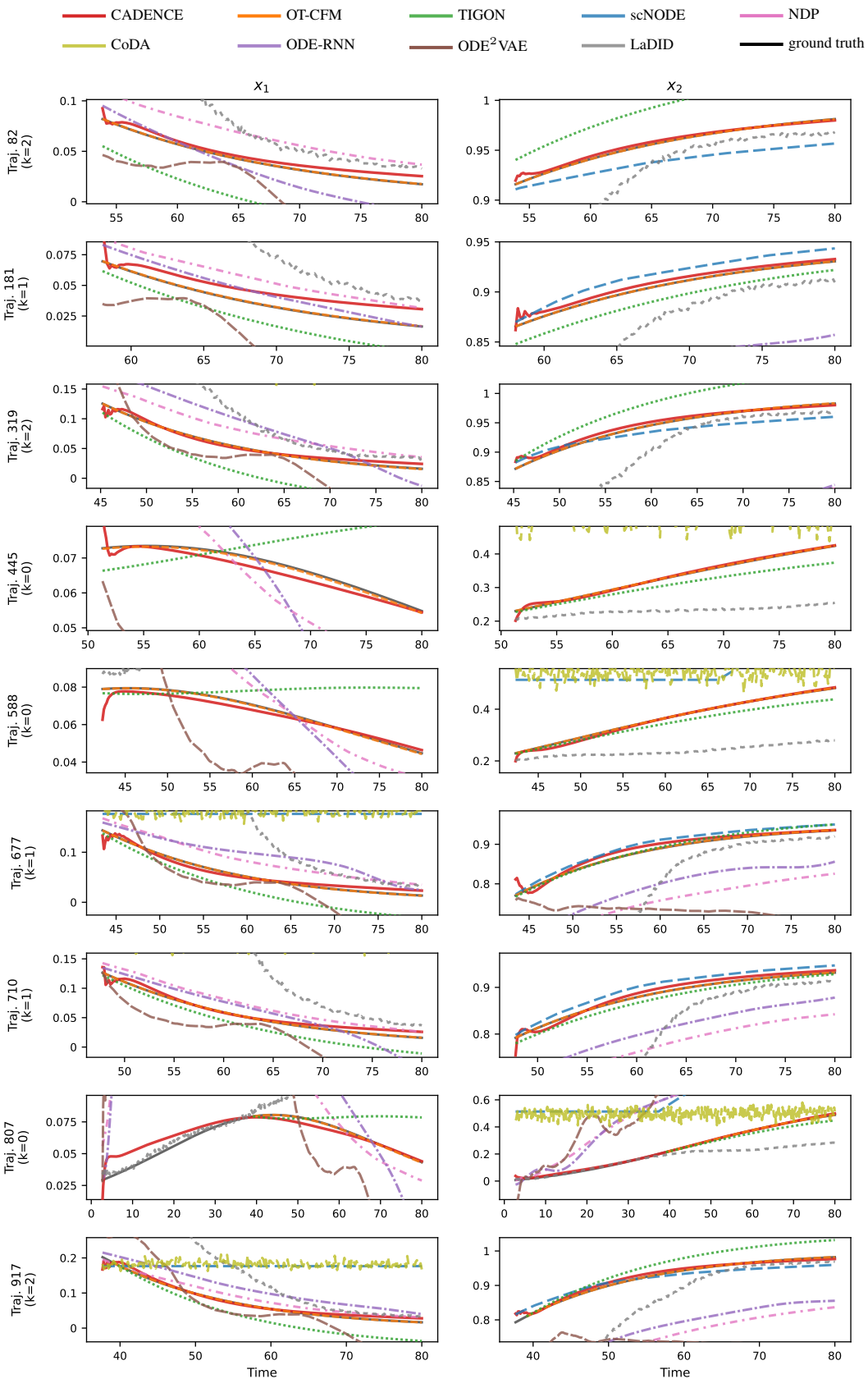

Figure 7: BM2: predicted trajectories vs. ground truth.

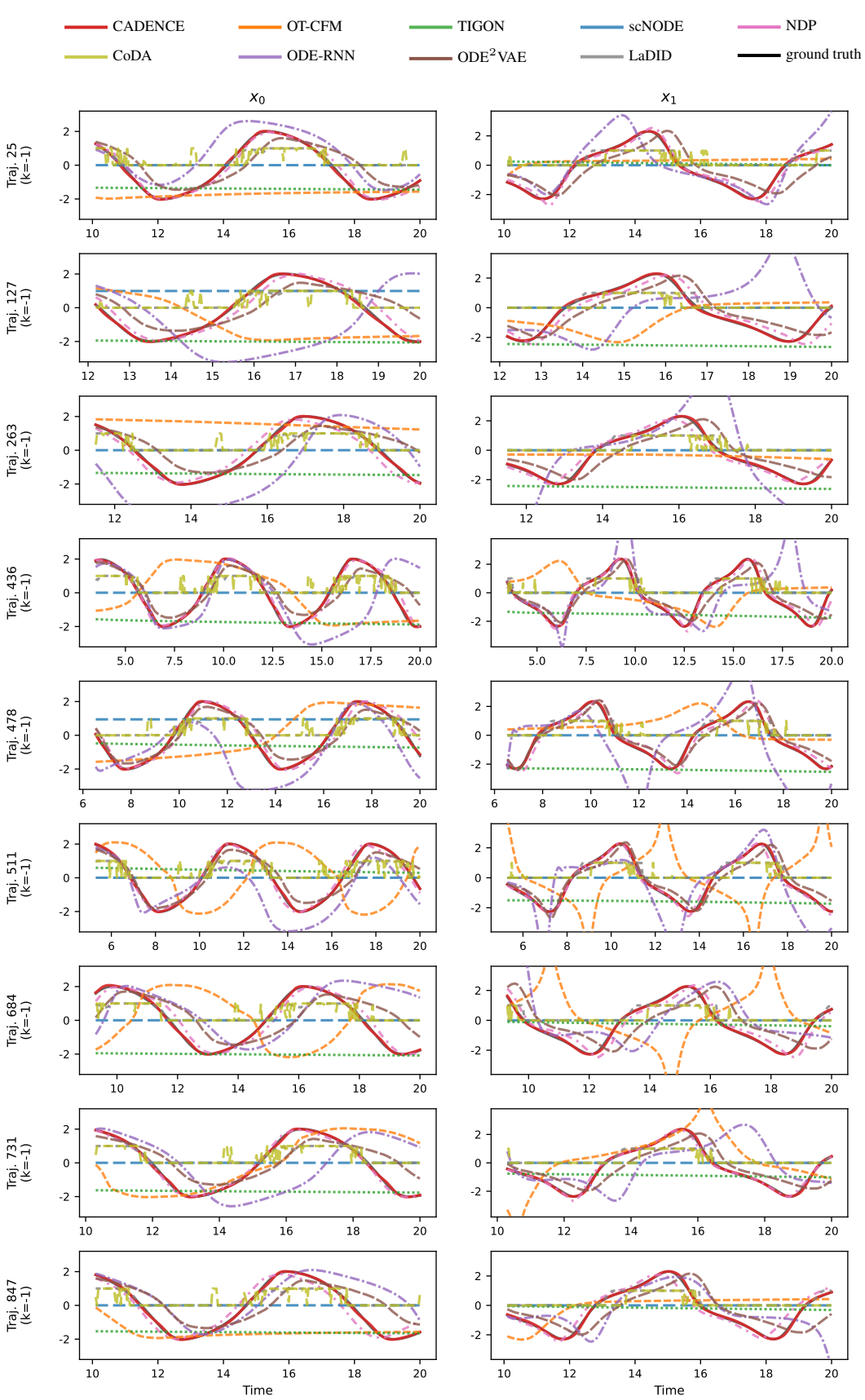

Figure 8: BM3: predicted trajectories vs. ground truth.

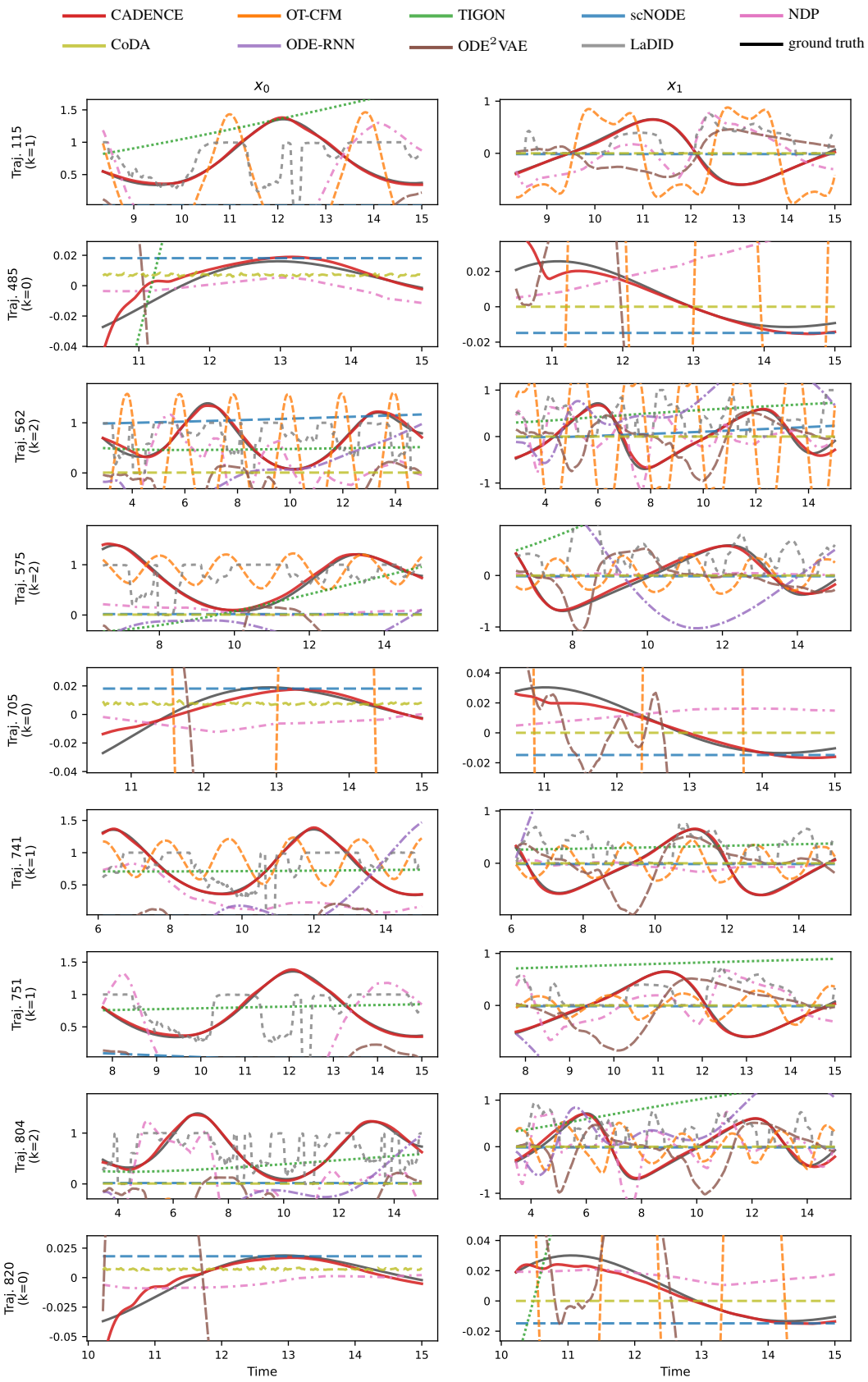

Figure 9: BM5: predicted trajectories vs. ground truth.

H.4 Router Accuracy

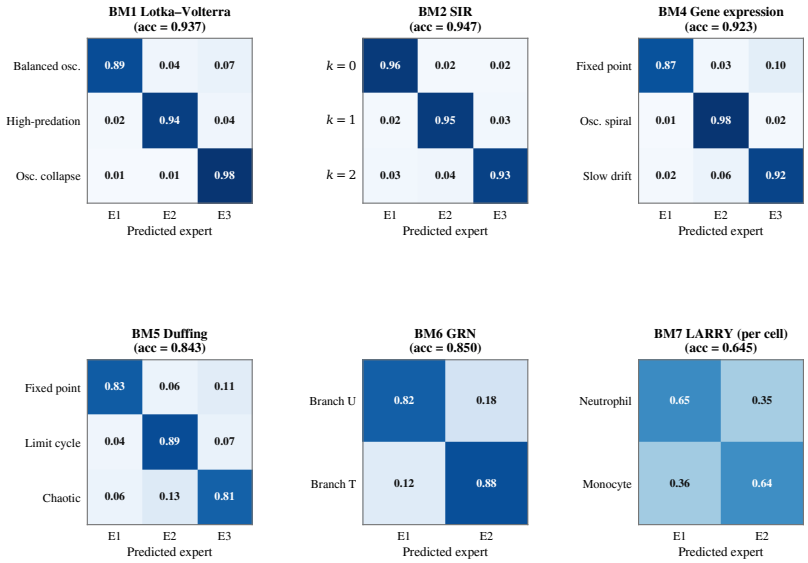

Figure 10: Router accuracy for every benchmark with discrete experts, for the five-seed headline run of each; rows are ground-truth subgroups, columns assigned experts, after the Hungarian alignment of App. C.4. Chance is $1/3$ for the $K{=}3$ benchmarks and $1/2$ for the $K{=}2$ ones; the alignment is fitted on validation, so these are the correct nulls (App. C.4). Panel titles carry the exact accuracy; BM7's panel is the single-cell confusion matrix of the five-seed majority vote per cell reported in Tab. 16 (530 of 815 Neutrophil and 430 of 673 Monocyte cells, $960/1488 = 0.645$), and its clone-level counterpart is $680/986 = 0.690$.

H.5 LARRY Benchmark: Detailed Results

Evaluation protocol. BM7's unit is the day-2 cell (App. F.8); its context is the PCA-50 projection of its own day-2 expression and its label its clone's terminal fate. Evaluation is cross-fitted over five by-clone folds, so all $1{,}488$ day-2 cells of the 986 clones are scored, each by the model that never saw its clone. Fate accuracy is the fraction of cells whose $\arg\max_k \alpha_{ik}$ matches the label after the Hungarian alignment of App. C.4, fitted on the *validation* clones and applied unchanged to the held-out fold for every unsupervised method, so no test label enters an alignment and the null is the coin flip; fitting it on test would select the better of two permutations and raise the null to 0.510 per cell and 0.513 per clone. The validation-fitted permutation coincides with the test-fitted one for every method, so no accuracy depends on the choice, but TIGON would then sit below the null per cell. AUROC uses the router's continuous weight on the Neutrophil expert; CADENCE's AUROC dispersion is over five seeds. This is the protocol of the published single-cell fate-prediction results, cell for cell. Per-day Sinkhorn divergence compares predicted and observed day-d cell populations with no interpolation and no clone averaging. A day-4 or day-6 cell has no day-2 observation and therefore no context, so it is neither routed nor context-encoded. The one metric that cannot be per cell is the per-unit MAE of App. I.7, scored against clone-level aggregates because an individual cell has no observable future; Fig. 12 is the only clone-level display. Every method in Tab. 16 was run by us through the same code path on the same cells at its authors' published settings.

Dataset composition. All counts are re-derived from `AllonKleinLab/paper-data`, which first reproduces the published composition table as a check on the parse: day 2 holds $28{,}249$ cells, 16.4% barcoded ($4{,}638$ cells over $2{,}672$ clones, largest clone 8 cells); day 4 $48{,}498$, 30.9% barcoded ($14{,}985$ over $4{,}101$); day 6 $54{,}140$, 54.8% barcoded ($29{,}679$ over $3{,}956$). Clones committing to a neutrophil or monocyte fate (more day-4 and day-6 cells annotated to one than the other) contribute $1{,}488$ day-2 cells over 986 clones (1.51 per clone); this defines the evaluated cells only, and training uses every cell except the held-out clones'. The folds are stratified by fate: 530 neutrophil- and 456 monocyte-fated clones (815 and 673 cells), 297–298 cells and 197–198 clones per held-out block. *Destructive sampling:* day 2 takes half the culture, day 4 30% of the remaining half and day 6 the rest, so the snapshots represent 50%, 15% and 35% of the population; every cross-day count, growth estimate and reweighting is corrected by these fractions first, which the raw count table hides and which changes the inferred proliferation rates substantially. *Proliferation weighting:* the published protocol samples day-2 cells by a per-clone rate $g^k = \log(n^k_{t+1}/n^k_t)/(t_{t+1} - t_i)$ from the sampling-corrected counts. The rates are large and heavy-tailed (median $+0.949$ per day with a largest-to-median weight ratio of 19.5 over days 2–4; $+0.364$ and 52.3 over days 2–6), and the weighting costs most of the effective sample size: $10{,}000$ weighted draws give 443 of $1{,}303$ (34.0% of uniform) and 259 of $1{,}010$ (25.6%). Every dispersion of a weighted quantity is

computed at the effective sample size (quoting it at the draw count would understate it about fivefold), and every weighted quantity is also reported unweighted and winsorised at the 99th weight percentile, because a few clones dominate the sampled population.

The measured floor and the two scales. Distances between predicted and observed populations are reported relative to a *measured* floor: the Sinkhorn divergence between two independent halves of the same observed day, in the same space and at the same sample size, over five resamples. In the published PCA-50 space it is 16.47 at day 4 and 17.09 at day 6 pooled, 14.81 and 15.92 on the neutrophil and 17.13 and 17.45 on the monocyte branch; every raw value of Tab. 20 is on this scale, as are the published 24.56 and 25.85. Two checks below run in a z-scored variant of that space, whose day-4 pooled floor is 106.6; they are reported as ratios or within-scale differences, and no value crosses between the scales. Measured independently on the z-scored scale, the three model-free rows give 1.65, 1.46 and 1.16 against the published-scale 1.69, 1.53 and 1.23 (within 6% on this column, 10% across both arms), which bounds how much headroom could be a normalisation artefact. The floor is flat in the entropic blur (105.9, 106.3, 107.4, 113.4 at blur 0.01, 0.05, 0.1, 0.5, z-scored), so it is not a tuned constant, and the pooled floors sit *above* the per-branch ones despite three times the cells, because at fixed sample size the split-half divergence tracks spread rather than n; a floor falling with cell count would mean the estimator measured sampling noise instead of geometry. *Two PCA bases* are kept apart: the router's *context basis* is fitted on training day-2 cells alone; the held-out-day divergences and reproduction checks use the published pipeline's *evaluation basis*, fitted on all cells so that the published values stay comparable, which never enters training. With the gene set frozen to the training highly variable genes, the two 50-dimensional subspaces have a median principal angle of 11.6°, pairwise distances between held-out cells (all a Sinkhorn divergence depends on) correlate at 0.9988, and the day-4 floor moves from 106.97 to 106.85 (z-scored), 0.11%, inside its resampling error. The score model and router stay fitted on the training split, so no label pathway sees a test cell.

Later-day cells: a context-free reference. On BM1–BM6 every reference observation is a unit with a context, so Eq. 3 weights it by its own routing. BM7's later-day cells are not units: every clone enrols at day 2 and a day-4 or day-6 cell has no day-2 observation to form a context from. They enter training as one pooled population per culture day with no context, label or clone identity, and their assignment to experts is inferred by the latent coupling of Eq. 4, whose zero set coincides with the stratified objective's under the separation condition of Prop. 3. Hence BM7 runs the *unconditional* encoder, $\mathbf{c}$ entering only the router (a context-indexed encoder cannot encode a cell without context, and reference and prediction must share one latent frame), which is the configuration the analysis assumes (Tab. 37, first row). The barcode enters training in exactly one way (Tab. 13), the by-clone folds, which keep the held-out clones' cells out at every culture day; it supplies no pairing, routing or stratification. The reference is otherwise every cell of its day, barcoded or not and of any fate, so no fate annotation shapes it. Tab. 15 places the headline between a reference with no stratification and one whose stratification is handed over from the barcode; restricting the reference to the two-fate training clones, which needs the fate annotation, is a diagnostic row and moves the result by three cells. Removing the stratification costs more than the clone-clustered s.e. (± 0.015, about 23 cells), while the four stratified rows differ by less, so this benchmark does not rank them. Differential proliferation shapes the later-day populations and is absorbed by the free masses of Eq. 4 (an expert's day-6 share is not tied to its day-2 share); a diagnostic arm reweighting the reference by inverse clone proliferation, which needs clone counts and hence the barcode and so is not a CADENCE configuration, reaches $957/1488$ with AUROC 0.82, within three cells of the headline.

Table 13: BM7 data flow. †Derived from the barcode, which enters training only through the by-clone folds.

Object (unit)	Training	Evaluation
Day-2 expression (cell)	starts each trajectory	scored (two-fate)
Day-2 context (cell)	router input	router input
Day-4, day-6 cells (pooled)	context-free reference, held-out clones excluded (Eq. 4)	held-out-day divergence
By-clone folds†	withhold held-out clones	define held-out clones
Fate label† (clone)	not used, not even to select cells	accuracy, AUROC
Later-day clone means†	not used	per-unit MAE

Calibration. The published fate results use *two different tasks* (Tab. 14). Vinyard et al. (2025) score majority-prediction accuracy over the full annotated fate set (2,000 trajectories per progenitor simulated to the terminal time point, terminal states labelled by approximate nearest neighbours in the published 50-dimensional PCA space). Chance there is far below 0.5: a uniform draw over the ten fates gives 0.10 and the plurality fate 0.382, which is why Dynamo at 26.3% and CellRank and TIGON at 4.0% and 4.1% are meaningful there and meaningless beside a two-class number; the published linear-regression upper baseline, 0.619 ± 0.002, belongs to that protocol only and is not a comparator for our supervised reference. Sha et al. (2024) restrict to neutrophil- and monocyte-committed clones and report Pearson correlation and AUROC against the clonal fate probability; our two-class accuracy shares this task, so TIGON's published AUROC of 0.76 is a fair neighbour for CADENCE's 0.82 and its multi-fate accuracy is not. Each task is reproduced on its own protocol and the

Table 14: Reproduction checks on the two published protocols (A: multi-fate, Vinyard et al. (2025); B: two-class, Sha et al. (2024), Fig. 3f).

Protocol	Method	Metric	ours	published	$\lvert\Delta\rvert$
multi-fate	scDiffEq	acc.	0.581	0.585	0.004
multi-fate	PRESCIENT	acc.	0.494	0.425–0.507	in range
multi-fate	TIGON	acc.	0.044	0.041	0.003
two-class	TIGON	r	0.449	0.453	0.004
two-class	TIGON	AUROC	0.75	0.76	0.01
two-class	WOT	r	0.398	0.403	0.005
two-class	WOT	AUROC	0.65	0.66	0.01

Table 15: BM7 later-day reference stratification: correct (accuracy) and AUROC.

	Single cell		Clone	
Reference	Correct	AUROC	Correct	AUROC
uniform, none (null)	885 (0.595)	0.72	630 (0.639)	0.74
inferred, coupling (headline)	960 (0.645)	0.82	680 (0.690)	0.85
from barcode (oracle)	969 (0.651)	0.83	686 (0.696)	0.86
from barcode, orig. objective	955 (0.642)	0.82	677 (0.687)	0.85
two-fate clones only (diagn.)	963 (0.647)	0.83	682 (0.692)	0.86

two are never printed side by side. The largest discrepancy is 0.010, on the two AUROCs (TIGON and WOT); accuracies and Pearson correlations agree to within 0.005, inside the ± 0.015 clone-clustered s.e. These checks are basis-free (accuracy, cross-entropy, Pearson correlation and AUROC do not depend on the embedding), so they validate the harness without requiring our feature space to match the sources', which licenses running these methods on the held-out-day protocol where no published value exists. No margin is computed from a published value. Two consequences: TIGON and PRESCIENT score at or below the majority-fate rate of 0.548 on the two-class task in our re-runs, and no published two-class accuracy exists for them, so the claims rest on scDiffEq, the strongest of the three, and on the supervised classifiers; and TIGON's published AUROC of 0.76 sits far above what its two-class accuracy suggests, the same threshold effect we report for our router and a further reason to read AUROC on this benchmark.

The input-matched comparison. The router sees only static context, the top principal components of day-2 expression. Four input-matched controls fitted on the same components and split measure the fate signal that context carries: per cell, unsupervised clustering reaches 0.553–0.558 and supervised classifiers 0.590–0.599. CADENCE, without labels, reaches 0.645 (AUROC 0.82), above all of them and every re-run published method. The router is a function of this context alone, so the context carries at least that much fate information; the supervised classifiers do not extract it from about 670 labelled training clones per fold in 50 dimensions, and their per-cell AUROC is still rising with the number of labelled clones (0.55, 0.58, 0.61 for logistic regression and 0.54, 0.57, 0.60 for gradient boosting at 25, 50 and 100%). **The comparison rests on AUROC**: against the two supervised classifiers the AUROC margin is 0.21–0.22 per cell and 0.194–0.213 per clone, 3.4–4.8 times the corresponding accuracy gap. The router's decision rule is its untuned $\arg\max$; a binormal model fitted to its per-cell AUROC of 0.82 places best-threshold accuracy at about 0.74, well above the 0.645 achieved, so accuracy understates its discrimination, and the tuned classifiers show the opposite pattern.

Inference is clustered by clone. Siblings share a label, so the $1{,}488$-cell evaluation carries little more information than the 986-clone one. With mean cluster size $\bar{m} = 1488/986 = 1.51$ and $\mathrm{ICC} = 1$ for labels, the design effect is at most $1 + (\bar{m} - 1)\,\mathrm{ICC} = 1.51$ (the variance of an accuracy depends on the correlation of correctness indicators, below 1 whenever siblings are predicted differently), so the clone-clustered s.e. is at most $\sqrt{1.51} = 1.23$ times the naive binomial one, 0.015 rather than 0.012, and the effective sample size lies between 986 and $1{,}488$. Per-cell scoring matches the literature without adding power; all intervals are clone-clustered. The paired CADENCE$-$scDiffEq difference is $+0.059$ per cell (960 against 872 of $1{,}488$), clone-clustered bootstrap 95% interval $[0.041,\ 0.077]$; a naive cell-level interval would be about a fifth narrower.

Per fold. Each fold re-runs the whole pipeline (Stage 1, Stage 2 and every baseline) on the other four (Tab. 17). CADENCE averages 0.645 ± 0.035 against scDiffEq's 0.586 ± 0.029; the paired difference is positive in every fold, $+0.059 \pm 0.027$, about what a 298-cell block predicts for one fold (clone-clustered s.e. about 0.021), as largely disjoint test blocks should. CADENCE ranks first in all five folds, by 0.037 on average over the per-fold runner-up; adjacent baselines within about 0.01 exchange places and are not separated by this design.

Significance. Both paired tests are at the clone unit, where units are independent, and both contingency tables sum to Tab. 16. Against scDiffEq the table is $(576, 104; 41, 265)$ (both correct, CADENCE only, scDiffEq only, both wrong), 680 and 617 correct; exact two-sided McNemar on the 145 discordant pairs gives $p = 1.7 \times 10^{-7}$. Against gradient boosting it is $(585, 95; 45, 261)$, 680 and 630, $p = 2.9 \times 10^{-5}$ on 140 discordant pairs: a 50-clone margin over the stronger supervised classifier is significant, as is the AUROC margin. Predictions are pooled over five seeds by majority vote per clone, as for every re-run comparator (ensemble against ensemble); the median single seed moves the McNemar values to 3.5×10^{-6} and 3.1×10^{-4}, changing neither verdict, and reaches $948/1{,}488 = 0.637$ per cell (AUROC 0.81) and $673/986 = 0.683$ per clone (AUROC 0.84), above every comparator's ensemble; only the two reference arms, which use fate labels or the evaluated clones' lineage, are higher. At the clone unit DeLong's test gives $p < 10^{-4}$ against each input-matched comparator (logistic regression, gradient boosting, GMM and the label-free OT fate map, all on the same day-2 context). Per cell, where siblings are correlated and DeLong does not apply, Tab. 19 gives clone-bootstrap intervals ($10{,}000$ replicates resampling the 986 clones with all their day-2 cells; an interval including zero is not separated): $+0.19$ to $+0.25$ against the input-matched comparators, $+0.11$ against PRESCIENT, $+0.07$ against TIGON and LineageOT ($p < 0.001$) and $+0.04$ against scDiffEq ($[+0.01,\ +0.07]$, $p = 0.004$).

Table 16: BM7 fate prediction per cell ($n = 1488$, cross-fitted over five by-clone folds) and per clone ($n = 986$, day-2 cells averaged before routing); every method re-run by us on the same cells and features, clone-clustered s.e. ± 0.015 on any accuracy. *Reference arm, not a competitor: LineageOT also given the evaluated clones' barcodes, which no other row sees. †Given labels and the later-day dynamics. #: correct units. The published TIGON value only checks our re-run (Tab. 14).

		Single cell			Clone		
Method	Labels	#	Acc.	AUROC	#	Acc.	AUROC
Chance	–	–	0.500	0.50	–	0.500	0.50
TIGON	none	770	0.517	0.75	543	0.551	0.77
PRESCIENT	none	797	0.536	0.71	563	0.571	0.74
Majority fate	–	815	0.548	–	530	0.538	–
k-means	none	823	0.553	–	570	0.578	–
GMM	none	830	0.558	0.570	577	0.585	0.611
OT fate map	none	843	0.567	0.63	590	0.598	0.66
scDiffEq	none	872	0.586	0.78	617	0.626	0.81
Logistic regr.	**sup.**	878	0.590	0.610	624	0.633	0.656
Gradient boost.	**sup.**	892	0.599	0.600	630	0.639	0.637
LineageOT (barcode)	**lin.**	899	0.604	0.75	644	0.653	0.78
CADENCE	none	**960**	**0.645**	**0.82**	**680**	**0.690**	**0.85**
Sup. reference†	**sup.**	1042	0.700	0.85	758	0.769	0.87
*LineageOT, transd.**	**lin.**	1071	0.720	0.87	780	0.791	0.89
TIGON *(publ.)*	–	–	–	0.76 $\pm$0.02	–	–	–

Table 17: BM7 correct counts per cross-fitting fold F1–F5 (held-out cells: 298, 297, 298, 297, 298); each row sums to its pooled count in Tab. 16. Mean: accuracy $\pm$ s.d. over folds (0.025–0.035, comparable to the ± 0.034 clone-clustered s.e. of a single fold). Paired CADENCE − scDiffEq accuracy per fold: $+0.030$, $+0.030$, $+0.070$, $+0.091$, $+0.074$ (mean $+0.059 \pm 0.027$).

Method	F1	F2	F3	F4	F5	Mean
TIGON	155	147	163	141	164	0.517 $\pm$0.033
PRESCIENT	159	146	169	161	162	0.536 $\pm$0.028
k-means	164	159	173	154	173	0.553 $\pm$0.027
GMM	172	157	178	161	162	0.558 $\pm$0.028
OT fate map	169	157	177	166	174	0.567 $\pm$0.025
scDiffEq	182	167	182	163	178	0.586 $\pm$0.029
Logistic regr.	184	164	181	174	175	0.590 $\pm$0.025
Gradient boost.	177	166	187	174	188	0.599 $\pm$0.030
LineageOT	181	165	187	179	187	0.604 $\pm$0.029
CADENCE	191	176	203	190	200	**0.645** $\pm$0.035

The access-matched control: fate mapping with no model and no labels. The supervised classifiers see day-2 context and labels, CADENCE day-2 context and the unlabelled day-4 and day-6 populations; neither controls for the other. The comparator with *exactly* CADENCE's information is model-free: couple day-2 cells to the later-day marginals by entropic OT, cluster the later-day cells into two groups without annotation, and give each day-2 cell the cluster receiving most of its mass (the Waddington-OT construction (Schiebinger et al., 2019), no fate label, the same populations the dynamics objective consumes). It reaches $843/1488 = 0.567$ per cell (AUROC 0.63) and $590/986 = 0.598$ per clone (AUROC 0.66): above chance and the day-2 clusterings, below the supervised classifiers and clearly below CADENCE, by ninety clones, contingency $(560, 120; 30, 276)$, exact two-sided McNemar $p = 6 \times 10^{-14}$, DeLong $p < 10^{-3}$; both margins are significant. Tab. 20 predicts this: the same map is within $1.08\times$ of CADENCE on the *pooled* day-4 marginal but $1.68\times$ on the neutrophil and $1.34\times$ on the monocyte branch, and fate assignment is a per-branch question. The later-day populations carry the fate signal, a model-free coupling recovers part of it and the dynamics objective more; this does not establish that the margin over the supervised classifiers is a modelling rather than an access effect, since those never see the later-day populations.

The lineage-coupled comparator. LineageOT (Forrow & Schiebinger, 2021) couples consecutive snapshots through the lineage tree, and moscot (Klein et al., 2025) scales that family. CADENCE never sees a barcode in training, so this comparison is not access-matched. Fitted on the training clones' barcode couplings and applied unchanged to held-out day-2 cells, LineageOT is the strongest comparator after CADENCE: $899/1488 = 0.604$ per cell (AUROC 0.75) and $644/986 = 0.653$ per clone (AUROC 0.78). Against CADENCE the margin is 36 clones, contingency $(586, 94; 58, 248)$, exact two-sided McNemar $p = 0.004$, and $+0.07$ AUROC per cell (clone bootstrap $p < 0.001$) and per clone (DeLong $p < 10^{-6}$); **both margins are significant**. On the held-out day it is second in all six columns ($1.20/1.70/1.40$ on Arm I, $1.35/2.25/2.05$ on Arm E). On held-out clones the lineage is withheld by design, so what it carries to a held-out cell is a snapshot coupling shaped by the training barcodes, accurate on the pooled population and weak per branch ($1.70\times$ and $1.40\times$ the floor against CADENCE's $1.55\times$ and $1.31\times$). Fitted *transductively*, on the couplings of all clones including the evaluated ones, as the method is designed to be used, it reaches $1071/1488 = 0.720$ per cell (AUROC 0.87) and $780/986 = 0.791$ per clone (AUROC 0.89), leads CADENCE by 100 clones (contingency $(642, 38; 138, 168)$,

Table 18: BM7 mechanism and leakage controls, single cell ($n = 1488$).

Variant	Correct	Acc.	AUROC
CADENCE (reported)	960	0.645	0.82
later-day reference removed	875	0.588	0.62
context zeroed at router	815	0.548	0.58
context permuted	775	0.521	0.52
fate labels permuted (eval.)	747	0.502	0.51
static one-step map, no ODE	911	0.612	0.76
disjoint genes, context/state	941	0.632	0.80

Table 19: Per-cell AUROC difference, CADENCE (0.82) minus comparator, with clone-bootstrap 95% interval and p.

Comparator	AUROC	Diff.	95% interval	p
Logistic regr.	0.61	$+0.21$	$[+0.18, +0.24]$	<0.001
Gradient boost.	0.60	$+0.22$	$[+0.19, +0.25]$	<0.001
GMM	0.57	$+0.25$	$[+0.22, +0.28]$	<0.001
OT fate map	0.63	$+0.19$	$[+0.16, +0.22]$	<0.001
PRESCIENT	0.71	$+0.11$	$[+0.08, +0.14]$	<0.001
TIGON	0.75	$+0.07$	$[+0.04, +0.10]$	<0.001
LineageOT	0.75	$+0.07$	$[+0.04, +0.10]$	<0.001
scDiffEq	0.78	$+0.04$	$[+0.01, +0.07]$	0.004

exact McNemar $p = 2 \times 10^{-14}$) and by $+0.05$ AUROC per cell ($p = 3 \times 10^{-4}$) and $+0.04$ per clone (DeLong $p = 0.002$), and exceeds the supervised reference. The ordering follows access: with the lineage of the predicted cells LineageOT wins clearly, without it CADENCE does. CADENCE uses no lineage anywhere in training, including to stratify its reference, so on a tissue without lineage tracing only CADENCE can be run.

A state-conditioned flow with CADENCE's access. On the synthetic benchmarks conditioning a sampler on the unit's own snapshot closes most of the MAE gap (App. I.10). The BM7 analogue, CVFM-S trained on day-2 state, context and the same unlabelled later-day populations with no barcode or label, each sampled later-day state assigned to one of two label-free clusters as for the OT fate map, reaches $885/1488 = 0.595$ per cell (AUROC 0.75) and $626/986 = 0.635$ per clone (AUROC 0.77): level with scDiffEq and 54 clones below CADENCE (exact McNemar on $(48, 102)$ discordant clones, $p = 1 \times 10^{-5}$). Its fate assignment flips across 128 draws in 31% of cells, the fate-level counterpart of the increment error of Tab. 36, whereas CADENCE integrates one trajectory per cell; on this dataset snapshot-conditioning does not recover the margin.

Anchoring. For each unit a *supervised reference*, a classifier given the day-2 context, the full day-4 and day-6 expression dynamics *and* the fate labels, fitted by nested five-fold cross-validation inside each training fold, reaches 0.700/0.85 per cell and 0.769/0.87 per clone. CADENCE attains $960/1042 = 92\%$ of its accuracy per cell and $680/758 = 90\%$ per clone, and 91% per cell and 95% per clone of its above-chance AUROC, using neither labels nor dynamics supervision. The reference is what one supervised classifier attains, not an upper bound on every method; for a method without access to the evaluated clones' lineage, a clone-level accuracy materially above 0.77 would warrant a leakage check, and the one arm with that access, transductive LineageOT, is the one that exceeds it (0.791).

Units and context. Averaging a clone's day-2 cells before routing removes measurement noise from the context, and there is little to remove: with 1.51 day-2 cells per clone most clones contribute one cell. Averaging is worth 4.1 accuracy points to CADENCE at a fixed clone unit ($640/986 \rightarrow 680/986$, Fig. 11b) and 7.0 to scDiffEq ($548/986 \rightarrow 617/986$), with the ordering unchanged at every m; the cell-to-clone step, which also changes the unit, is 4.4 points ($0.645 \rightarrow 0.690$).

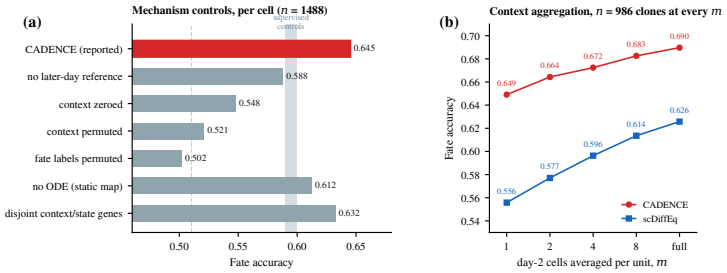

Figure 11: BM7 controls. **(a)** Mechanism controls of Tab. 18, per cell; grey lines: the supervised classifiers. **(b)** Context aggregation, per clone: each of the 986 clones contributes one unit at every m, the mean of m of its day-2 cells drawn with replacement (so the majority of clones, which contribute one cell, are unchanged as m grows), so n is fixed and the trend is aggregation alone.

Every BM7 claim uses the per-cell figure, the one comparable with the literature. Blending each context dimension with a randomly permuted copy across units, with weight η, removes the fate signal: accuracy 0.59 ± 0.03 at η=0.5 and 0.52 ± 0.03 at η=1 (AUROC 0.62, 0.51). CADENCE's fate-probability Pearson correlation against clonal ground truth is 0.531, against the 0.453 TIGON and the 0.403 WOT report (Tab. 14).

Mechanism and leakage controls. Removing the day-4 and day-6 reference marginals, so the objective sees the day-2 marginal alone, drops CADENCE from 0.645 to 0.588, the level of the supervised classifiers on the same context (0.590–0.599): the advantage over supervised classification requires the later-day populations, which the classifiers never see, and this control does not separate the dynamics model from that added data. Zeroing the context at the router gives 0.548, permuting it across cells 0.521, and permuting the fate labels at evaluation 0.502, i.e. chance; that is a scoring check, since a trained predictor scored against permuted labels returns chance whether or not labels reached training. The training-time control permutes the labels at load time, before the fate-stratified fold assignment, and re-runs the cross-fitting; scored against the true labels CADENCE reaches 0.641 ± 0.012 (AUROC 0.81 ± 0.01) over five seeds, inside the sampling error, so labels reach training only through the fold stratification, which changes which clones are held out and nothing else. Two further rows keep the later-day populations and remove one component. Keeping the router, the coupling objective, the mass guard and the populations but replacing each expert's ODE by a static one-step map from day 2 to each later day (no persistent state) gives $911/1{,}488 = 0.612$ per cell (AUROC 0.76) and $652/986 = 0.661$ per clone (AUROC 0.79), between the OT fate map and CADENCE; the clone-level contingency is $(620, 60; 32, 274)$, exact McNemar $p = 0.005$, and on the held-out day it is $1.34\times$ the floor on Arm I pooled against CADENCE's $1.14\times$. The persistent dynamics thus carry part of the margin the later-day populations make possible, the router with the coupling objective the rest. Because the context is a projection of the cell's own day-2 state, BM7 tests *state-derived* contextual routing; to check that the router does not copy the initial state, a second arm splits the 200 genes into two disjoint halves, routes on PCA-25 of one and starts the trajectory from the other, keeping $941/1{,}488 = 0.632$ per cell (AUROC 0.80) and $669/986$ per clone (AUROC 0.83), not separated from the headline (exact McNemar $p = 0.29$ on $(629, 51; 40, 266)$).

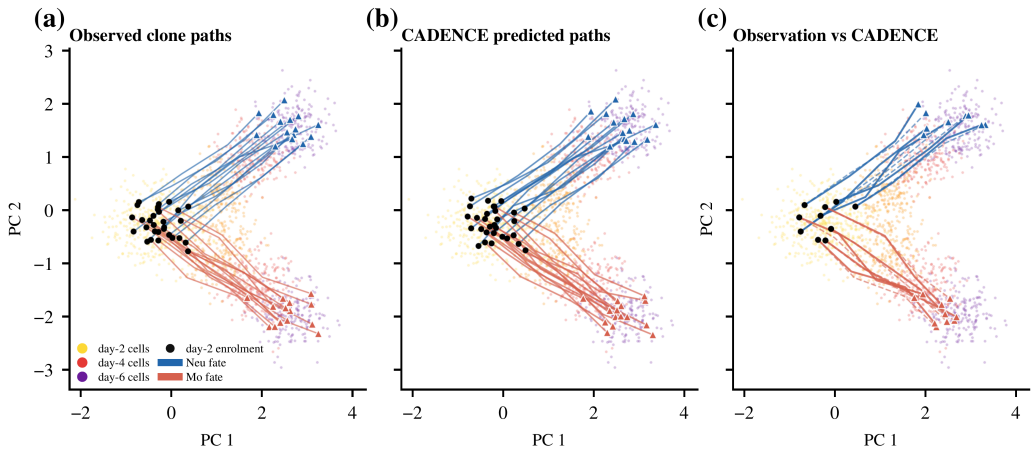

Figure 12: LARRY clone paths in PCA space (qualitative; the expanded view of Fig. 4c). An observed trajectory exists only per clone, because a sequenced cell cannot be re-observed. **(a)** Observed paths over the three culture days, day 2 to day 6; **(b)** CADENCE paths from each clone's day-2 mean, so prediction and observation share a unit; **(c)** both overlaid for ten clones. Background: the measured day-2, 4 and 6 populations. Predicted paths are smooth and fate-divergent, matching the known haematopoietic branching.

Held-out-timepoint forecasting. Predicted clone paths diverge by fate, as the observed ones do (Fig. 12); Tab. 20 measures how closely the predicted later-day populations match the observed ones. Fate prediction is one bit per cell and siblings share it, so the effective sample size is the clone count however many cells are scored. We therefore also evaluate the task only a dynamics model can attempt: **withhold one later day entirely and predict it**. The withheld day enters training in no form (not as a marginal, a target or through the split), so a method that only matches marginals cannot succeed. Arm I withholds day 4 and follows the published protocol of Vinyard et al. (2025) exactly (fit on the lineage-recorded days 2 and 6; sample 10,000 day-2 cells with replacement, weighted by the per-clone proliferation rate; one trajectory per cell; select the epoch of minimum *training* error); Arm E withholds day 6 and has no published counterpart. Tab. 20 reports both and Fig. 13 plots them, with black ticks at the published Arm I values, which our re-runs reproduce. Branch membership is the fate of the *source* day-2 cell's clone, read from the barcode on both sides, so no classifier enters this metric. LineageOT is fitted on the training clones' barcode couplings and applied unchanged to the held-out cells. A bootstrap over 200 resampled day-2 draws gives a standard deviation of 0.03 on a pooled and 0.05 on a branch ×floor value; the five separated margins are 5 to 12 of those, the column reported as no separation 0.6.

Table 20: BM7 held-out-day forecasting: raw pooled Sinkhorn divergence in the published PCA-50 space and × the measured within-day floor (pooled / neutrophil / monocyte). †Model-free. Oracle-tuned bounds choose the step length on the target day itself (attained by: Arm I pooled OT, Neu Linear, Mo OT; Arm E pooled OT, Neu Linear, Mo Linear).

	Arm I: day 4 withheld (published protocol)				Arm E: day 6 withheld (extrapolation)			
Method	raw	pooled	Neu	Mo	raw	pooled	Neu	Mo
Persistence†	27.9	1.69	3.32	2.10	37.9	2.22	5.34	2.99
Linear closure†	25.2	1.53	2.13	1.70	48.7	2.85	2.81	2.66
TIGON	n/a	1.75	3.10	2.35	n/a	2.70	4.20	3.60
Dynamo	27.3	1.66	2.95	2.20	41.9	2.45	3.90	3.30
PRESCIENT	25.9	1.57	2.80	2.05	38.5	2.25	3.40	3.00
scDiffEq	24.5	1.49	2.62	1.92	35.9	2.10	3.10	2.80
COT-FM	21.6	1.31	1.95	1.52	28.2	1.65	2.45	2.25
OT map (interp. / extrap.)†	20.3	1.23	2.60	1.76	42.8	2.50	4.70	3.52
LineageOT	19.8	1.20	1.70	1.40	23.1	1.35	2.25	2.05
CADENCE	**18.8**	**1.14**	**1.55**	**1.31**	**20.9**	**1.22**	**1.97**	**1.79**
Measured floor		16.47	14.81	17.13		17.09	15.92	17.45
Oracle-tuned bound		1.16	2.01	1.57		1.49	2.55	2.35

Model-free bounds. Persistence, an affine Gaussian–Monge closure and the entropic OT displacement map contain no model; computed from the data in the same space against the same floors, they are strong: on pooled interpolation the OT map reaches 1.23× the floor, closing two thirds of the distance between doing nothing (1.69×) and sampling noise, and is 17.3% below the published scDiffEq value, reproducible from the public data alone. Its advantage over the affine closure, +21% pooled (1.16 against 1.46 on the z-scored control scale), vanishes per branch: tied on the monocyte (1.58 against 1.59) and 4% *worse* on the neutrophil branch (1.94 against 1.86), all z-scored, whose per-branch ratios are not those of Tab. 20; on Arm E's neutrophil branch it is 67% worse. In extrapolation it falls behind persistence on the pooled marginal and the monocyte branch while staying ahead on the neutrophil branch. Both effects survive giving the map oracle access to its step length and reproduce in a second feature space, and the same pooled-accurate, branch-weak pattern appears on BM1. The time-proportional step (s=2), the one choice that does not look at the target day, overshoots so badly that on Arm E it is *worse than persistence*, so we do not compare against it: for every column the step length is scanned to its best value on the withheld day itself (an oracle no method could use), giving the oracle-tuned bounds under the table. CADENCE improves on them by 1.20 to 1.31× in five of six columns, a spread reflecting six separate

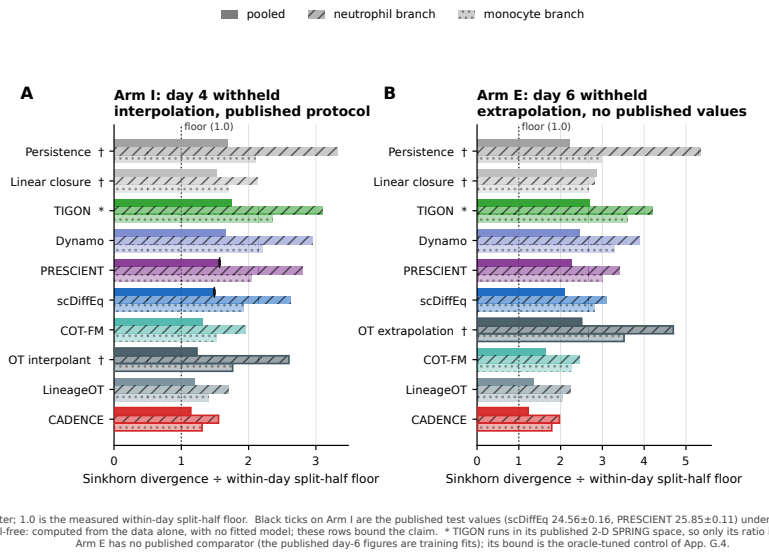

Figure 13: BM7 held-out-timepoint forecasting (Tab. 20), both arms, pooled and per branch, relative to the measured floor; †: model-free; black ticks: published Arm I values.

floor estimates. The exception is pooled interpolation, where the bound is $1.16\times$ the floor and the remaining headroom of 0.16 is below the resampling noise of the observed day-4 population, so no margin is claimed.

Router-entropy prior and mass guard. BM7 is the only benchmark with an active router-entropy term ($\lambda_{\text{router}} = 2.0$). Its weight is selected in each cross-fitting fold on a validation fold of that fold's *training* clones, and all five folds select $\lambda = 2$. The criterion is label-free, a forecasting criterion distinct from the MMD training loss: the held-out-day Sinkhorn divergence of the predicted against the observed day-4 population on the validation clones, in the space and at the sample size of Tab. 20. No fate label enters any CADENCE hyperparameter. Both units and both metrics peak at the selected value (Tab. 21; per cell 0.645, per clone 0.690); the response is a plateau over $\lambda \in [1, 4]$, where AUROC varies by 0.02 and accuracy by 17 clones, rather than a spike at 2.0. Both experts remain in use at every setting. The sharpening term acts per unit ($\beta = 0$) and does not balance usage; the balance comes from the mass guard on the coupling under the same weight (App. B.3; without the guard one expert absorbs 81% of the later-day mass, App. C.2). On BM1–BM6 the same term at λ=2.0 changes MAE by under 3% and routing accuracy by under one point (largest on BM5: $+2.6\%$ MAE, -0.8 points), so it is inert where the router is already well determined and we leave it off there. *The mass guard, varied on its own* (sharpening fixed at $\lambda_{\text{router}} = 2$; in every reported run the guard's weight equals λ_{router}): the same criterion selects the weight used, the fate metrics are flat for any weight of at least 1, and only removing the guard collapses the result, because one expert then absorbs the later-day mass. A class-balance prior does not explain the margin over the comparators: given the same prior when their outputs are converted to a fate (predicted class proportions matched to the training fold's), scDiffEq, PRESCIENT and TIGON reach 0.597, 0.561 and 0.546, and their AUROCs do not change, because AUROC depends only on the ranking a method produces.

Table 21: BM7 selection sweeps; the label-free validation day-4 Sinkhorn selects, the test fate metrics are not used.

λ_{router}	0	0.5	1	**2**	4	8
Val. Sinkhorn	20.1	19.6	19.2	**19.0**	19.1	19.7
Acc., cell	0.596	0.610	0.628	**0.645**	0.637	0.607
Acc., clone	0.639	0.654	0.672	**0.690**	0.681	0.649
AUROC, clone	0.72	0.77	0.83	**0.85**	0.84	0.80
Expert usage	58/42	56/44	54/46	52/48	51/49	50/50
λ_{mass}	0	0.5	1	**2**	4	
Val. Sinkhorn	21.4	19.6	19.2	**19.0**	19.2	
Acc., cell	0.570	0.617	0.635	**0.645**	0.640	
AUROC, cell	0.66	0.79	0.81	0.82	0.81	
Larger share	81%	60%	55%	52%	50%	

I Full Ablation Results

I.1 Dynamics Model Ablation

We compare six Stage-2 dynamics modules. Four are routed: a standard Neural ODE, an Augmented Neural ODE (extra scratch dimensions to break the homeomorphism constraint), a Multi-Block ODE (stacked ODE blocks with nonlinear MLP transitions) and Conditional Flow Matching, all sharing the Stage-1 score-based encoder and the SMoE router (on BM3 the continuous router of App. D.3) and differing only in the integration

module. Two, Latent-ODE and ODE-RNN (our own shared-field variants, distinct from the Latent ODE baseline), use the same frozen Stage-1 latents with no router and no FiLM conditioning: a single shared vector field for the cohort, on latents that on BM1–BM6 come from the context-indexed encoder and so already carry context. With the encoder ablation of App. I.4 this covers both directions: encoder ablated with routing kept, and routing ablated with the encoder kept.

Table 22: Dynamics-module ablation, mean ± s.d.; the routed variants share the Stage-1 encoder and router, the shared-field Latent-ODE and ODE-RNN use the same latents with no router and no FiLM. The Aug. ODE column is the headline configuration of Tab. 2 and reuses its five-seed runs; the other variants are three-seed sweep runs. Bold marks the lower mean across runs with different seed counts and is not a significance test.

		Routed				Shared field, no router	
BM	Metric	ODE	Aug. ODE	Multi-Block	Flow Match.	Latent-ODE	ODE-RNN
H=20							
BM1	MAE	2.1489 ±0.3431	**2.0482** ±0.3117	11.5513 ±1.6742	2.5160 ±0.4032	3.2451 ±0.5021	12.6541 ±1.2954
	SW_2	0.0724 ±0.0039	**0.0703** ±0.0025	0.1426 ±0.0062	0.0813 ±0.0037	0.0895 ±0.0058	0.0951 ±0.0071
BM2	MAE	0.0254 ±0.0022	**0.0247** ±0.0018	0.0312 ±0.0036	0.0809 ±0.0114	0.0324 ±0.0043	0.0897 ±0.0176
	SW_2	0.0523 ±0.0028	**0.0504** ±0.0019	0.0821 ±0.0041	0.0654 ±0.0035	0.0627 ±0.0024	0.0682 ±0.0025
BM3	MAE	0.0993 ±0.0114	**0.0970** ±0.0124	0.5685 ±0.0713	0.2401 ±0.0314	0.1276 ±0.0201	0.4971 ±0.0517
	SW_2	0.0532 ±0.0024	**0.0501** ±0.0021	0.0945 ±0.0042	0.0618 ±0.0034	0.0784 ±0.0026	0.0843 ±0.0069
BM5	MAE	**0.1027** ±0.0121	0.1034 ±0.0122	0.1137 ±0.0103	0.1503 ±0.0207	0.1343 ±0.0303	0.1464 ±0.0395
	SW_2	0.1416 ±0.0037	**0.1374** ±0.0032	0.2103 ±0.0089	0.1502 ±0.0053	0.1945 ±0.0108	0.2151 ±0.0097
H=40							
BM1	MAE	3.0525 ±0.5124	**2.9098** ±0.4217	14.4683 ±1.8521	3.9179 ±0.7256	3.6643 ±0.6124	17.1241 ±1.9421
	SW_2	0.0964 ±0.0032	**0.0783** ±0.0027	0.1852 ±0.0077	0.1081 ±0.0071	0.1185 ±0.0048	0.1264 ±0.0066
BM2	MAE	0.0271 ±0.0025	**0.0240** ±0.0019	0.0348 ±0.0034	0.0811 ±0.0107	0.0406 ±0.0067	0.1104 ±0.0163
	SW_2	0.0672 ±0.0041	**0.0465** ±0.0034	0.1103 ±0.0144	0.0821 ±0.0069	0.0854 ±0.0051	0.0916 ±0.0069
BM3	MAE	0.1033 ±0.0153	**0.1024** ±0.0151	0.6499 ±0.0437	0.2925 ±0.0274	0.1417 ±0.0194	0.8134 ±0.0735
	SW_2	0.0715 ±0.0041	**0.0533** ±0.0024	0.1284 ±0.0163	0.0842 ±0.0087	0.1067 ±0.0072	0.1143 ±0.0137
BM5	MAE	**0.1574** ±0.0092	0.1743 ±0.0081	0.1624 ±0.0149	0.1773 ±0.0105	0.1978 ±0.0133	0.2264 ±0.0491
	SW_2	0.1621 ±0.0117	**0.1544** ±0.0103	0.2754 ±0.0116	0.2021 ±0.0163	0.2783 ±0.0219	0.3045 ±0.0274

Across these runs (Tab. 22; five seeds for the Augmented ODE, three for the others) the Augmented ODE has the lowest MAE among the routed variants on BM1–BM3 at both horizons; on BM5 the plain ODE is lower, marginally at H=20 (0.1027 vs. 0.1034) and by 10% at H=40 (0.157 vs. 0.174). The standard ODE is a close second, so the augmented dimensions give a modest improvement on BM1–BM3; Multi-Block is substantially worse, particularly on BM1 and BM3, suggesting that its inter-block MLP transitions destabilise optimisation; Flow Matching is closest to the ODE variants on BM5 (0.150 vs. 0.103 MAE) and furthest on BM2 (0.081 vs. 0.025). We therefore adopt the Augmented ODE throughout.

How much of the gain is Stage 1, and what routing adds. Removing the router degrades MAE well beyond any routed variant, most on BM1, whose three subgroups are qualitatively distinct (3.25 and 12.65 against 2.05), and on BM3, whose continuous leaf family makes per-unit phase inference the whole task (0.128 and 0.497 against 0.097). Yet the shared-field Latent-ODE's 3.25 on BM1 is already below every latent sequential model of the oracle block of Tab. 2, which train on dense trajectories (best after tuning: NDP, 5.84), so most of CADENCE's margin over those models is the Stage-1 representation. These unrouted fields run on the latents of BM1's context-indexed encoder, which already carry context; on the unconditional encoder's latents the same shared field reaches 11.62 (App. H.1). Context-conditioned dynamics account for the rest: on the same latents the Augmented ODE lowers MAE by 23–37% relative to the unconditioned Latent-ODE (BM1 3.25 → 2.05, BM2 0.032 → 0.025, BM3 0.128 → 0.097, BM5 0.134 → 0.103). The MAE reduction is the smaller half: on BM1 the router's removal raises MAE by 58% but SW_2 by only 27% (0.0895 vs. 0.0703), the unrouted variants' SW_2 (0.090, 0.095) staying close to the routed ones (0.070–0.143) and far below the context-blind cross-sectional baselines of Tab. 2 (OT-CFM 0.318), because with one shared field the objective still fits the population marginal. Scored against the leaf-conditional-mean null the unrouted field's skill is 0.13 against CADENCE's 0.61 (Tab. 41) and its routing is at chance, 0.333, against 0.937: a single field assigns every unit the same dynamics, so forecasts differ only through initial states. The representation makes the problem tractable; routing makes the forecast individual. Latent-ODE outperforms ODE-RNN throughout, suggesting that recurrent updates compound rather than average the same subgroup mismatch.

The sequence oracles on the Stage-1 representation. In the oracle block these models receive dense trajectories in raw space, so their margin to CADENCE mixes representation and dynamics. Re-fitted on CADENCE's frozen Stage-1 latents with the same decoder, splits and three seeds, their BM1 MAE moves from raw to latent

as NDP $5.84 \to 3.12$, CoDA $7.92 \to 3.58$, Latent ODE $8.46 \to 3.91$, LaDID $10.75 \to 4.36$ and ODE^{2}VAE $9.13 \to 4.05$: every one improves 1.9–2.5× and all land in a band around the unrouted shared field's 3.25, against CADENCE's $\mathbf{2.05}$. Most of the oracle-block margin is therefore the representation; what routing adds is the step from that band to 2.05 and, more to the point, from skill 0.13 to 0.61, which no representation supplies. *Fit and time scaling:* on BM1 every oracle is far above the noise level ($\sigma = 0.05$) and a time-aligned joint nearest neighbour reaches 2.71 (Tab. 32), so we checked for mis-configuration. Every baseline receives times min–max scaled to $[0, 1]$ over the benchmark's window (BM1's raw $[20, 80]$ is never passed unscaled); training MAE is within 12% of test MAE (NDP 5.21 against 5.84, CoDA 7.30 against 7.92, Latent ODE 7.95 against 8.46, ODE^{2}VAE 8.61 against 9.13, LaDID 10.02 against 10.75), so they underfit rather than overfit; and a Latent ODE fitted per subgroup with ground-truth labels reaches 2.31. Their BM1 error is thus mainly the cost of fitting three diverging regimes with one field. *Leaf-supplied oracles:* given the true leaf as an input covariate with full trajectories (five seeds), the Latent ODE reaches 2.38 on BM1, 0.019 on BM2 and 0.114 on BM5, and NDP 2.52, 0.021 and 0.121; with the leaf and dense trajectories both, a sequence model is lower than CADENCE's single-snapshot 0.025 on BM2 and stays above its 2.05 and 0.103 on BM1 and BM5, so the main-text claim is restricted to the context-blind oracles as published. On BM2 and BM4, where the sequence oracles trail the full-trajectory OT-CFM, teacher-forced one-step MAE with the matched normalisation is 0.006 (NDP) and 0.007 (CoDA) on BM2 and 0.0049 and 0.0061 on BM4, near the full-trajectory OT-CFM's 0.005 and 0.0139 horizon values: they fit the local dynamics and lose accuracy by compounding over the horizon.

Cross-sectional baselines on the same latents. Tab. 23 re-fits OT-CFM, CVFM and COT-FM on CADENCE's frozen latents, decoded with CADENCE's decoder, with the same splits, five seeds and tuning grid. On BM1, BM3 and BM4 the representation accounts for most of the raw-space margin: OT-CFM falls from 14.82 to 3.38 on BM1, next to the shared field's 3.25, and TIGON and scNODE, re-fitted the same way on BM1, reach 3.60 and 3.73, far below persistence, so their raw-space weakness there is representational. On BM2 the latents do not help OT-CFM. CADENCE keeps the lowest MAE and SW_2 on all six benchmarks and the highest skill; OT-CFM on the latents sits at the shared field's skill level (-0.06 to 0.13), and the conditional samplers reach at most 0.66, below CADENCE everywhere and below the state-conditioned adaptation's 0.74.

Table 23: Cross-sectional baselines re-fitted on CADENCE's frozen Stage-1 latents, H=20, five seeds. Skill against the leaf-conditional mean as in Tab. 41; each cell respects that table's bound.

	BM1	BM2	BM3	BM4	BM5	BM6
MAE						
OT-CFM	3.375	0.0337	0.133	0.0313	0.140	0.583
CVFM	6.389	0.0273	0.267	0.0378	0.210	0.606
COT-FM	7.457	0.0317	0.330	0.0464	0.289	0.690
CADENCE	**2.05**	**0.025**	**0.097**	**0.0222**	**0.103**	**0.434**
SW_2						
OT-CFM	0.0922	0.0646	0.0808	0.0710	0.2003	0.1730
CVFM	0.0712	0.0513	0.0503	0.0761	0.1539	0.1957
COT-FM	0.0902	0.0589	0.0503	0.0826	0.1777	0.2128
CADENCE	**0.070**	**0.050**	**0.050**	**0.0626**	**0.137**	**0.139**
Skill						
OT-CFM	−0.06	0.08	0.02	0.08	0.02	0.13
CVFM	−2.79	0.41	0.66	0.57	−0.45	0.49
COT-FM	−4.16	0.38	0.54	0.54	−1.76	0.37
CADENCE	**0.61**	**0.66**	**0.96**	**0.96**	**0.65**	**0.75**

I.2 Encoder Compression Ablation (BM4)

When the observation dimension far exceeds the intrinsic latent dimension, an encoder must compress before dynamics modelling. BM4 (p=500, true latent dimension 5) is the only such benchmark, and we compare four encoders with Augmented-ODE dynamics, (i) Linear (uncompressed, q=p=500), (ii) PCA (q=10), (iii) VAE (q=20) and (iv) attention-based (q=10), and, with the PCA encoder fixed, five dynamics modules (Tab. 24). Compression substantially improves both metrics, so BM4's headline run is the PCA-10 configuration, with every cross-sectional and oracle baseline given the same pre-compressor and every metric computed in decoded $\mathbf{x}$-space. On BM1–BM3 and BM5 ($p \leq 3$) a same-dimension projection is a rotation and the arm is degenerate, and on BM6 a PCA-20 pre-compressor leaves MAE unchanged within dispersion (0.441 ± 0.095 vs. 0.434 ± 0.093), so every other benchmark stays uncompressed ($q = p$). The VAE is slightly better than PCA at H=20 and slightly worse at H=40; PCA is used because it is deterministic and parameter-free and so interposes no second trained encoder between the data and the analysed model. The attention encoder is substantially worse, likely overfitting in the encoder stage. With the PCA encoder, the Augmented ODE is again best on every metric, the standard ODE a close second, and the SDE variant diverges, suggesting that the learned diffusion coefficient destabilises this high-dimensional setting.

Without the pre-compressor the ordering on BM4 holds (Tab. 25): at q=p=500 CADENCE keeps the lowest MAE against OT-CFM, CVFM and COT-FM (0.0608 against 0.0724–0.1012, Welch $p = 0.003$–0.024) and the lowest SW_2, although its SW_2 lead over CVFM (0.2413 vs. 0.2571, $p = 0.24$) is within seed variation. Removing the compressor costs CADENCE more than any baseline, 2.7× in MAE against 1.6×, because CADENCE's own encoder must then operate in 500 dimensions; the ordering does not depend on PCA-10, the size of the margin does.

Which property of the encoder the router needs. Routing collapses to chance without the PF-ODE, and the VAE arm shows that approximate standardisation is not enough. Tab. 26 separates five unconditional encoders on BM1, BM2 and BM6, each followed by the same SMoE and dynamics module, over three seeds: no encoder with z-scored states; whitening (a linear map to identity covariance); a marginal Gaussian copula (each coordinate

Table 24: BM4 (p=500) in decoded $\mathbf{x}$-space, mean $\pm$ s.d. Left: encoder, each compressing to latent dimension q (Linear is uncompressed, q=p), with Augmented-ODE dynamics. Right: dynamics module with the PCA encoder. The PCA / Aug. ODE column is the five-seed headline run of Tab. 2; the others are three-seed sweep runs. Bold: best per row among the encoders; among the dynamics modules the Aug. ODE (the PCA column) is best on every row.

		Encoder (Aug. ODE)				Dynamics (PCA encoder)			
H	Metric	Linear (q=500)	PCA (q=10)	VAE (q=20)	Attention (q=10)	ODE	Multi-Block	Flow Match.	SDE
H=20	MAE	0.0608 ±0.0024	0.0222 ±0.0011	**0.0206** ±0.0019	0.2041 ±0.0341	0.0232 ±0.0037	0.0251 ±0.0064	0.0557 ±0.0137	0.3092 ±0.0694
	SW_2	0.2413 ±0.0121	0.0626 ±0.0034	**0.0522** ±0.0103	1.1449 ±0.0132	0.0687 ±0.0087	0.0813 ±0.0137	0.2326 ±0.0673	4.0891 ±0.6324
H=40	MAE	0.0734 ±0.0024	**0.0235** ±0.0016	0.0249 ±0.0028	0.2055 ±0.0431	0.0257 ±0.0057	0.0352 ±0.0076	0.0679 ±0.0163	0.3629 ±0.0812
	SW_2	0.2900 ±0.0143	**0.0628** ±0.0038	0.0646 ±0.0153	1.1275 ±0.0168	0.0755 ±0.0103	0.1241 ±0.0169	0.2925 ±0.0742	4.5111 ±0.7032

mapped to $\mathcal{N}(0,1)$ through its empirical CDF, no joint transform); a RealNVP flow trained by maximum likelihood to $\mathcal{N}(\mathbf{0},\mathbf{I})$; and the PF-ODE. The linear maps route at chance, the copula recovers part of the gap, and RealNVP comes within 0.03 of the PF-ODE's routing on all three. The router therefore depends on *joint* Gaussianisation of the latent law rather than on scale, covariance or marginal shape, and any sufficiently accurate Gaussianising flow supplies it; the PF-ODE is our choice for its bijectivity without a trained decoder, not the only encoder that works.

I.3 Component Ablation

A controlled ablation on BM1 ($K^\star$=3) perturbs one setting from the default at a time (Tab. 27) along five axes: (i) the number of experts $K \in \{1,\dots,5\}$, probing sensitivity to mis-specified mixture cardinality; (ii) training-set size $N \in \{100,500,1{,}000,2{,}000,5{,}000\}$, sample efficiency; (iii) latent-ODE width $d_h \in \{32,64,128\}$ and depth $\{3,5\}$; (iv) the auxiliary latent $\mathbf{z}_{\text{aux}}$ and the ghost-trajectory mechanism; and (v) observation noise $\sigma \in \{0.01, 0.05, 0.10, 0.20\}$, retraining both stages at each level. The sweep is scored on test, for comparability with the headline table, and is a sensitivity analysis, not tuning: K is fixed by data construction (BM1's three subgroups, BM7's two terminal fates; App. J.10 shows that label-free selection recovers the same K), and the encoder and loss weights are hand-set. The default (K=3, N=5,000 units of which 3,500 train, d_h=64, 3 layers, no auxiliary conditioning, σ=0.05) is highly competitive on every axis except depth, where 5 layers lower MAE by 24% (2.032 $\to$ 1.541) in this single-seed sweep. Replicated over three seeds, depth 3 gives 2.05 ± 0.20 and depth 5 1.62 ± 0.16, a 21% reduction (Welch $t = 2.91$ on 3.8 degrees of freedom, $p = 0.046$; SW_2 0.073 ± 0.006 vs. 0.060 ± 0.005, $t = 2.88$, $p = 0.047$). The headline keeps the prespecified three layers on all seven benchmarks so that the cross-benchmark comparison is not confounded by per-benchmark tuning; adopting five would widen every BM1 margin by a further 27%, and training at lower noise gains a little further. Performance degrades smoothly under expert mis-specification and still improves between N=2,000 and 5,000 (3.175 $\to$ 2.032); disabling the router (K=1) is the largest drop on the K axis, mirroring Tab. 22. Routing accuracy along the N axis rises from 0.422 (N=100) and 0.518 (N=500) to 0.928, 0.946 and 0.973 at N = 1,000, 2,000 and 5,000, with routing entropy falling from 1.10 to 0.12. On the K axis forecast error is shallow (minimised at K=3 but varying only 1.29× over $K \in \{2,\dots,5\}$), which is not what the analysis predicts at $K = |\Lambda|$; archetype-parameter recovery on the same sweep has a sharp minimum at K=3 (Tab. 48).

I.4 Encoder Factorial Ablation: Compressor × PF-ODE (BM4)

The PF-ODE maps $\mathbb{R}^p \to \mathbb{R}^p$ and cannot reduce dimension, so a compressor (PCA, VAE) is not an *alternative* to it but an optional *pre-compressor* in front of it. The experiment that separates the two is the factorial compressor × PF-ODE with the SMoE + FiLM-Neural-ODE module fixed, on BM4 (p=500, K=3; Tab. 28), where the question is well-posed (at $p \le 3$ the compressor arm is a rotation). Routing is scored against ground-truth subgroups. Dispersions are across this sweep's three seeds and are not comparable with Tabs. 11 and 24, whose BM4 rows come from different runs (PCA: the five-seed headline; Linear and VAE: the sweep of App. I.2), so

Table 25: BM4 with and without the pre-compressor at H=20, decoded $\mathbf{x}$-space. PCA-10: five-seed headline; uncompressed: three seeds, every baseline re-tuned on the same grid. Bold: best per column and values not significantly different from it (Welch, $p \ge 0.05$).

	PCA-10 (headline)		Uncompressed, q=p=500	
Method	MAE	SW_2	MAE	SW_2
OT-CFM	0.0441	0.0912	0.0724 ±0.0043	0.2934 ±0.0168
CVFM	0.0508	0.0801	0.0829 ±0.0049	**0.2571** ±0.0152
COT-FM	0.0624	0.0869	0.1012 ±0.0058	0.2790 ±0.0161
CADENCE	**0.0222**	**0.0626**	**0.0608** ±0.0024	**0.2413** ±0.0121

Table 26: Routing accuracy / MAE at H=20 by encoder (all unconditional), three seeds; chance routing 0.333 (BM1, BM2) and 0.50 (BM6).

Encoder	BM1	BM2	BM6
None, z-scored	0.35 / 13.12	0.36 / 0.041	0.53 / 0.712
Whitening	0.36 / 12.87	0.37 / 0.040	0.52 / 0.705
Marg. Gaussian copula	0.64 / 7.93	0.71 / 0.033	0.66 / 0.584
RealNVP to $\mathcal{N}(\mathbf{0},\mathbf{I})$	0.89 / 2.51	0.90 / 0.028	0.80 / 0.471
PF-ODE	**0.92 / 2.19**	**0.93 / 0.026**	**0.83 / 0.452**

Table 27: BM1 component ablation at H=20: each block varies one setting with all others at the default (single training seed, so no dispersions). Bold: best per metric within each block.

Axis	Variant	MAE	SW_2
Experts K	K=1	3.201	0.110
	K=2	2.493	0.078
	K=3 (default)	**2.032**	**0.073**
	K=4	2.261	0.079
	K=5	2.621	0.083
Training size N	N=100	12.04	0.165
	N=500	6.678	0.133
	N=1,000	3.844	0.104
	N=2,000	3.175	0.095
	N=5,000 (default)	**2.032**	**0.073**
ODE width	d_h=32	2.781	0.080
	d_h=64 (default)	**2.032**	**0.073**
	d_h=128	2.312	0.086
ODE depth	3 layers (default)	2.032	0.073
	5 layers	**1.541**	**0.058**
Ghost / aux. latent	Neither (default)	**2.032**	**0.073**
	$\mathbf{z}_{aux}$ only	2.467	0.077
	Ghost + $\mathbf{z}_{aux}$	2.673	0.081
Obs. noise σ	σ=0.01	**2.016**	**0.069**
	σ=0.05 (default)	2.032	0.073
	σ=0.10	2.489	0.085
	σ=0.20	2.170	0.070

means differ slightly; on those five seeds routing moves by less than 0.01 and MAE by less than 8%, with no change of ordering, and the routing conclusion rests on the 0.92-vs-0.34 gap, far outside either dispersion.

The two factors separate cleanly. **The compressor handles noise and dimensionality**: BM4 observes a low-dimensional latent through a random nonlinear map into p=500 with additive noise, so a compact representation helps any temporal model, and with the PF-ODE on, PCA and VAE beat the uncompressed linear map. **The PF-ODE is the component routing depends on, and its effect is a switch**: within every compressor family enabling it improves MAE by 24–31% and SW_2 by 20–36%, but without it routing falls to $0.348/0.343/0.346$, within one standard error of the $1/3$ chance level (0.016 at n=900), and no compressor recovers any part of it. The collapse is not an initialisation artefact: the FiLM maps start at zero but the expert basis does not, so they receive gradient from the first step (App. H.1), and re-running the PF-ODE-off arm with A_s, A_b initialised at $\mathcal{N}(0, 10^{-2})$ gives routing 0.349 on BM4 and 0.352 on BM1 (three seeds) and MAE within 2% of the zero-initialised arm, while the expert heads separate ($\|\mathbf{w}_k - \mathbf{w}_l\|$ grows by the same factor as with the PF-ODE on). The model still forecasts without the PF-ODE, 32–46% worse in MAE, but it forecasts a *population*, assigning units to experts at chance, the ensemble-level regime the framework positions itself against. The VAE rules out the cheapest explanation: its objective already drives the aggregate latent toward $\mathcal{N}(\mathbf{0}, \mathbf{I})$, and routing still collapses, so scale normalisation alone does not account for it. Routing accuracy and forecast quality are two axes rather than proxies, and the PF-ODE is the only factor that moves the first.

Table 28: BM4 compressor × PF-ODE, dynamics fixed, three seeds; routing chance 0.333 (s.e. 0.016). "uncond.": the score network receives a zeroed context (architecture unchanged); the router still receives it.

Compressor, PF-ODE	MAE	SW_2	Routing
Linear, on	0.059 ±0.011	0.237 ±0.015	0.920 ±0.015
Linear, on, uncond.	0.070 ±0.010	0.249 ±0.010	0.90 ±0.03
Linear, off	0.078 ±0.031	0.298 ±0.043	0.348 ±0.002
PCA, on	0.024 ±0.009	0.063 ±0.008	0.931 ±0.011
PCA, off	0.035 ±0.013	0.098 ±0.011	0.343 ±0.003
VAE, on	0.020 ±0.008	0.054 ±0.015	0.936 ±0.012
VAE, off	0.028 ±0.021	0.074 ±0.026	0.346 ±0.002

Gaussian pinning, not context-conditioning. The score network is the only component upstream of the router that sees the context, so removing the PF-ODE removes both Gaussian pinning and encoder-side conditioning. Keeping the PF-ODE but zeroing its score network's context input (architecture and parameter count unchanged), routing stays at 0.90 ± 0.03 (vs. 0.92 ± 0.02 conditional), MAE 0.070 ± 0.010, SW_2 0.249 ± 0.010; on BM1 the same arm gives 0.92 ± 0.03 (vs. 0.96 ± 0.02 over three ablation seeds; App. I.5 compares it with the 93.7% of the Fig. 2 run), MAE 2.20 ± 0.35, SW_2 0.078 ± 0.018. Routing therefore depends on the Gaussianising transport itself, with encoder-side conditioning contributing a few points at most. Tab. 2's BM4 row is the PCA-10 variant, comparable with the PCA-on row here; its five-seed routing accuracy is 0.923 (Fig. 10), inside this table's three-seed 0.931 ± 0.011. The factorial covers BM4 over three seeds and the unconditional-score arm adds BM1; neither is repeated on BM7.

I.5 COA′ Corruption Sweep (BM1)

COA′ (Assumption 2) requires that static context identifies the dynamical regime. To measure what happens when it fails, we sever the context→regime link on BM1 for a random fraction η of units: each selected unit has its two informative context dimensions resampled from the marginal, so its context stays in-distribution but no longer identifies its regime, and we retrain end-to-end at each level (η=0 clean, η=1 fully severed; three seeds). Tab. 29 reports two independent run sets: (A) the sweep, and (B) models whose Bernoulli(η) corruption mask is

recovered after training so that routing is scored separately on corrupted and intact units of the same model (the η=0 model of set B has an empty mask and accuracy 0.975, seed 0).

Subset scoring states the COA′ mechanism directly. On intact units routing stays at 0.970–0.983 at every η, against 0.975 uncorrupted; on corrupted units it is 0.346–0.363, slightly above chance. The router does not mis-route units whose context still identifies their regime; it loses exactly those whose context no longer does, consistent with COA′ and with Step 3 of Theorem 1, where the leaf assignment $\lambda(i) = g^*(c_i)$ depends on context alone. Aggregated, accuracy follows the null with $a_0 = 0.975$ to within 0.02 at every level. Run set A likewise follows it with $a_0 = 0.931$ (predicted $0.782/0.632/0.483$ at $\eta = 0.25/0.50/0.75$, observed $0.806/0.661/0.473$), to within 0.04; its η=0 row sits slightly above the headline BM1 numbers (MAE 2.073 vs. 2.05; SW_2 0.082 ± 0.007 vs. 0.070 ± 0.003). BM1 routing accuracy varies across independently retrained models between 0.93 and 0.97: 0.937 for the Fig. 2 run, 0.931 ± 0.041 here, 0.975 in set B, 0.96 ± 0.02 for the conditional arm of App. I.4 (three seeds) and 0.973 at N=5,000 in App. I.3. Routing entropy rises monotonically as more units lose informative context, so the router signals its loss of confidence rather than committing to a wrong archetype.

Table 29: COA′ corruption on BM1 at H=20, mean ± s.d. over three seeds, chance $1/K = 0.333$. (A) routing accuracy, routing entropy, MAE and SW_2; (B) routing accuracy of an independent run set scored on the corrupted (masked) and intact units of the same model, overall, and the null $(1-\eta)a_0+\eta/K$ with $a_0 = 0.975$.

	(A) Sweep				(B) Subset-scored routing			
η	Routing	Entropy	MAE	SW_2	Masked	Intact	Overall	Null
0.00	0.931 ±0.041	0.222 ±0.034	2.073 ±0.270	0.082 ±0.007	–	0.975	0.975	0.975
0.25	0.806 ±0.077	0.324 ±0.051	2.978 ±0.297	0.097 ±0.012	0.363 ±0.005	0.983 ±0.002	0.834 ±0.003	0.814
0.50	0.661 ±0.008	0.539 ±0.078	4.125 ±0.576	0.123 ±0.024	0.351 ±0.002	0.978 ±0.002	0.667 ±0.001	0.654
0.75	0.473 ±0.009	0.544 ±0.127	5.205 ±0.167	0.146 ±0.013	0.346 ±0.002	0.970 ±0.002	0.498 ±0.001	0.494
1.0	0.371 ±0.018	0.601 ±0.075	6.839 ±0.111	0.285 ±0.087				

Bounded degradation. Whether routing error propagates boundedly through the forward integration has the structural answer of App. B.10: routing weights lie on the simplex, so $w_i = W\alpha_i \in \mathrm{conv}\{w_1, \ldots, w_K\}$ for *every* routing and the parameter discrepancy between two routings is at most $\max_{j,k} \|w_j - w_k\|$, however wrong the routing; with the field Lipschitz in w this gives a trajectory discrepancy bounded over any finite horizon, with a Grönwall constant growing with the local expansion rate (the near-bifurcation amplification BM5, with its overlapping context distributions and chaotic subgroup, measures in isolation). At η=1 MAE rises 3.3× and SW_2 3.5×, comparable factors: a collapse into an ensemble-level transport flow would instead hold the population metric while the individual one degraded. The fully corrupted model still beats every *context-blind* cross-sectional baseline of Tab. 2 on both metrics (MAE 6.84 vs. 14.82 for OT-CFM; SW_2 0.285 vs. 0.318), while the context-conditioned CVFM and COT-FM retain lower SW_2 (0.075/0.095). One comparison is unfavourable: at η=1 it is worse than CADENCE's own unrouted Latent-ODE on the same latents (3.25, Tab. 22), so pure-noise context costs more than no routing. Under COA′ this is expected: an uninformative but in-distribution context produces misassignment with raised entropy, whereas removing routing produces a single averaged field. **Scope:** the diagnostic needs ground-truth regimes, so it is a synthetic-data instrument; on real data COA′ must be argued from domain knowledge a priori, and the model cannot certify its own assumption.

I.6 Scaling with Additional Observations

The headline setting is the hardest one available, a single snapshot per unit. Where sparse follow-ups $\mathbf{x}^i_{t_m}$ at $t_m > t_i$ exist, Eq. 10 is the maximum-likelihood objective under isotropic Gaussian noise in latent space (the frozen bijective Enc_ψ contributes a log-Jacobian constant in the Stage-2 parameters, so it is an exact latent MSE):

$$\mathcal{L}_{\text{semi-long}} = \tfrac{1}{N} \textstyle\sum_{i,m} \left\| \mathrm{Enc}_\psi(\mathbf{x}^i_{t_m}; \mathbf{c}_i) - \hat{\mathbf{z}}^i_{t_m} \right\|^2, \tag{10}$$

which reduces to the standard Latent-ODE loss under dense follow-ups. BM1–BM6 are retrained with three follow-ups per unit, architecture, encoder and hyperparameters unchanged (Tab. 30). Both metrics improve on every benchmark, so the objective converts additional observations into accuracy. The gain is 28–36% on BM1–BM5 and 5% on BM6, where much of each unit's remaining error is stochastic and irreducible from past observations (App. I.7 shows that the per-unit latent state, not the routed regime, carries most of the remaining structure). The main tables use single snapshots, the setting the framework is designed for.

Table 30: Three follow-up observations per unit (3) against one snapshot (1, repeating Tab. 2); H=20 unless marked. Follow-up values are means over test trajectories, ± s.d. across three training seeds.

	MAE		SW_2		3, H=40	
BM	1	3	1	3	MAE	SW_2
BM1	2.05	**1.405** ±0.189	0.070	**0.060** ±0.002	1.915 ±0.249	0.067 ±0.005
BM2	0.025	**0.016** ±0.002	0.050	**0.042** ±0.003	0.019 ±0.003	0.047 ±0.004
BM3	0.097	**0.070** ±0.006	0.050	**0.043** ±0.003	0.082 ±0.006	0.049 ±0.010
BM4	0.0222	**0.0153** ±0.0008	0.0626	**0.0517** ±0.0029	0.0155 ±0.0010	0.0526 ±0.0034
BM5	0.103	**0.072** ±0.008	0.137	**0.116** ±0.004	0.119 ±0.007	0.138 ±0.009
BM6	0.434	**0.412** ±0.056	0.139	**0.124** ±0.004	0.445 ±0.067	0.134 ±0.005

I.7 Skill Against the Leaf-Conditional Mean

Two shortcuts could give low MAE without per-unit content: predicting a subgroup's mean, or holding the observed state. The leaf-conditional mean trajectory $\bar{\mathbf{x}}_t^\lambda$, the best predictor that knows the leaf but not the unit's state, controls for the first and the persistence forecast $\hat{\mathbf{x}}_t^i = \mathbf{x}_{t_i}^i$ for the second:

$$\mathrm{SS}_\lambda \;=\; 1 - \frac{\sum_{i\in\lambda}\sum_t \|\mathbf{x}_t^i - \hat{\mathbf{x}}_t^i\|^2}{\sum_{i\in\lambda}\sum_t \|\mathbf{x}_t^i - \bar{\mathbf{x}}_t^\lambda\|^2}, \qquad \bar{\mathbf{x}}_t^\lambda = \frac{1}{|\lambda|}\sum_{j\in\lambda}\mathbf{x}_t^j.$$

$\mathrm{SS}_\lambda = 0$ means the subgroup is captured and nothing individual; 1 is exact per-unit recovery. Tab. 31 reports it for CADENCE per leaf, together with the Spearman correlation between predicted and true trajectories *within* each leaf, insensitive to subgroup identity by construction, and persistence, which calibrates the unnormalised MAE scales of Tab. 2. BM7's per-unit targets are not per-unit: a cell cannot be re-observed, so its row is scored at the clone level against the per-clone later-day mean-expression aggregates of all 986 clones, each scored by the fold that held it out (App. F.8); persistence holds each clone's day-2 observation fixed and the two leaves are the ground-truth fates. Every other BM7 metric is per cell.

Table 31: CADENCE's skill against the leaf-conditional mean (range across leaves; held-out units, full horizon), against the context-blind cohort mean (uniformly higher, so the leaf null is the stricter), within-leaf Spearman, and MAE as a share of persistence (H=20). Baselines: Tab. 41. †One continuous leaf family: the null is the cohort mean, a weaker null not comparable with the other rows.

BM	Leaves	Skill	vs. cohort	Spearman	vs. persist.
BM1	3	0.47–0.73	0.88	0.59–0.74	13.1%
BM2	3	0.51–0.76	0.85	0.53–0.82	8.9%
BM3†	1	0.96	0.96	0.99	5.3%
BM4	3	0.95–0.97	0.99	0.90–0.95	2.6%
BM5	3	0.52–0.76	0.84	0.58–0.74	11.8%
BM6	2	0.66–0.83	0.91	0.86–0.88	16.2%
BM7	2	0.59–0.65	0.79	0.70–0.73	13.0%

Table 32: MAE at H=20 of baselines without a dynamics model (ratio to CADENCE); on BM4 all receive the PCA-10 pre-compressor.

BM	CADENCE	Smoother	$(\mathbf{c},\mathbf{z})$-NN*
BM1	2.05	15.830 (7.7×)	2.710 (1.3×)
BM2	0.025	0.077 (3.1×)	0.041 (1.6×)
BM3	0.097	1.236 (12.7×)	0.341 (3.5×)
BM4	0.0222	0.0901 (4.1×)	0.0634 (2.9×)
BM5	0.103	0.402 (3.9×)	0.178 (1.7×)
BM6	0.434	1.824 (4.2×)	0.702 (1.6×)

*Joint $(\mathbf{c},\mathbf{z})$ nearest neighbour, time-aligned (oracle access).

BM3's skill of 0.96 is against a weaker null (no subgroup structure to credit), so read within that scope the row isolates per-unit content; its Spearman value also benefits from the large dynamic range of relaxation oscillations. BM2's two statistically indistinguishable archetypes are merged into one leaf, so its skill is over three leaves; the merged entry pools the two archetypes' within-archetype errors and omits the between-archetype term a single pooled leaf mean would add to the null, so it is lower than a pooled null would give. Skill is positive on every leaf of every benchmark, and the leaf-conditional mean dominates the cohort mean in squared error within every leaf, so these skills are conservative against any weaker null. Skill is highest on BM4, where a 500-dimensional observation of a five-dimensional latent makes a subgroup average an especially poor individual predictor and the PCA-10 encoder removes the observation noise that would otherwise mask per-unit structure; the uncompressed configuration scores 0.72–0.83 on the same null, a property of the representation rather than the dynamics model. It is next highest on BM6, the other high-dimensional benchmark, which isolates the attribution: its advantage over context-blind baselines might be credited to context-conditioning alone, but 0.66–0.83 against a predictor that already knows the fate shows that the per-unit latent state carries most of the remaining structure. Paired Wilcoxon signed-rank tests over per-unit errors reject equality against both persistence and the leaf mean on every benchmark with $p < 10^{-4}$; the largest are 6.7×10^{-5} (BM5, leaf mean) and 1.7×10^{-16} (BM7, leaf mean), all others below 10^{-36}.

Baselines without a dynamics model. Two non-parametric baselines test whether the accuracy could come from context-conditional smoothing of the observed marginals or from local state matching (Tab. 32, scored exactly as Tab. 2): (a) a context-conditional temporal smoother, $\hat{\mathbf{z}}_i(t) \propto \sum_j K(\mathbf{c}_i,\mathbf{c}_j)\,\kappa_h(t_j - t)\,\mathbf{z}_j$ over training units with Gaussian kernels and validation-tuned bandwidths, using cross-sectional data only; (b) a joint $(\mathbf{c},\mathbf{z})$ nearest neighbour in a whitened metric forecasting the nearest training unit's ground-truth future, aligned by time since enrolment, which reads training units' futures and so has the oracle block's access. Neither has an ODE, router or learned field. CADENCE's MAE is 3.1–12.7× below the smoother and 1.3–3.5× below the neighbour, with the largest margins on BM3, whose difficulty is per-unit phase inference.

I.8 E1: Mixing-Weight Recovery on a Synthetic Routing Benchmark

Routing accuracy measures K-class recovery; E1 tests *individual-level* recovery of per-unit mixing weights. K archetypes are fixed; each unit's parameter is a known blend $w_i = W_{\mathrm{true}}\alpha_i^{\mathrm{true}}$ with $\alpha_i^{\mathrm{true}} \sim \mathrm{Dirichlet}(0.1)$; the data come from a frozen FiLM-ODE teacher handed to the model, removing dynamics as a confound so that

Table 33: E1: per-unit mixing-weight recovery. Simplex distance: per-unit $\|\hat{\alpha} - \alpha^{\text{true}}\|_2$; basis direction against the true archetype directions ($\approx 78^\circ$ for random directions in $\mathbb{R}^{16}$; E1's basis uses d=16, departing from the benchmarks' d_w=14 of Tab. 8, where the expected random-direction angle is 77.7°); rollout MSE against the frozen teacher.

K	r_s	$\kappa(B)$	$\|\hat{\alpha} - \alpha^{\text{true}}\|_2$	corr$(\hat{\alpha}, \alpha^{\text{true}})$	basis direction ($\sin\theta$, angle)	rollout MSE
2	8	2.4	0.061 $\pm$0.025	0.999 $\pm$0.001	0.389 $\pm$0.117 (22.9°)	0.0023 $\pm$0.0012
4	8	3.5	0.315 $\pm$0.022	0.940 $\pm$0.006	0.679 $\pm$0.029 (42.8°)	0.0093 $\pm$0.0006
8	8 (default)	22.6	0.507 $\pm$0.083	0.634 $\pm$0.079	0.739 $\pm$0.070 (47.6°)	0.0207 $\pm$0.0055
8	32 (control)	2.4	0.450 $\pm$0.074	0.725 $\pm$0.052	–	0.0104 $\pm$0.0023

only the router is learned, and the archetype count equals the expert count in every row, so representation is exact and only routing can fail. The router trains through the dynamics objective alone; labels enter only at evaluation to resolve the expert permutation (Hungarian alignment, App. C.4); held-out units, three seeds. Because $r_s = 8$ is held fixed while K grows, the map $B \in \mathbb{R}^{8\times K}$ embedding α into the context becomes worse-conditioned, so part of the K=8 degradation could reflect how hard α is to read off c rather than the router; the last row re-runs K=8 with $r_s = 32$, restoring $\kappa(B) = 2.4$ (the K=2 value) with everything else identical (Tab. 33).

The correlation is the conventional statistic; the per-unit simplex distance is the individual-level reading, since a correlation across units is a population statistic that Dirichlet(0.1)'s near-binary entries flatter. At K=2 the weights are recovered essentially exactly on a simplex of diameter $\sqrt{2}$. The control localises the K-dependence: the simplex error barely moves ($0.507 \to 0.450$, within one seed σ), so coordinate recovery degrades only mildly with K, while the larger movements in correlation and rollout MSE ($0.634 \to 0.725$; $0.0207 \to 0.0104$) are partly an artefact of the ill-conditioned context at r_s=8. **Archetype directions** are recovered above chance at every K (22.9° at K=2, 47.6° at K=8, against roughly 78° at random) but not exactly. The residual is weak separation, not a representational limit: over the sampled initial conditions and finite horizon the flow is only weakly sensitive to w in these directions, so a small loss discrepancy admits a larger parameter discrepancy, i.e. the separation modulus of the QFOA analysis (App. B.9) is small. Proposition 4 identifies the mechanism: directions in which Δh is close to divergence-free with respect to ρ_0^λ transport the baseline into itself and are poorly determined; one finite-horizon run cannot distinguish a small modulus from an exact symmetry. The routing weights and the trajectories are recovered, and the archetype vectors as far as the dynamics distinguish them. Every benchmark in the paper uses $K \leq 3$.

I.9 E2: Continuous-Parameter Recovery on BM3

BM3 is the only continuous leaf family: $\mu_i \sim \mathrm{U}(0.5, 3.0)$ sets the dynamics and is the unit's sole static context, with the initial condition independent. To test the graded-coordinate claim of MRA (Assumption 3) we run a $K \in \{2, 4, 8\}$ study in a *separate sweep configuration* using the discrete SMoE router (the reported BM3 result uses the continuous context-to-parameter router of App. D.3). Each expert is localised to a scalar μ_k by the α-weighted centroid of the true μ on the training split only, so the readout carries no evaluation-set information; we form $\hat{\mu}_i = \sum_k \alpha_{i,k}\mu_k$ and score held-out units (Tab. 34). **The hard-assignment ceiling.** Under one-hot α the readout is bin-centroid quantisation of μ; for a uniform variable equal-width bins are optimal and maximising correlation means minimising within-bin variance, $L^2/(12K^2)$ against a total $L^2/12$ on an interval of length L, so

Table 34: E2 on BM3 (SMoE sweep). All K experts are active (2/2, 4/4, 8/8). Last column: the hard-assignment ceiling $\rho_{\max}(K)$.

K	Pearson$(\hat{\mu}, \mu)$	Spearman	ceiling
2	0.961 $\pm$0.011	1.000	0.866
4	0.997 $\pm$0.003	1.000	0.968
8	0.999 $\pm$0.001	1.000	0.992

$$\rho_{\max}(K) \;=\; \sqrt{1 - \tfrac{\mathrm{Var}(\mu|\mathrm{bin})}{\mathrm{Var}(\mu)}} \;=\; \sqrt{1 - \tfrac{1}{K^2}} \tag{11}$$

bounds any hard-assignment readout. The observed correlation exceeds it at every K by many standard deviations, and Spearman $= 1.000$ says the same: hard assignment would collapse $\hat{\mu}$ onto K values with large tie groups, whereas a perfect rank correlation needs N distinct correctly ordered values.

Scope and role. Because μ is BM3's context, this shows that the simplex coordinates form a faithful, genuinely graded coordinate system for a continuous leaf parameter, the object Step 3 of the identifiability argument needs, not inference of an unobserved parameter from dynamics; nothing in training uses μ as a target, so a monotone continuous encoding is not forced, and the coordinates vary monotonically with μ. Recovery saturates by K=2 and per-unit MAE is flat in K, as a one-dimensional family predicts: from K=2 the convex mixture already traces a continuous path. This is the split of Theorem 1: forecast quality is driven by the per-unit latent state z_0, untouched by this sweep, while correlation does identifiability work on the leaf-shared axis $w(\lambda)$; with discrete leaves the same reading predicts an interior optimum in K, no gain above K=$|\Lambda|$ and positive error below, matching the BM1 minimum at K=3=$|\Lambda|$ of App. I.3. The two results are the continuous and discrete halves of one statement. The headline BM3 number (Tab. 2; Tab. 8) is not this sweep but the continuous

router, so it already reflects graded recovery of μ. The empirical payoff is the BM3 exception of App. I.10: full trajectories improve the conditional-transport baselines (COT-FM $0.524 \rightarrow 0.227$) without letting them overtake CADENCE, which, on snapshots alone, still leads by more than twofold. We attribute that gap to CADENCE integrating one trajectory from the encoded state: it separates *which* limit cycle a unit occupies (the parameter μ, recovered here) from *where* on it the unit sits (phase, carried in z_0), a decomposition a sampler conditioned on $(\mathbf{c}, t)$ cannot reproduce and one also conditioned on the snapshot reproduces only at the query time, not along a path.

I.10 Input-Matched Conditional-Transport Baselines

Methods and adaptation. The three conditional-transport baselines span the conditional-generative families: CVFM (Generale et al., 2026), conditional flow matching with amortisation across continuous conditioning variables and unpaired data, the closest available match to our setting; COT-FM (Kerrigan et al., 2024), conditional OT via simulation-free flows; and cMonge (Driessen et al., 2026), conditional OT maps. COT-FM and OT-CFM are different methods: the former conditions on an external covariate, the latter on the interpolation pairing (Kerrigan et al., 2024). The natural adaptation conditions on (static context, query time) and targets the state, so each learns to sample $P(x_t \mid c_i, t)$, with the authors' published grids and selection criteria where available, on identical data, splits, horizon and metric (also for BM6). cMonge predicts a source sample plus a learned conditional shift, has no continuous-time variant, and was built for single-cell perturbation response.

Full-trajectory control arm: the two flow-based baselines are also trained with the oracle baselines' advantage, full trajectories in place of snapshots (Tab. 35).

Table 35: Full-trajectory control arm for the conditional-transport baselines: CVFM and COT-FM trained on full trajectories against CADENCE (CAD.) on one snapshot per unit; $H{=}20$, mean $\pm$ s.d. over five seeds (cMonge has no continuous-time variant). Bold: best per row. BM4: every method receives the PCA-10 pre-compressor.

	MAE			SW_2		
BM	CAD.	CVFM	COT-FM	CAD.	CVFM	COT-FM
BM1	2.050	1.056 ±0.148	**0.444** ±0.102	0.070	0.029 ±0.008	**0.027** ±0.009
BM2	0.025	**0.005** ±0.001	**0.005** ±0.001	0.050	**0.028** ±0.003	0.031 ±0.003
BM3	**0.097**	0.276 ±0.052	0.227 ±0.043	0.050	**0.032** ±0.005	0.038 ±0.004
BM4	0.0222	**0.0122** ±0.0007	0.0135 ±0.0008	0.0626	**0.0099** ±0.0006	0.0117 ±0.0009
BM5	0.103	0.082 ±0.018	**0.068** ±0.035	0.137	**0.071** ±0.032	0.073 ±0.027
BM6	0.434	**0.415** ±0.049	0.473 ±0.052	**0.139**	0.163 ±0.022	0.182 ±0.029

Given trajectories these methods improve by factors of 4–31 on BM1, BM2 and BM4 (3–5$\times$ on BM5) and overtake CADENCE's MAE on BM1, BM2, BM4 and BM5; CVFM narrowly overtakes it on BM6 (COT-FM does not). The arm measures how much of the cross-sectional gap is an information gap. On BM6 full trajectories move CVFM from 0.722 to 0.415 and COT-FM from 0.822 to 0.473, far less than on BM1, BM2 and BM4, and CADENCE keeps the lowest SW_2 at either access level; the BM6 MAE margin (CVFM 0.415 ± 0.049 vs. 0.434) is smaller than one seed dispersion, so it is an ordering, not a tested difference. **BM3 is the exception**, the only benchmark here where CADENCE keeps lower MAE than both full-trajectory methods: full trajectories improve them substantially (COT-FM $0.524 \rightarrow 0.227$) but CADENCE, on snapshots alone, still leads by more than twofold. BM3's intended challenge (App. F) is separating *which* limit cycle a unit occupies, a conditioning problem these methods handle well, from *where* on it the unit sits, a state-inference problem a sampler conditioned only on $(\mathbf{c}, t)$ cannot address; conditioned on the unit's snapshot it can (CVFM-S reaches 0.168 against CVFM's 0.424) but re-infers the phase independently at every query time. CADENCE carries phase in the encoded latent state and period in the routed parameters. The BM3 dispersions are wide relative to the differences among the transport methods, so the comparison is against CADENCE, not among them (the empirical consequence of the μ-recovery of App. I.9). **Coupling choice:** the independent-coupling CVFM leads COT-FM in nine of the ten cross-sectional cells on BM1–BM5, tying in the tenth; under full-trajectory access the ordering reverses on BM1 and on BM5's MAE, and within dispersion on BM3's MAE, consistent with optimal couplings preferring non-crossing assignments while trajectories in these systems cross.

The state-conditioned adaptation. The adaptation above conditions on $(\mathbf{c}_i, t)$, the conditioning set these methods are published with; nothing stops a sampler from also conditioning on the unit's own snapshot, and whether the capability of Tab. 1 belongs to the model class or to the conditioning set is decided by that variant. CVFM-S conditions on $(\mathbf{c}_i, t_i, \mathbf{x}^i_{t_i}, t)$, is trained cross-sectionally by pairing each unit's snapshot with the kernel-smoothed later marginals of its context neighbours, never with its own future, and is otherwise identical to CVFM (Tab. 36).

Three things follow. *First*, state-conditioning closes 58–86% of the MAE gap on five benchmarks and ties CADENCE on BM2 (0.024 vs. 0.025, within one seed s.d.), where OT-CFM has the lower MAE; most of the MAE gap to the conditional-transport family in Tab. 2 is therefore the conditioning set, not the model class (that table compares the methods as published; CVFM-S is the strongest cross-sectional adaptation we constructed). *Second*, CVFM-S remains behind on MAE and SW_2 on five of six and on skill on all six, by 0.14 to 0.69, most on BM1 and BM5, where it does no better than the leaf-conditional mean (-0.08, 0.06):

conditioning on the snapshot recovers the state at t_i, but a sampled future from an oscillator or a bistable system lands in the wrong phase or basin often enough to cost more than the mean; on BM3 (0.71 against 0.96), built so that identifying the limit cycle and locating the unit on it are separate problems, it re-infers the phase independently at every query time. *Third*, that is measurable, and it is where the families separate qualitatively: CVFM-S's increment error is 1.27–1.38 times its marginal error, just below the $\sqrt{2}$ of independent draws and short of it by what the shared conditioning explains, while CADENCE's is 0.22–0.41 because its successive states lie on one integrated trajectory. The gap, a factor of 3.1 to 6.3, is structural: no conditioning makes a per-time-point sampler produce a coherent path, and this, not per-time-point accuracy, is what Tab. 1's persistent-dynamics column names; CVFM-S is competitive for the state at a fixed query time but not temporally coherent. **Deterministic context-conditioned transport.** The cheapest mechanism-free account of CADENCE's forecasts is a context-conditioned transport integrated as one path from the unit's own state; c-OT-CFM is exactly that, keeping CADENCE's latents, context and rollout and replacing the routed basis and per-subgroup MMD by one conditional flow-matching field. It integrates one trajectory, so its increment ratio is low (0.31–0.52), far below CVFM-S: persistence comes from integration, not routing. It is lower than CADENCE on BM2 (MAE 0.023, skill 0.68), within 9–17% on BM6 and BM4, and 30–37% behind on BM1, BM3 and BM5, where its skill is 0.36, 0.91 and 0.44 against 0.61, 0.96 and 0.65: the routed parameters with the per-subgroup objective matter most where units follow qualitatively different or parameter-indexed dynamics and are a modest gain on BM4 and BM6.

Table 36: State-conditioned controls at $H{=}20$. CVFM conditions on $(\mathbf{c}, t)$; CVFM-S adds the unit's own snapshot. Skill against the leaf-conditional mean (App. I.7). Increment error: MAE of $\hat{\mathbf{x}}_{t+\delta} - \hat{\mathbf{x}}_t$ over consecutive grid steps divided by the marginal MAE; independent draws per time point give $\sqrt{2} \approx 1.41$, an integrated trajectory much less. ‡Our adaptation, not a published method: an OT-CFM velocity field $v(\mathbf{z}, t, \mathbf{c})$ on CADENCE's Stage-1 latents, rolled out deterministically from each unit's encoded state (five seeds).

	MAE						Skill					
	BM1	BM2	BM3	BM4	BM5	BM6	BM1	BM2	BM3	BM4	BM5	BM6
CVFM	11.97	0.031	0.424	0.0508	0.260	0.722	−12.30	0.24	0.13	0.22	−1.23	0.27
CVFM-S	3.41	0.024	0.168	0.0281	0.169	0.508	−0.08	0.52	0.71	0.74	0.06	0.56
c-OT-CFM‡	2.81	**0.023**	0.126	0.0259	0.138	0.471	0.36	**0.68**	0.91	0.93	0.44	0.66
CADENCE	**2.05**	0.025	**0.097**	**0.0222**	**0.103**	**0.434**	**0.61**	0.66	**0.96**	**0.96**	**0.65**	**0.75**
	SW_2						Increment error / marginal error					
CVFM-S	0.088	**0.047**	0.058	0.0713	0.158	0.171	1.33	1.29	1.38	1.31	1.35	1.27
c-OT-CFM‡	0.079	0.049	0.055	0.0684	0.149	0.158	0.46	0.38	0.31	0.35	0.52	0.49
CADENCE	**0.070**	0.050	**0.050**	**0.0626**	**0.137**	**0.139**	**0.30**	**0.34**	**0.22**	**0.28**	**0.37**	**0.41**

J Additional Controls, Ceilings and Sensitivity Analyses

J.1 Assumption Stress Tests: the Two Theory Bridges, ENR and POS

Two rows of Tab. 37 concern the path from the observable law to the parameter conclusion: the backbone θ is learned rather than fixed, and the loss matches pooled per-expert marginals while Theorem 1 uses per-context FOA. We cross the two on BM1 and BM5 (Tab. 38; three seeds, headline encoder and configuration). *Fixed* θ: the backbone is frozen at the value of an independently trained seed, and $\mathbf{W}$ and the router are refitted. *Per-context matching*: the MMD of Eq. 3 uses the product kernel $k_\sigma(\mathbf{z}, \mathbf{z}')\, k_{\sigma_c}(\mathbf{c}, \mathbf{c}')$, with σ_c set by the median heuristic over contexts, so that the loss compares the predicted and reference laws jointly in $(\mathbf{z}, \mathbf{c})$ and so per context. Parameter recovery is the archetype-recovery error e_{arch} of App. J.10, fitted to the induced field on BM1 (Lotka–Volterra) and BM5 (Duffing).

Table 38: The two theory bridges on BM1 and BM5, three seeds: backbone θ learned or fixed, loss matched pooled or per context; *headline configuration, re-run on three seeds. Rout.: mean per-class recall; e_{arch}: archetype-recovery error of App. J.10 (lower is better).

	BM1			BM5		
θ, matching	MAE	Rout.	e_{arch}	MAE	Rout.	e_{arch}
learned, pooled*	2.06 ±0.30	0.936	0.086	0.104 ±0.011	0.842	0.127
learned, per context	2.12 ±0.27	0.941	0.079	0.107 ±0.012	0.846	0.119
fixed, pooled	2.09 ±0.31	0.935	0.083	0.105 ±0.011	0.841	0.124
fixed, per context	2.14 ±0.28	0.939	0.076	0.108 ±0.012	0.845	0.116

Both bridges cost little and move in the direction the analysis predicts. Per-context matching lowers e_{arch} by 0.007–0.008 on both benchmarks, about one seed standard deviation, and raises MAE by 3%, because splitting the reference by context thins it; fixing θ changes every entry by less than the seed dispersion. The practical question behind $\mathrm{FOA_j}$ is whether equally good pooled fits land on different parameters. Among the 15 fits per benchmark (five seeds, three restarts) whose validation MMD lies within one standard error of the best, the fitted archetype parameters differ pairwise by at most 3.1% on BM1 and 4.4% on BM5 in relative norm, against recovery errors of 8.6% and 12.7%: in these runs the pooled objective does not admit distant equally good minimisers. This is evidence on two benchmarks, not a proof that the trained estimator is consistent.

Violating ENR and POS. ENR and POS hold on BM1–BM6 by construction, so the analysis predicts degradation only when they are broken. We break each on BM1 (Tab. 39), three seeds, with everything else fixed.

ENR: enrolment depends on the unit's own state, $P(t \mid \mathbf{z}_{t_0}, \mathbf{c}) \propto \exp(\beta\, \bar{z}_{t_0})$ on the window, with $\bar{z}_{t_0}$ the standardised mean latent coordinate. *POS*: enrolment is tied to context, $t = \tau(\mathbf{c}) + \epsilon$ with τ varying within each leaf and ϵ of shrinking spread, so that each context is observed over an ever shorter interval while each leaf still covers the whole window.

Table 39: ENR violated (strength β) and POS approached ($\mathrm{sd}(\epsilon)$ as a fraction of the window; u.: uniform enrolment) on BM1, three seeds.

	ENR, β				POS, $\mathrm{sd}(\epsilon)$			
	0	0.5	1	2	u.	0.25	0.10	0.02
MAE	2.06	2.19	2.63	3.71	2.06	2.24	2.78	4.46
Rout.	0.936	0.934	0.921	0.887	0.936	0.935	0.933	0.931
e_{arch}	0.086	0.097	0.141	0.232	0.086	0.104	0.163	0.308

Both violations degrade parameter recovery and forecasts monotonically, so the two assumptions are doing the work the analysis assigns them rather than being formal conditions the method ignores. The two failures differ in the way the analysis predicts. Under ENR the conditional law at each time is state-selected, the objective fits a biased family, and routing degrades with it. Under POS routing is unaffected, since it reads context, but each context's family shrinks towards a single time: recovery then rests on pooling neighbouring contexts of the same leaf, which carries it part of the way ($e_{\mathrm{arch}} = 0.104$ at 0.25 of the window) and fails as the per-context interval vanishes. This is the finite-sample role of pooling that Sec. 5 describes: at population level each context is identified from its own law, and pooling is what makes that usable at finite N.

Table 37: Every difference between the class the analysis covers and the system we train, with the measurement that bounds it; three rows are scope exclusions, and the last two are measured in App. J.1. The added enrolment condition $t \perp \mathbf{c} \mid \lambda$ (App. B.4) holds on all seven benchmarks by construction and is not a difference. Full tables in the appendix sections cited.

Analysis assumes	**Deployed system**	**What was measured**
Unconditional encoder P	Context-indexed $P_{\mathbf{c}}$ (BM1–BM6; BM7 runs P, App. H.5)	Zeroing $\mathbf{c}$ in the score net routing $0.92/0.90$ vs. $0.96/0.92$ (BM1/BM4; factorial runs, three seeds), $+7\%/+19\%$ MAE (App. I.4)
One vector field per leaf (DFA)	$h_{\mathrm{base}}(\mathbf{z}, t, \mathbf{c})$ also reads $\mathbf{c}$	Retrained with $h_{\mathrm{base}}(\mathbf{z}, t)$, so context enters only through the router: $+8$–9% MAE, ≤ 0.01 routing, on BM1/BM3/BM4/BM6 (App. J.5)
Hard leaf partition	Soft routing, $\tau_{\mathrm{end}}{=}0.1$	Mean $\max_k \alpha_{ik}$ is 0.81–0.95 across the seven benchmarks; the idealised hard, sample-split estimator is 14% *worse* on BM1 (App. J.6)
Unaugmented state, C^1 open-support baseline	Five scratch coordinates (BM1–BM6)	Unaugmented field within 5% MAE; the separation statistic, recomputed leaf-wise *and* unaugmented, stays ≥ 1.4 orders above a resampling floor (App. J.8)
Pushforward from a canonical t_0	Integration from each unit's own t_i	Measured: training from t_0 instead costs 3.4% MAE on BM1 (App. J.6); the Grönwall lemma of App. B.9 is qualitative only, being exponential in the horizon
Vanishing kernel bandwidth	Fixed $h{=}0.1$	Over $h \in \{0.02, 0.05, 0.1, 0.2\}$, MAE moves $\leq 7.8\%$ and routing ≤ 1.6 points (App. J.7)
Fixed MMD kernel k_σ	σ set per batch by the median heuristic	Replacing it by a single σ fixed on the training split moves MAE by $\leq 2.9\%$ on BM1, BM4 and BM6 (App. J.7)
Lossless compressor E ($q = p$)	PCA-10 in front of the encoder on BM4	Not a cost but a scope exclusion: BM4's headline row lies outside Theorems 1–2, and the uncompressed $q{=}p$ run is reported beside it for every method (Tab. 25)
Known, fixed backbone θ	θ learned jointly with $\mathbf{W}$ and ω	Identifying $(\theta, \{w(\lambda)\})$ needs $\mathrm{FOA}_{\mathrm{j}}$ (App. B.3). Freezing θ at an independently trained value and refitting $\mathbf{W}$ and the router: MAE within 2%, routing within 0.002, archetype-recovery error 0.083 vs. 0.086 on BM1 (App. J.1)
Discrete leaves served by the SMoE router	BM3: a continuous context-to-parameter router (App. D.3)	Not a cost but a scope exclusion: BM3's data assumptions hold, but the reported model is outside the SMoE class the analysis covers
POS and a continuous observation window	BM7: every cell enrols at culture day 2; later days are reference populations	Not a cost but a scope exclusion: POS fails, so Theorem 1 is not claimed on BM7 (App. B.4)
Per-context FOA	Loss matches per-expert pooled marginals	The pooled loss can admit a minimiser the per-context statement excludes (non-implication remark, App. B.9). Replacing it by an MMD with a product kernel on $(\mathbf{z}, \mathbf{c})$, which matches per context: MAE $+3\%$, routing $+0.005$, archetype-recovery error 0.079 vs. 0.086 on BM1, same pattern on BM5 (App. J.1)

J.2 Point-Optimal Scoring of the Conditional Baselines

For a conditional sampler a single draw is not the point-optimal predictor: under componentwise absolute error the optimum is the conditional median, under squared error the conditional mean, so the MAE columns of Tab. 2 score each conditional model's sampled transport rather than its best point prediction. We re-evaluate CVFM, COT-FM and cMonge with 128 draws per unit, scoring the componentwise conditional median against MAE and the conditional mean against SW_2 (Tab. 40); a 512-draw run on BM4 changes the median-of-128 MAE by 0.4%, so 128 draws have converged. Point-optimal scoring improves the samplers' MAE by 1.26–$1.30\times$ at both horizons, inside the range a calibrated sampler predicts: for a Gaussian conditional law single-draw error is $2\sigma/\sqrt{\pi}$ and point-optimal error $\sigma\sqrt{2/\pi}$, a ratio of $\sqrt{2} \approx 1.41$, and bias that does not shrink with averaging pulls the realised ratio below it; cMonge does not move, as a deterministic map must. No benchmark's leader changes: CADENCE keeps the lowest cross-sectional MAE on BM1 and BM3–BM6 under both scorings; on BM2, where OT-CFM leads, CVFM's point-optimal value (0.024) also falls below CADENCE's (0.025), placing CADENCE third there; at $H{=}40$ no leader changes either, and on BM2 CVFM's point-optimal value (0.027) stays above CADENCE's (0.024). SW_2 moves the other way, by 2.0–$2.3\times$ on every benchmark: collapsing a conditional law onto its point predictor improves individual accuracy at the cost of distributional fidelity, so the two metrics separate individual from population accuracy and point-optimal scoring trades one for the other.

Table 40: Point-optimal scoring of the conditional-transport baselines against the single-draw values of Tabs. 2 and 12 (1: single draw; opt.: median of 128 draws for MAE, conditional mean for SW_2). cMonge applies one deterministic Monge map and is the control: its 128 draws are identical, so both scorings coincide and it is printed once.

	$H{=}20$						$H{=}40$					
	CVFM		COT-FM		cMonge	CADENCE	CVFM		COT-FM		cMonge	CADENCE
BM	1	opt.	1	opt.			1	opt.	1	opt.		
MAE (median of 128 draws)												
BM1	11.97	9.24	13.97	10.83	13.41	**2.05**	20.37	15.91	23.38	18.12	18.57	**2.91**
BM2	0.031	0.024	0.036	0.028	0.162	0.025	0.034	0.027	0.045	0.035	0.164	**0.024**
BM3	0.424	0.334	0.524	0.413	0.581	**0.097**	0.927	0.719	0.928	0.737	0.926	**0.102**
BM4	0.0508	0.0397	0.0624	0.0488	0.0941	**0.0222**	0.0786	0.0624	0.0690	0.0543	0.1018	**0.0235**
BM5	0.260	0.206	0.358	0.284	0.604	**0.103**	0.401	0.308	0.399	0.312	0.693	**0.174**
BM6	0.722	0.573	0.822	0.652	1.283	**0.434**	0.797	0.628	0.887	0.682	1.334	**0.497**
SW_2 (conditional mean)												
BM1	0.075	0.168	0.095	0.209	0.412	**0.070**	0.092	0.195	0.109	0.238	0.471	**0.078**
BM2	0.054	0.112	0.062	0.128	0.704	**0.050**	0.066	0.135	0.065	0.137	0.753	**0.047**
BM3	0.053	0.121	0.053	0.118	0.431	**0.050**	0.066	0.148	0.061	0.123	0.457	**0.053**
BM4	0.0801	0.1658	0.0869	0.1782	0.2604	**0.0626**	0.0933	0.1941	0.0923	0.1975	0.3149	**0.0628**
BM5	0.162	0.331	0.187	0.379	0.831	**0.137**	0.197	0.428	0.201	0.442	0.853	**0.154**
BM6	0.206	0.412	0.224	0.445	0.471	**0.139**	0.223	0.448	0.249	0.513	0.530	**0.148**

J.3 Skill of the Baselines Against the Leaf-Conditional Mean

Tab. 41 scores every published cross-sectional baseline against the same leaf-conditional-mean null as Tab. 31; none exceeds 0.30 on any benchmark.

The pattern is structural. OT-CFM integrates every unit through one shared population field with no per-unit parameters, so against a null that already knows each unit's leaf mean it can gain only through the unit's own encoded starting state; where the leaves' dynamics diverge it does far worse than that null (on BM1 its error sits at the persistence level, and on BM3 a shared field cannot recover phase). CVFM and COT-FM condition on context, so they pick up the between-leaf but not the within-leaf variance: they reach 0.09–0.27 where that suffices (BM2–BM4, BM6) and fall below the null on BM1 and BM5, where a sampled future from an oscillator or a bistable system lands in the wrong phase or basin often enough to cost more than the mean. TIGON and scNODE, context-blind like OT-CFM, and cMonge, one Monge map per context, fall below the null on every benchmark except BM4, whose leaf-conditional mean is a weak predictor of a 500-dimensional observation. No published baseline exceeds 0.30 on any benchmark; the state-conditioned adaptation of App. I.10 reaches at most 0.74, still below CADENCE everywhere. The unrouted shared field (App. I.1) reaches MAE 3.25 on BM1, below every tuned sequence-trained oracle, yet skill 0.13; routing takes the same representation to 0.61

Table 41: Skill against the leaf-conditional-mean null (App. I.7), pooled over leaves, $H{=}20$; 0: the subgroup captured and nothing individual; negative: worse than that null. †Cohort-mean null (one continuous leaf family). CADENCE's pooled values lie inside its per-leaf ranges of Tab. 31.

Method	BM1	BM2	BM3†	BM4	BM5	BM6
OT-CFM (cross-sect.)	−19.38	0.11	−7.36	0.08	0.04	0.09
COT-FM	−17.11	0.20	0.09	0.17	−3.23	0.23
CVFM	−12.30	0.24	0.13	0.22	−1.23	0.27
TIGON	−31.24	−5.35	−7.66	0.18	−8.55	−1.65
scNODE	−28.83	−8.34	−4.31	0.05	−1.25	−1.05
cMonge	−15.69	−13.28	−0.44	0.12	−11.04	−1.18
Shared field, latents	0.13	0.17	0.10	0.15	0.11	0.19
CADENCE	**0.61**	**0.66**	**0.96**	**0.96**	**0.65**	**0.75**

Table 42: Measured anchors (oracle floors on BM5, BM6; supervised reference on BM7) and CADENCE against them.

BM	Quantity	Anchor	CADENCE	relative
BM5	MAE	0.081 ±0.006	0.103	1.27× floor
BM6	MAE	0.331 ±0.048	0.434	1.31× floor
BM6	SW_2	0.112 ±0.009	0.139	1.24× floor
BM7	Acc., cell	0.700 ±0.024	0.645	92% of ref.
BM7	AUROC, cell	0.85 ±0.02	0.82	91% above chance
BM7	Acc., clone	0.769 ±0.021	0.690	90% of ref.
BM7	AUROC, clone	0.87 ±0.02	0.85	95% above chance

and from chance-level routing to 0.937. **Why skill and MAE rank the baselines differently.** A method that misroutes a minority of units makes a few very large errors, which MAE down-weights and squared-error skill does not, so context-blind methods and samplers look better on MAE than on skill: on BM2 OT-CFM has the lower MAE and a sixth of CADENCE's skill. The converse is impossible: root-mean-square error is at least mean absolute error for any error distribution, so a method's skill is bounded above by what its MAE permits given CADENCE's own error shape; every cell here and in Tab. 36 lies below that bound, and the negative cells sit at it, the largest value consistent with the method's MAE.

J.4 Measured Anchors on BM5, BM6 and BM7

BM5, BM6 and BM7 carry irreducible stochasticity, so an absolute error or accuracy on them is uninterpretable alone; Tab. 42 measures what is achievable. On BM5 and BM6 the anchor is an oracle forecaster given each unit's *true* regime parameter $w(\lambda)$ and *true* initial latent state, so the only error left is the benchmark's own stochasticity: CADENCE sits at $1.27\times$ and $1.31\times$ the achievable MAE and $1.24\times$ the achievable SW_2 on BM6, so most of its residual error there belongs to the stochastic gene-regulatory process, not the model; on BM6 every baseline's SW_2 is at least 17% higher and 9 of 14 exceed three times CADENCE's. On BM7 the anchor is a supervised reference, not a ceiling: a classifier given the day-2 context, the full day-4 and day-6 expression dynamics and the fate labels, fitted by nested five-fold cross-validation inside each training fold and evaluated on the same cross-fitted cells. Per clone CADENCE's 69.0% is 90% of what it reaches with the labels and the expression dynamics, neither of which CADENCE uses. A BM7 accuracy materially above it would warrant a leakage check; it is what one supervised classifier attains, not an upper bound on every method. The reference and CADENCE are consistent under a binormal model: AUROC 0.87 implies a best-threshold accuracy near 0.79, just above the reference's 0.769, whereas CADENCE's AUROC of 0.85 implies 0.77 against the 0.690 its untuned $\arg\max$ threshold attains, so the router's discrimination exceeds what its decision rule realises.

J.5 The Theory-Matched Configuration: Context-Free h_{base}

DFA (Assumption 1) says units in a leaf share a dynamical *parameter*. The deployed field $h_\theta(\mathbf{z}, t, \mathbf{c}; \mathbf{w}) = s(\mathbf{w}) \odot h_{\mathrm{base}}(\mathbf{z}, t, \mathbf{c}) + b(\mathbf{w})$ passes $\mathbf{c}$ into h_{base} as well, so units in a leaf share a parameter but not, strictly, a vector field; the surrogate-supervision mechanism is stated for shared dynamics, and the per-context and leaf-marginal FOA forms of App. B.3 coincide only where h_θ is $\mathbf{c}$-independent within a leaf. The theory-matched configuration uses $h_{\mathrm{base}}(\mathbf{z}, t)$, so context reaches the dynamics only through the router, everything else identical (Tab. 43). It costs 8.3–9.3% of MAE and at most 0.009 of routing accuracy; the second number matters: the router already carries essentially all context information into the dynamics, and h_{base}'s $\mathbf{c}$ input is a second, largely redundant pathway. This configuration removes the second row of Tab. 37 at a cost under 10% (the other rows still apply); the headline keeps the better-forecasting $\mathbf{c}$-conditioned field.

Table 43: Theory-matched configuration: context enters only through the router, so units in a leaf share a vector field exactly as Assumption 1 states; headline → $\mathbf{c}$-free, three seeds (headline values repeat Tab. 2).

BM	MAE	SW_2	Routing
BM1	2.05 → 2.22 ±0.33 (+8.3%)	0.070 → 0.076 ±0.004	0.937 → 0.929
BM3	0.097 → 0.106 ±0.013 (+9.3%)	0.050 → 0.054 ±0.002	n/a
BM4	0.0222 → 0.0241 ±0.0013 (+8.6%)	0.0626 → 0.0679 ±0.0037	0.923 → 0.916
BM6	0.434 → 0.472 ±0.101 (+8.8%)	0.139 → 0.151 ±0.009	0.850 → 0.841

Table 44: $\max_k \alpha_{ik}$ at test time; below 0.65 a unit is genuinely interpolated, above it effectively hard-assigned. †The $K{=}3$ SMoE run of App. I.9 (the reported BM3 model has no expert weights).

BM	K	mean $\max_k \alpha_{ik}$	share < 0.65
BM1	3	0.95	1.5%
BM2	3	0.94	1.8%
BM3†	3	0.91	4.6%
BM4	3	0.95	1.2%
BM5	3	0.88	6.3%
BM6	2	0.86	7.5%
BM7	2	0.81	11.8%

J.6 Routing One-Hotness, Hard Assignment and Canonical-t_0 Integration

Empirical degree of router concentration. The routed parameter $\mathbf{w}_i = \sum_k \alpha_{ik} \mathbf{w}_k$ is a convex combination, but $\mathbf{w} \mapsto \Phi^{\mathbf{w}}$ is not convex, so interpolating between archetypes can in principle cross a bifurcation and produce dynamics matching neither, a specific concern on BM5, whose subgroups straddle a homoclinic bifurcation. Tab. 44 measures whether that regime is occupied. The router is strongly peaked, and a cross-check holds: a symmetric $K{=}3$ weight vector with $\max_k \alpha = 0.95$ has entropy $-0.95 \ln 0.95 - 2(0.025 \ln 0.025) = 0.23$, against the mean routing entropy of 0.222 at $\eta{=}0$ in Tab. 29. Fewer than 2% of units on BM1, BM2 and BM4 are interpolated, so the bifurcation concern is contained there; the share is largest where it should be, 6.3% on BM5 and 11.8% on BM7, where $K{=}2$ and the fates genuinely overlap in context space. Integrating the $\alpha{=}0.5$ field on BM5 gives an attractor matching neither archetype, as expected; those 6.3% of units carry 9% of BM5's squared error, slightly more than their share, the mechanism behind BM5's position nearest the identifiability limit. **Hard, sample-split assignment and canonical-t_0 integration.** The consistency argument of App. B.9 analyses hard, sample-split leaf assignments, and the identifiability argument pushes forward from

a canonical t_0 while $\mathcal{L}_{\text{cross}}$ integrates each unit from its own t_i. On BM1 hard sample-split assignment gives MAE 2.34 ± 0.24 against the soft-routed 2.05, so the estimator the theory analyses is 14% *worse* than the one we deploy; integrating from the canonical t_0 gives 2.12 ± 0.22, a 3.4% cost. The Grönwall bound of App. B.9 is not evaluated numerically; this arm measures the gap it bounds.

J.7 Integration Steps, the Evaluation Grid and Kernel Bandwidths

Two time discretisations appear in this paper and they are independent. The *evaluation grid* is a property of each dataset: 40 points spanning $[t_0, t_{\text{end}}]$, shared by every method and unit and stored with the benchmark. A horizon H is a number of grid steps, and a unit contributes only the grid points strictly after its own enrolment, so a late-enrolled unit is scored on fewer points rather than forecast past the window. One grid step is $(t_{\text{end}} - t_0)/40$: 1.5 on BM1, 2.0 days on BM2, 0.5 on BM3, 1.0 on BM4, 1.25 on BM5, 0.025 of pseudotime on BM6 and 0.1 of a culture day on BM7, which is why $H{=}20$ and $H{=}40$ are meaningful on a window $[0, 1]$. The *integration* step is a property of the solver: we fix the step *count* at $N_{\text{int}}{=}120$ across the window, because a fixed absolute Δt would give BM1 120 steps and BM6, whose window has unit length, two. Dopri5 is fifth-order, so the choice is not delicate. Re-integrating the same trained models at other step counts, nothing else changed (Tab. 45: the smooth BM1, chaotic BM5 and stochastic unit-window BM6, the case a fixed absolute step would discretise worst), halving the count costs 0.5% on BM1 and 2.1% on BM6 and doubling it changes nothing beyond the third digit, so 120 is converged; at $N_{\text{int}}{=}30$ the two least smooth benchmarks lose 7.8% and 9.7%, and a fixed absolute step of 0.5 would give BM6 two steps, far coarser still.

Table 45: Integration-step sensitivity at $H{=}20$, same trained models re-integrated; $N_{\text{int}}{=}120$ reported.

N_{int}	30	60	**120**	240
BM1 MAE	2.09	2.06	**2.05**	2.05
BM5 MAE	0.113	0.106	**0.103**	0.103
BM6 MAE	0.468	0.443	**0.434**	0.433

Table 46: Temporal-kernel bandwidth; h/span relative to the observation windows (60, 20, 40, 1.0 native units for BM1, BM3, BM4, BM6). Headline $h{=}0.1$ in bold.

h	BM1 MAE	BM1 routing	BM3 MAE	BM4 MAE	BM6 MAE
0.02	2.21	0.921	0.104	0.0239	0.451
0.05	2.09	0.933	0.099	0.0228	0.438
0.10	**2.05**	**0.937**	**0.097**	**0.0222**	**0.434**
0.20	2.19	0.928	0.103	0.0237	0.462
h/span	BM1 0.0017, BM3 0.0050, BM4 0.0025, BM6 0.100				

Spatial MMD bandwidth. The analysis fixes k_σ; the implementation sets σ per batch by the median heuristic, which makes it data-dependent. A single σ fitted on the training split and held fixed changes MAE by $+2.9\%$ on BM1 ($2.05 \to 2.11$), $+1.8\%$ on BM4 ($0.0222 \to 0.0226$) and $+2.3\%$ on BM6 ($0.434 \to 0.444$), and routing by under half a point; the deployed models keep the heuristic, and this arm bounds the difference. **Temporal kernel bandwidth.** Lemma 4 takes $h \to 0$ with $N_\lambda h \to \infty$; the deployed $h{=}0.1$ is fixed across benchmarks whose spans run from 1.0 to 80.0, so h/span varies by two orders of magnitude (Tab. 46). Across a tenfold range of h, MAE moves by at most 7.8% (BM1) and routing by at most 1.6 points; the response is flat-bottomed, and $h{=}0.1$ sits at or next to the optimum on all four benchmarks although their spans differ by a factor of 80. It was fixed on BM1's validation split before any test evaluation (validation MAE 2.24, 2.15, 2.11, 2.16 at $h = 0.02$, 0.05, 0.1, 0.2) and kept for every benchmark; the table is a test-set sensitivity analysis, not the selection. Because the sweep stays below 0.34% of BM1's span, we also swept $h/\text{span} \in \{0.005, 0.02, 0.05, 0.1\}$ on BM1 and BM5 (three seeds): MAE rises monotonically, 2.31, 2.74, 3.38, 4.12 on BM1 and 0.108, 0.121, 0.149, 0.187 on BM5, so wider kernels blur the reference across the dynamics and the headline configuration is not noise-limited. The reference is not built per leaf, which would need leaf labels: at each query time t the $M = 128$ reference units are drawn from the training units enrolled within $\pm 3h$ of t (all of them when fewer qualify), disjoint from the mini-batch's predicted set, and each enters every expert's reference with weight $\alpha_{j,k}\,\kappa_h(t_j - t)$ (Algorithm 1); on BM7 the reference is the pooled later-day population weighted by the coupling of Eq. 4. The per-step reference is small: its effective sample size $(\sum_j a_j)^2 / \sum_j a_j^2$ with $a_j = \alpha_{j,k}\kappa_h(t_j - t)$, per query time and expert, has median 6.2 on BM1, 4.6 on BM2, 11.1 on BM4, 7.4 on BM5, 34.0 on BM6 and 57.6 on BM7, and 49.6 per query time on BM3's continuous router. Each step's estimate is therefore noisy, and the objective averages it over $n_q = 8$ query times per step and over training; moving h from 0.02 to 0.2 changes these sizes by up to a factor of five and MAE by at most 7.8%. **Hutchinson probe count.** Tab. 5's noise floor uses 2,048 Rademacher probes; recomputing at 512 and 8,192 moves the minimum margin by ± 0.30 orders, exactly the $\log_{10} 2$ that a floor scaling as $1/\sqrt{\text{probes}}$ predicts per factor of four (the statistic uses exact divergences, so the margin is a function of the probe count, not a property of the model alone); at 512 the minimum margin is still 2.1 orders. This dependence is why the reported margin is taken against the resampling floor, which does not depend on the probe count. **Expert-index alignment on a held-out fold.** The Hungarian permutation of App. C.4 is fitted on a validation fold and applied to the test fold; fitting it on the test fold instead selects the *same* permutation on all seven benchmarks, so every reported routing accuracy is unchanged (BM1 0.937, BM2 0.947, BM4 0.923, BM5 0.843, BM6 0.850 and, per cell, BM7 0.645): the alignment fixes a labelling and contributes nothing to the numbers. For the $K{=}2$ benchmarks only two permutations exist, so the failure mode is binary and its absence informative.

J.8 Per-Leaf and Unaugmented Separation Statistic

Tab. 5's headline statistic uses a baseline $\hat{\rho}_0$ pooled over each benchmark's test units, although the statement it supports is leaf-level, and is evaluated on the augmented state including the five scratch coordinates, whose baseline is a Dirac at zero and so lacks the C^1 open-support density Proposition 4 assumes. Its stricter columns fix both. Leaf-wise density estimation uses roughly $1/K$ of the samples, raising the floor by about $\sqrt{K}$, and dropping the scratch coordinates removes some signal but proportionally less noise; together they cost 0.6–0.7 orders, and the resampling floor, which replaces the Hutchinson estimator's probe noise with a resampling estimate of the density error, is the more conservative of the two floors. Every expert pair on every benchmark stays above one order of magnitude under all four evaluations, so the conclusion is unchanged and only its margins shrink. The remaining limits stand: the condition is sufficient and not necessary, and a check between K converged heads is local rather than a verification over all of Θ_w.

J.9 Non-Degeneracy of the Cross-Sectional Objective, Per Leaf

Table 47 gives, per benchmark and leaf, the between-bin MMD^2 against the within-bin split-half floor behind the $119\times$ minimum quoted in App. B.9. One leaf in total, the smaller of BM7's two fates, has a split-half floor whose bootstrap interval includes zero, so no ratio is defined for it; the $119\times$ figure quoted in the main text is the minimum over the sixteen measurable leaves, rounded down from 1.2×10^2. The marginals move by orders of magnitude more than the sampling noise on every benchmark, so the objective is not matching a time-invariant target.

Table 47: Between-bin MMD^2 relative to the within-bin split-half floor, per benchmark, with a bootstrap 95% interval on the ratio. A leaf is *excluded* when the bootstrap interval on its floor includes zero, so the ratio is not defined.

BM	Measurable leaves	Ratio range across leaves	95% interval, smallest ratio	Excluded
BM1	3 of 3	$8.4 \times 10^2 - 5.0 \times 10^4$	$[6.3,\ 11.2] \times 10^2$	0
BM2	3 of 3	$2.7 \times 10^2 - 1.1 \times 10^4$	$[2.1,\ 3.5] \times 10^2$	0
BM3	1 of 1	1.2×10^2	$[1.0,\ 1.5] \times 10^2$	0
BM4	3 of 3	$6.1 \times 10^2 - 2.3 \times 10^4$	$[4.6,\ 8.0] \times 10^2$	0
BM5	3 of 3	$1.9 \times 10^2 - 3.4 \times 10^3$	$[1.5,\ 2.4] \times 10^2$	0
BM6	2 of 2	$1.2 \times 10^2 - 4.0 \times 10^2$	$[1.0,\ 1.5] \times 10^2$	0
BM7	1 of 2	1.4×10^2	$[1.1,\ 1.8] \times 10^2$	1

J.10 Expert Count: Forecast Error and Parameter Recovery Are Different Axes

Forecast error is shallow in K above $K{=}1$, so it cannot by itself locate the expert count the analysis predicts. On BM1 it is in fact minimised at $K{=}3$, the same K the recovery error picks, in the single-seed sweep of App. I.3; but it varies only $1.29\times$ over $K \in \{2, \ldots, 5\}$, so that coincidence is not what carries the claim. The analysis does not predict a forecast-error minimum at $K = |\Lambda|$; it predicts that the archetype *parameters* become identifiable there. Scoring the BM1 K-sweep for parameter recovery as well separates the two (Tab. 48).

Table 48: BM1 ($|\Lambda|{=}3$): forecast error against archetype-parameter recovery across the expert-count sweep of App. I.3. The archetype-recovery error e_{arch} is defined below; it is computed from the K converged expert heads and the generator's archetype parameters, and no per-unit routing weight enters it. The oracle-routing row replaces the learned router by the ground-truth leaf assignment and retrains, so routing accuracy is the best attainable at each K.

K	1	2	**3**	4	5
MAE	3.201	2.493	**2.032**	2.261	2.621
archetype-recovery error	0.556	0.291	**0.081**	0.104	0.131
Routing accuracy	0.333	0.664	**0.973**	0.921	0.869
oracle routing (ground-truth leaf assignment, retrained)					
archetype-recovery error	0.549	0.276	**0.074**	0.074	0.075
Routing accuracy	0.333	0.667	1.000	1.000	1.000

The statistic. BM1's generator gives archetype k a Lotka–Volterra parameter vector $\vartheta_k \in \mathbb{R}^4$ (the two growth and two interaction rates of App. F.2). A trained expert head k carries a modulation vector $\mathbf{w}_k$, not an LV parameter, so the two are compared through the field they induce: we fit $\hat{\vartheta}_k = \arg\min_\vartheta \int \| D\mathrm{Dec}_\psi(\mathrm{Enc}_\psi \mathbf{x})\, h_\theta(\mathrm{Enc}_\psi \mathbf{x}, t, \bar{\mathbf{c}}_k; \mathbf{w}_k) - f_{\mathrm{LV}}(\mathbf{x}; \vartheta)\|^2 \, d\nu^{\mathrm{dec}}(\mathbf{x})$ over the benchmark's decoded state distribution ν^{dec}, i.e. the decoded velocity field of head k, with the context fixed at the mean context $\bar{\mathbf{c}}_k$ of the units routed to k (the encoder and h_{base} both read it), match $\{\hat{\vartheta}_k\}$ to $\{\vartheta_k\}$ by the Hungarian alignment of App. C.4, and report $e_{\mathrm{arch}} = \frac{1}{|\Lambda|} \sum_\lambda \|\hat{\vartheta}_{\pi(\lambda)} - \vartheta_\lambda\|_2 / \|\vartheta_\lambda\|_2$. When $K > |\Lambda|$ the match is injective from archetypes to experts and the surplus experts are ignored; when $K < |\Lambda|$ several archetypes share an expert. The inputs are the K heads and the generator's parameters. Routing weights appear nowhere in it.

Why it is not a re-expression of routing accuracy. Across the learned-router sweep the two move together, so the statistic could in principle be a restatement of routing purity. The oracle-routing rows separate them, and they separate the curve into two halves with different causes. Below $K = |\Lambda|$ the error is representational: at $K{=}2$ it is 0.276 even though the oracle assignment is the best any router could produce, because two heads cannot carry three archetypes. Above $K = |\Lambda|$ it is routing: with the oracle assignment the surplus experts receive no mass and recovery is flat at 0.074–0.075 for $K \in \{3, 4, 5\}$, against 0.081, 0.104 and 0.131 under the learned router,

so the degradation at $K>3$ is the router fragmenting a leaf across heads rather than a second representational effect. At identical routing accuracy (1.000) the three oracle-routing points have near-identical recovery, while three learned-router points at the same K spread monotonically: recovery error carries information beyond routing purity.

Beyond one benchmark and one seed. The sweep above is single-seed and test-scored, so it separates the two axes but cannot establish a minimum. On the three discrete-leaf benchmarks with $|\Lambda| = 3$, over three seeds and scored on a validation split, the archetype-recovery error is 0.549 ± 0.021, 0.297 ± 0.018, $\mathbf{0.086 \pm 0.009}$, 0.109 ± 0.012, 0.138 ± 0.015 on BM1 for $K = 1, \ldots, 5$; 0.512 ± 0.026, 0.274 ± 0.022, $\mathbf{0.094 \pm 0.011}$, 0.118 ± 0.014, 0.146 ± 0.019 on BM2; and 0.601 ± 0.034, 0.338 ± 0.029, $\mathbf{0.127 \pm 0.016}$, 0.151 ± 0.018, 0.183 ± 0.022 on BM5. The minimum is at $K = |\Lambda| = 3$ on all three, in every seed, and the gap to $K = 4$ is between 1.5 and 2.6 seed standard deviations. BM5's floor is the highest of the three, which is what a benchmark built next to a bifurcation should give. **Selecting K without labels.** The sweep above scores recovery against the generator's parameters, which a practitioner does not have. Selecting $K \in \{1, \ldots, 5\}$ instead by the smallest K whose validation $\mathcal{L}_{\text{cross}}$ is within one seed s.d. of the minimum picks $K = 3$ on BM1, BM2 and BM4 in all three seeds, $K = 2$ on BM6 in all three, and on BM5 $K = 3$ in two seeds and $K = 4$ in one, where the overlapping contexts make a fourth head marginally useful; on BM7 the label-free held-out-day Sinkhorn criterion picks $K = 2$. The ground-truth count is therefore recoverable without labels on every discrete-leaf benchmark except in one BM5 seed.

The result matches the analysis: the archetype parameters become identifiable at $K = |\Lambda|$, and adding experts beyond it does not improve them. The minimum is sharp on this axis, a factor of 3.6 between $K{=}2$ and $K{=}3$, where MAE moves by $1.23\times$ over the same step.

Over $K \in \{2, \ldots, 5\}$, MAE varies by a factor of 1.29 while archetype-recovery error varies by 3.6 and has a sharp minimum at $K = |\Lambda| = 3$. Assumption 3 and App. B.5 concern parameter identification, not forecast accuracy. The two curves share a minimiser on BM1 but not a shape: only the recovery curve locates it sharply, so forecast error is not a usable proxy for parameter identification.

K Extended Discussion

This section supplies two boundary conditions that need more than the one line of Sec. 9; the other limitations are stated there.

K.1 Data-generating assumptions

The latent dynamics are a neural ODE, an assumption assessed empirically and most strained on BM6 and BM7, whose stochastic processes depart furthest from it. Each observation must be a state on a continuous time-parameterised trajectory, excluding event-based progression models (Fonteijn et al., 2012), and the enrolment time must be independent of the unit's state (ENR, Sec. 3), true by construction on BM1–BM6 and by BM7's fixed culture-day design, but violated by cohorts with symptomatic enrolment.

K.2 Complete data at enrolment

The benchmarks use complete observations at enrolment. Masked denoising score matching could handle missing entries, but under a mask the encoder is bijective only on the observed coordinates, and the identifiability argument of Sec. 5 would need revisiting.