UNPAIRED DOMAIN-TO-DOMAIN TRANSPORT WITH DYNAMICAL ALIGNMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

For decades, numerical models have been central to forecasting and understanding physical systems. Yet unresolved scales, simplified physics, and uncertain parameters make them imperfect. Over long trajectories, model states and observations therefore lie on different attractors, described by different invariant measures. Relating these measures is a transport problem for which generative models provide expressive maps. Standard transport objectives, however, rely on geometric proximity and may pair states whose futures are dynamically incompatible. We introduce a framework for transporting invariant measures while aligning their dynamics. We define a discrepancy that quantifies how well two systems can be aligned up to an admissible change of coordinates. We then bound the map-dependent commutation residual by a pairwise ground cost independent of the transport map, reducing dynamical alignment to a linear optimal transport problem. From unpaired trajectories, intertwining Extended Dynamic Mode Decomposition (iEDMD) learns this cost by aligning reduced representations of the two dynamics. The resulting coupling supervises a Flow Matching model without requiring integration of the transport during training. The learned map can then be wrapped around an unchanged numerical solver in a transport–solve–transport procedure. Experiments on chaotic systems and fluid flows show that the dynamical cost improves conjugation when geometry and dynamics disagree; when the systems already share their coordinates, geometric costs or simpler maps remain preferable.

1 INTRODUCTION

Numerical models, whether physical solvers (Bauer et al., 2015) or learned emulators (Pathak et al., 2022; Lam et al., 2023; Rasp et al., 2024), are used to forecast systems that they describe only approximately. Unresolved scales, simplified physics or inaccurate parameters can give a model dynamics that differ from those of the observed system. In data assimilation, this discrepancy has been linked to initialisation shocks (Mulholland et al., 2015) and to a gradual drift of forecasts toward the model's own attractor (Volpi et al., 2021). An analysis that is accurate for the real system may thus be poorly suited to the model that propagates it. The mismatch can be particularly marked when the model places coherent structures differently from the real system, as reported for the separation latitude of the Gulf Stream in ocean models (Chassignet and Marshall, 2008; Chassignet and Xu, 2017). In such situations, the best estimate of the true state need not be the most suitable initial condition for the model.

This misalignment appears in the long-term statistics: the model and the observed system sample different invariant measures, μ_X and μ_Y. Relating them is a transport problem, arising whenever one seeks to correct a model's statistical bias or to ask what a simulated state—a hurricane, for instance—would correspond to in the observed system. Optimal transport (Villani, 2009; Peyré and Cuturi, 2019) and generative transport (Rezende and Mohamed, 2015; Song et al., 2021; Lipman et al., 2022; Tong et al., 2023) offer scalable ways to construct such maps and have proved effective for unpaired translation between image domains (Zhu et al., 2017). Because the two measures are known only through unpaired samples, any method must decide which source state x_i to pair with which target state y_j.

The usual choice is geometric: pair states that are close. For dynamical systems, however, closeness is a poor guide. Nearby states may have different futures, while a change of coordinates may separate

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.

states that evolve alike. We instead pair states by their dynamical compatibility. This matters most when geometry and dynamics disagree: the systems are observed in different coordinates, or the model places coherent structures elsewhere, so nearby states need not correspond. When the systems already share their coordinates and differ only by a parameter, the identity or a simple affine map may suffice; we use these cases as controls.

We therefore seek a transport that matches the distributions and respects the dynamics. Let $\mathcal{F} : X \to X$ denote the source (model) dynamics and $\mathcal{M} : Y \to Y$ the target (observed system) dynamics over the same time interval. Besides transporting μ_X onto μ_Y, the map $T : X \to Y$ should make the following diagram approximately commute:

$$\begin{array}{ccc} X & \xrightarrow{\mathcal{F}} & X \\ {\scriptstyle T}\downarrow & & \downarrow{\scriptstyle T} \\ Y & \xrightarrow{\mathcal{M}} & Y \end{array} \qquad T \circ \mathcal{F} \approx \mathcal{M} \circ T.$$

In words, evolving then transporting should give the same result as transporting then evolving. When T is invertible and the relation is exact, T is a conjugacy: the two systems are identical up to a change of coordinates.

This requirement is map-dependent: evaluating the residual $T \circ \mathcal{F} - \mathcal{M} \circ T$ requires applying T both to states and to their successors. When T is the flow of a learned velocity field, as in Flow Matching, each evaluation requires integrating that field. The residual can be used as a penalty, but training must then differentiate through the transport, losing the simulation-free regression of Flow Matching. We instead move the dynamics into the ground cost. We bound the commutation residual of any map by the integral of a pairwise cost $c(x, y)$ evaluated on observed states and their successors, independently of T. Minimising this bound over couplings is a standard linear OT problem. We learn the required alignment, and hence the cost, from unpaired transitions with intertwining Extended Dynamic Mode Decomposition (iEDMD). The resulting OT plan provides dynamically consistent pairs from which Flow Matching learns the transport.

Once learned, the map can be used around the unchanged model. An observation is transported into the model's state space, advanced by its solver, and transported back. This provides a practical test of the alignment. It also suggests judging a model up to a change of coordinates, and not only by its direct forecasts. Figure 1 illustrates this point on Lorenz-63 observed through increasingly nonlinear coordinates, so that the two systems remain exactly conjugate. The direct forecast error and the Wasserstein distance between the invariant measures both grow with the deformation, whereas the discrepancy we define, over a sufficiently rich class of maps, stays at its floor; it grows only when the dynamics themselves change (Section 4.1). A model with poor forecasts may therefore still be close to the observed system once its states are correctly identified.

Contributions.

- **A discrepancy modulo changes of representation.** We measure how well two systems, with their invariant measures, can be aligned by a map from an admissible class Ω. The usual criterion, comparing forecasts from the same state, is the special case $\Omega = \{\mathrm{Id}\}$; richer classes reveal that a model with poor direct forecasts can still be close to the observed system (Section 3.1).
- **A dynamics-aware ground cost.** We bound the commutation residual of any transport by a pairwise cost independent of the map, learned from unpaired trajectories with iEDMD. Dynamical alignment then reduces to linear OT, and the resulting plan keeps Flow Matching simulation-free (Sections 3.2–3.4).
- **Conjugated forecasting.** The learned map wraps an unchanged solver: the observation is transported into the model's state space, advanced, and transported back. On chaotic ODEs and fluid flows, including a transport from a full model state to a delay embedding of partial observations, better alignment can yield better forecasts; shared-coordinate controls show when transport is unnecessary or harmful (Section 4).

2 Related work

Optimal transport and dynamical cost. Optimal transport selects a coupling between two measures by minimising a ground cost (Villani, 2009; Peyré and Cuturi, 2019), and entropic regularisation,

solved with Sinkhorn iterations, makes this selection tractable at scale (Knopp and Sinkhorn, 1967; Cuturi, 2013). The resulting coupling can in turn supervise a generative transport: OT-CFM (Tong et al., 2023) trains Flow Matching (Lipman et al., 2022; Albergo and Vanden-Eijnden, 2022) on pairs drawn from it. In such a pipeline, the learned map is only as meaningful as the cost that selected the pairs, which is usually the squared Euclidean distance. When the two spaces are not registered, Gromov–Wasserstein (Mémoli, 2011; Vayer et al., 2019) compares distances within each domain instead of across domains. This removes the need for shared coordinates, but the criterion remains geometric. Ground costs can also be learned, for instance end-to-end through a Monge map estimator (Howard et al., 2024). Our cost is learned as well, but separately from the transport, from transitions observed within each system, and it is motivated by a bound on the commutation residual.

Generative transport and unpaired translation. Most generative models can be read as transports of a reference measure onto data (Kingma and Welling, 2013; Goodfellow et al., 2014; Rezende and Mohamed, 2015; Chen et al., 2018; Ho et al., 2020; Song et al., 2021). When both endpoints are data and no correspondence is observed, many maps transport one measure onto the other, and a prior must select one. Cycle consistency (Zhu et al., 2017) asks the map to be invertible in practice; Schrödinger bridges (De Bortoli et al., 2021) favour paths close to a reference diffusion. Neither uses the temporal evolution of the data. Recycle-GAN (Bansal et al., 2018) does, by requiring that translating, predicting and translating back agree with a per-domain predictor. This is a commutation constraint, enforced as a loss on the generator. We impose it through the cost instead, using observed successors, so that it selects the pairs before the transport is trained.

Koopman operators and learned dictionaries. Koopman operators describe a nonlinear dynamics through its linear action on observables, functions of the state. EDMD approximates this action from data by a matrix acting on a finite dictionary of observables (Rowley et al., 2009; Williams et al., 2015; Brunton et al., 2021; Colbrook, 2023), possibly learned by a neural network (Li et al., 2017; Lusch et al., 2018; Otto and Rowley, 2019; Azencot et al., 2020). iEDMD extends EDMD to two systems: the source predictor is not fitted independently, but expressed through the target predictor and an invertible alignment between the two dictionaries. Fitting each system to its own transitions then yields this alignment without paired data. Koopman representations have already been used to relate systems. They serve to learn a conjugacy from a single system to a simpler one, such as its linearisation (Bevanda et al., 2022) or a normal form (Bramburger et al., 2021), to construct a conjugacy between two given systems by matching their eigenfunctions (Bollt et al., 2018), and to compare systems through their operators (Germain et al., 2026). These approaches start from the dynamics themselves, or from one of them. We only observe unpaired trajectories of both, and measure how far they are from conjugate through a transport, in the spirit of the homeomorphic defect of Skufca and Bollt (2008). At the operator level, conjugacy is an intertwining relation, classical for Markov kernels when the link between the two chains is known (Diaconis and Fill, 1990). Here the link is the unknown.

Model error. In data assimilation, the mismatch between model and system is often addressed by learning a correction to the model (Farchi et al., 2021). We leave the model unchanged and transport its states instead. The two approaches are complementary.

3 Transport with dynamical alignment

3.1 Setting

Let $\mathcal{F} : X \to X$ and $\mathcal{M} : Y \to Y$ denote the model and the observed system, each advanced over the same time lag τ, with invariant measures μ_X and μ_Y. For each system we observe transitions $(x, \mathcal{F}x)$ and $(y, \mathcal{M}y)$, but no correspondence between x and y. Both state spaces are represented in $\mathbb{R}^d$, with coordinates that need not be registered. We describe these dynamics with two families of operators (Koopman, 1931; Williams et al., 2015). The Koopman operators $Uf = f \circ \mathcal{F}$ and $Vh = h \circ \mathcal{M}$ act linearly on observables. Their adjoints, the Perron–Frobenius operators $P^X\nu = \mathcal{F}_{\#}\nu$ and $P^Y\eta = \mathcal{M}_{\#}\eta$, propagate probability measures.

Problem. We seek a map $T : X \to Y$ that transports the invariant measures and commutes with the dynamics,

$$T_{\#}\mu_X = \mu_Y, \qquad T \circ \mathcal{F} \approx \mathcal{M} \circ T. \tag{1}$$

The marginal constraint alone does not determine T: many maps carry μ_X onto μ_Y, and we look for the one that best commutes with the dynamics. Conversely, the constraint may exclude maps that would commute better but distort the distribution. We accept this trade-off and seek the best commutation compatible with the marginals. When T is invertible and the commutation is exact, T is a conjugacy: the two systems coincide up to a change of coordinates. On measures, this reads $T_{\#}P^X = P^Y T_{\#}$: transporting a distribution and then evolving it gives the same result as evolving it and then transporting it.

Commutation residual. For a measure-transporting map T and a discrepancy $d_{\mathcal{P}}$ on $\mathcal{P}(Y)$, define

$$\mathcal{L}(T) := \mathbb{E}_{X\sim\mu_X}\, d_{\mathcal{P}}\big(T_{\#}P^X\delta_X,\ P^Y\delta_{T(X)}\big)\,, \tag{2}$$

where δ_X is a point mass at X. The two arguments describe the outcomes of evolving then transporting, and transporting then evolving. For deterministic dynamics, both are point masses. We compare them using a dictionary of observables $\Psi : X \sqcup Y \to \mathbb{R}^n$, learned in Section 3.3. The feature discrepancy $d_\Psi(\mu,\nu) = \|\mathbb{E}_\mu\Psi - \mathbb{E}_\nu\Psi\|$ gives

$$\mathcal{L}_\Psi(T) = \mathbb{E}_{X\sim\mu_X}\big\|\Psi(T(\mathcal{F}X)) - \Psi(\mathcal{M}(TX))\big\|. \tag{3}$$

Dynamical transport discrepancy. Given a class Ω of admissible maps $X \to Y$, we define

$$\mathcal{D}_\Omega(\mathcal{F}\to\mathcal{M}) := \inf\big\{\mathcal{L}(T) :\ T\in\Omega,\ T_{\#}\mu_X = \mu_Y\big\}, \tag{4}$$

the best commutation achievable within Ω under the marginal constraint. The class Ω specifies which changes of representation are considered harmless. An affine class forgives rotations and rescalings, but not a nonlinear deformation; richer classes forgive more, and enlarging Ω can only decrease $\mathcal{D}_\Omega$. The discrepancy is directed, since the residual is measured in the target space.

An exact measure-preserving conjugacy in Ω gives $\mathcal{D}_\Omega = 0$; the converse requires further conditions, discussed in Appendix A.1.

3.2 From the commutation residual to a linear transport problem

The commutation residual (3) depends on T through $T(x)$ and $T(\mathcal{F}x)$. As discussed in the introduction, minimising it directly would require integrating and differentiating through the learned transport. We look instead for a cost $c(x,y)$ that can be evaluated on a candidate pair of observed states, one source and one target, without knowing T. Once such a cost is fixed, the transport is selected by the Kantorovich problem

$$\min_{\pi\in\Pi(\mu_X,\mu_Y)} \int_{X\times Y} c(x,y)\, d\pi(x,y). \tag{5}$$

This problem is linear in the coupling π, even if c is not linear in x or y. Both endpoints are drawn from the datasets, so no transported state is ever generated: the optimal coupling directly supplies the training pairs for Flow Matching (Section 3.4).

Bounding the residual by a pair cost. To relate 5 to the residual, we bound it by the integral of a pointwise cost against the distribution of its input–output pairs. For any measurable map T, write $\pi_T := (\mathrm{Id}, T)_{\#}\mu_X$, which pairs each x with $T(x)$; when $T_{\#}\mu_X = \mu_Y$, this is a Monge coupling in $\Pi(\mu_X,\mu_Y)$. The key ingredient is an alignment $G : \mathbb{R}^n \to \mathbb{R}^n$ that plays the role of T in feature space: for a good transport, $\Psi(Tx)$ should be close to $G\Psi(x)$.

Proposition 1 (Successor bound)**.** *Assume $\mathcal{F}_{\#}\mu_X = \mu_X$. For any measurable $G : \mathbb{R}^n \to \mathbb{R}^n$ and any measurable T for which the quantities are integrable,*

$$\begin{aligned} \mathcal{L}_\Psi(T) \ &\le \ \int_{X\times Y} c^G_{\mathrm{succ}}(x,y)\, d\pi_T(x,y), \\ c^G_{\mathrm{succ}}(x,y) &:= \|\Psi(y) - G\Psi(x)\| + \|\Psi(\mathcal{M}y) - G\Psi(\mathcal{F}x)\|. \end{aligned} \tag{6}$$

The proof is given in Appendix A.2, as the Monge case of Proposition 3.

Reading the cost. The cost $c^G_{\mathrm{succ}}(x, y)$ is small when y matches the image of x under G, and when the successor $\mathcal{M}y$ matches the image of the successor $\mathcal{F}x$. In other words, x and y correspond now and one lag later. The cost uses only observed states and their observed successors, and never involves T. The bound holds for every G: it requires neither an exact alignment nor one that transports the feature marginals. It is tight when G is the feature-space version of T, $G \circ \Psi = \Psi \circ T$: the first term then vanishes and the second equals the residual (Appendix A.4). We therefore learn G to approximate such an alignment (Section 3.3). The marginal constraint is enforced by the coupling, not by G.

From maps to couplings. For a fixed G, minimising the bound over transport maps means minimising $\int c^G_{\mathrm{succ}}\, d\pi_T$ over Monge couplings. Problem (5) relaxes this to all couplings. When μ_X has no atom and the cost is continuous, the relaxation does not lower the optimal value (Pratelli, 2007). The optimal value of (5) therefore bounds the smallest commutation residual achievable by a transport map. A second gap appears when the map is learned. Flow Matching is trained on a coupling π_B that approximates the solution of (5) (Section 3.4), and reproduces its marginals, but not necessarily the coupling itself. The bound (6) still applies to the learned map, through the coupling it induces; Appendix H.5 measures how closely that coupling follows π_B.

3.3 Learning the feature-space alignment with iEDMD

The successor cost requires a dictionary Ψ and an alignment G such that $G\Psi(x)$ approximates $\Psi(Tx)$ for a good transport. No pairs are available to fit G directly. We rely instead on the property that characterises a conjugacy: seen through G, the source dynamics should look like the target dynamics. We therefore fit both systems jointly with a reduced model in which the target dynamics are represented by a learned operator K_V, and the source dynamics are expressed through G as

$$K_U := G^{-1} \circ K_V \circ G. \tag{7}$$

Each system is fitted only to its own transitions, so no correspondence is needed. We call this extension of EDMD (Williams et al., 2015) *intertwining EDMD* (iEDMD).

Dictionary. We use $\Psi(z) = [\, z;\, \xi(z);\, \varphi_\eta(z)\,] \in \mathbb{R}^n$. The first block is the state itself, which makes Ψ injective and prevents the dictionary from collapsing to a constant. The optional block ξ contains features computed from recorded trajectories, for instance a velocity estimate regressed from finite differences (Appendix I.2); no access to the vector field is required. The last block φ_η is a neural network shared by the two systems: the same features are computed on both sides, and the mismatch between their coordinates is absorbed by G. For low-dimensional representations such as POD coefficients, we simply use $\Psi = \mathrm{Id}$.

Learning the alignment. The *conjugator* G is an invertible matrix or an affine coupling flow, which has an exact inverse (Dinh et al., 2016). By construction (7),

$$G \circ K_U = K_V \circ G, \tag{8}$$

which is the reduced counterpart of the commutation relation $T \circ \mathcal{F} = \mathcal{M} \circ T$. When G is linear, K_U is a matrix similar to K_V; when G is nonlinear, K_U is a nonlinear reduced map rather than an EDMD matrix. We learn Ψ, G and K_V by minimising

$$\begin{aligned}\mathcal{L}_{\mathrm{iEDMD}} = \mathbb{E}_{\mu_X} \|\Psi(\mathcal{F}X) - K_U\Psi(X)\|^2 + \mathbb{E}_{\mu_Y} \|\Psi(\mathcal{M}Y) - K_V\Psi(Y)\|^2 \\ + \lambda\, \mathrm{MMD}^2\big(G_\#\Psi_\#\mu_X,\ \Psi_\#\mu_Y\big).\end{aligned} \tag{9}$$

The second term fits the target dynamics, as in standard EDMD. The first term fits the source dynamics, but through the target operator K_V seen through G: it is small only when G carries the source dynamics onto the target ones. The maximum mean discrepancy (Gretton et al., 2006) is a soft penalty that encourages the aligned source features to overlap with the target features; the marginal constraint itself is left to the coupling. Appendix A.4 relates the prediction errors and this feature mismatch to the optimal value of the transport problem. Multistep prediction, conditioning terms and the variant used on Lorenz-96 are specified in Appendix B.

iEDMD cost. Once Ψ and G are learned, we use the successor cost of Proposition 1 on observed transitions $x^+ = \mathcal{F}x$, $y^+ = \mathcal{M}y$:

$$c^{\mathrm{succ}}_{\mathrm{iEDMD}}(x, y) := \|\Psi(y) - G(\Psi(x))\| + \|\Psi(y^+) - G(\Psi(x^+))\|. \tag{10}$$

This cost uses only recorded states and their successors: neither reduced predictor is applied, and the dynamics are never queried. We use this form for the ODE couplings, with each norm squared.

Remark 1 (Operator form)**.** *When successors are not available for the states being coupled, the reduced predictors can replace them, $\Psi(y^+) \approx K_V\Psi(y)$ and $\Psi(x^+) \approx K_U\Psi(x)$. Writing $u := \Psi(y) - G(\Psi(x))$, relation* (8) *and the linearity of K_V turn the second term into $\|K_V u\|$, which gives*

$$c^{\mathrm{op}}_{\mathrm{iEDMD}}(x, y) := \|u\| + \|K_V u\|. \tag{11}$$

This cost is evaluated from (x, y) alone, although learning iEDMD still requires transitions within each system. Its residual bound includes the reduced prediction errors (Appendix A.3).

We write c_{iEDMD} for either form. The ODE pairs use the successor form, and the full-state fluid pairs the operator form (Table 4). Appendix A.3 relates squared and unsquared costs.

3.4 Learning the transport with Flow Matching

The coupling is computed on minibatches of observed states. For each batch of source and target samples, we form the matrix of c_{iEDMD} costs and solve the entropically regularised transport problem with Sinkhorn iterations (Cuturi, 2013), run until the plan matches both batch marginals (Appendix E.1). We denote by π_B the resulting coupling, and draw one target partner for each source state. Flow Matching turns this coupling into a map. It learns a velocity field $v_\theta(s, z)$, where $s \in [0, 1]$ is the transport time, by regression along straight segments between paired states (Lipman et al., 2022; Tong et al., 2023):

$$\mathcal{L}_{\mathrm{FM}}(\theta) = \mathbb{E}_{(x,y)\sim\pi_B,\, s\sim\mathcal{U}[0,1]} \big\| v_\theta(s, (1-s)x + sy) - (y - x) \big\|^2. \tag{12}$$

Both the input $(1 - s)x + sy$ and the target $y - x$ are computed from the endpoints, so training requires no integration. The transport T_θ is obtained by integrating $\dot{z}_s = v_\theta(s, z_s)$ from $s = 0$ to $s = 1$, and its inverse by integrating backward from $s = 1$ to $s = 0$. Any coupling of μ_X and μ_Y yields a transport between them: at the population level, the marginal velocity field learned by conditional Flow Matching carries μ_X onto μ_Y, under the usual regularity and well-posedness assumptions (Lipman et al., 2022; Tong et al., 2023). What the coupling determines is which states are sent where. With the iEDMD cost, this correspondence is guided by the dynamics, within the limits of the marginal constraint: to respect both distributions, the coupling may have to depart from the alignment suggested by G. As noted in Section 3.2, the learned map then reproduces the coupling only approximately. Coupling and training settings are given in Appendices B and F.

3.5 Evaluation and forecasting

We evaluate a learned transport on the two requirements of (1), then use it for forecasting. Throughout, $\mathcal{F}_t$ and $\mathcal{M}_t$ denote evolution over a physical horizon t, which may exceed the lag τ used for training.

Conjugation error. The commutation residual is measured in feature space, at the training lag. To validate the learned maps, we measure commutation directly in state space, at several horizons:

$$E_{\mathrm{conj}}(t) := \mathbb{E}_{X\sim\mu_X} \left[d^{-1/2} \big\| T(\mathcal{F}_t X) - \mathcal{M}_t(TX) \big\|_2 \right]. \tag{13}$$

This is the root-mean-square gap per coordinate between the two routes of the diagram. Evaluating $\mathcal{M}_t(TX)$ requires a solver for the target system, which is available in our controlled experiments; it is used for validation only, and never during training.

Marginal fidelity. A map could lower the conjugation error by distorting the transported distribution. We therefore also report $\mathrm{MMD}^2(T_\#\mu_X, \mu_Y)$, divided by the same statistic computed between two independent target samples, with kernels distinct from those used in training. A ratio near 1 means that the error is comparable to this finite-sample reference; it does not prove that the measures coincide.

Forecasting. In practice, forecasts start from observations of the target system. Given an observation y_0, we transport it into the model's state space with T^{-1}, evolve it with the model, and transport the result back:

$$\widehat{y}_t = T\big(\mathcal{F}_t(T^{-1}y_0)\big). \tag{14}$$

If T is an exact conjugacy at horizon t and its inverse is computed exactly, then $\hat{y}_t = \mathcal{M}_t y_0$. Otherwise, both the alignment error and the numerical inversion error affect the forecast. We compare $\hat{y}_t$ with the direct forecast $\mathcal{F}_t y_0$, which corresponds to $T = \mathrm{Id}$. All methods start from the same observed states and are scored by their RMSE against the same true future $\mathcal{M}_t y_0$, which only requires recorded observations of the target system. The conjugation error measures alignment from source states; the forecast error measures what this alignment brings from target observations.

4 Experiments

Experiments proceed in three stages. Controlled coordinate changes first show that the discrepancy is insensitive to admissible diffeomorphic reparameterizations, so a poor direct model may remain good under a suitable alignment. We then benchmark conjugation and marginal fidelity on chaotic ODEs and fluids, and finally test conjugated forecasts with unchanged imperfect solvers.[1]

4.1 A discrepancy up to changes of representation

The discrepancy $\mathcal{D}_\Omega$ compares dynamics modulo the maps in Ω, much as shape analysis compares shapes modulo rigid motions. We observe Lorenz-63 through a diffeomorphism h_s of growing nonlinearity, so that source and target stay exactly conjugate (Figure 1). The direct forecast error and the Wasserstein distance grow with s, and so does $\mathcal{D}_\Omega$ over affine maps, which undo only the rotation. Over residual MLPs, it stays at the fitting floor: the nonlinear change of coordinates is absorbed. A change of dynamics is not: moving ρ from 28 to 40 raises it from $8 \cdot 10^{-4}$ to $2.6 \cdot 10^{-2}$. A model with poor direct forecasts can thus have low dynamical discrepancy from the target; what matters is whether it can be aligned.

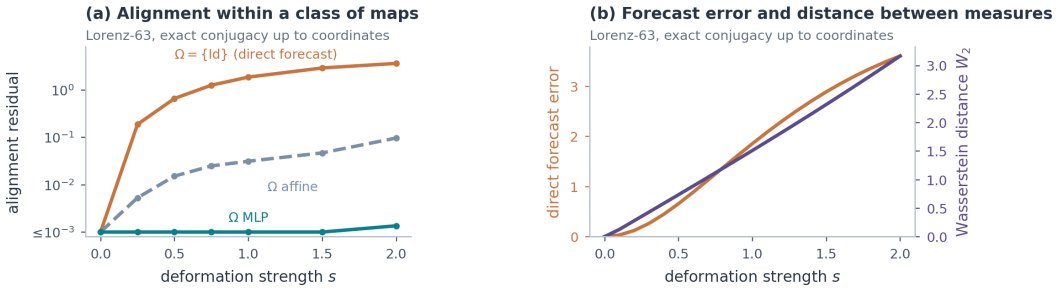


Figure 1: Lorenz-63 observed through a change of coordinates h_s of increasing strength s, from the identity ($s = 0$) to a nonlinear diffeomorphism ($s = 2$); source and target stay exactly conjugate. **(a)** Commutation residual minimised over a class Ω of maps with a soft moment constraint; $\Omega = \{\mathrm{Id}\}$ is the direct forecast error; values below 10^{-3}, the fitting floor on 1,500 samples, are shown at the bottom of the axis. **(b)** Direct forecast error (left axis) and Wasserstein distance between the invariant measures, corrected for its finite-sample floor (right axis).

4.2 Conjugation performance

For benchmarking purposes, we evaluate the three couplings by their conjugation error (13) at a short horizon, under a fixed Flow Matching architecture (Table 1).

The dynamical cost pays off when geometry and dynamics disagree. When the two systems are observed in unregistered coordinates, geometric proximity is a poor guide to correspondence: iEDMD coupling then lowers the conjugation error by 43–62% relative to Euclidean coupling on the four three-dimensional coordinate-deformed ODE pairs, and by 32–74% on the fluids, where Euclidean coupling

[1] Code and reproduction instructions: anonymous repository.

can even fall behind random pairing. The advantage fades when geometry becomes informative again: on pairs that share their coordinates and differ only by a parameter, the couplings perform alike, and a training-free affine map can be best (0.559 against 0.614 for iEDMD on Chen b). It also shrinks with dimension, to 6% on Lorenz-96 ($d = 40$), where the learned conjugator fits the dynamics less accurately. On the pure Chen rotation, a GW plan followed by a rigid Procrustes map performs better still, as expected when geometry alone identifies the correspondence.

Table 1: Conjugation error (13) (↓) on representative pairs under the same Flow Matching architecture; mean ± std over three seeds, bold: lowest per row. ODEs: RMSE at $t = 0.24$ (Rössler: $t = 1.20$), aggregated over 200 test starts by the mean, except Rössler by the segment median; conjugator and transport are retrained per seed.

Pair	random	Euclidean	iEDMD
h(L63), $28 \to 34$	13.607±0.408	7.862±0.016	**2.995±0.120**
Chen a, $h_{\mathrm{rot+nl}}$	9.192±0.494	5.653±0.058	**2.496±0.181**
Rössler a, $h_{\mathrm{rot+nl}}$	3.885±0.124	0.669±0.048	**0.284±0.079**
Chen a, h_{rot}	7.046±0.323	2.386±0.001	**1.351±0.097**
Cylinder wake (parametric)	0.238±0.055	0.280±0.032	**0.083±0.0001**
Cylinder wake (structural)	0.301±0.020	0.441±0.038	**0.114±0.004**
Fluidic pinball (structural)	0.386±0.032	0.375±0.003	**0.254±0.002**
Lorenz-96 ($d = 40$), h_{ring}	4.364±0.024	4.115±0.001	**3.883±0.023**
Chen b	0.834±0.267	0.672±0.009	**0.614±0.022**

On the coordinate-deformed ODE pairs, the iEDMD transports remain within 2–5 times the marginal sampling floor, comparable to the Euclidean ones, whereas random coupling can depart by two orders of magnitude. The Flow Matching step is also useful: the conjugator G defines a state map of its own, but it approaches the target measure only through the soft MMD penalty of (9). The transport has better marginal fidelity than this direct readout on nine of twelve pairs.

Relaxing simulation-free training, we can add $c_{\mathrm{iEDMD}}(x, T(x))$ as a regularizer. It improves every coupling, including random pairing, confirming that the cost carries dynamical information; however, it requires integrating and differentiating through T and can degrade marginal fidelity. On deformed Lorenz-63, random pairing with this penalty even beats the unregularized iEDMD coupling (2.06 versus 2.995), illustrating the trade-off between alignment and simulation-free training.

4.3 Conjugated forecasting

Data assimilation seeks the best estimate y_0 of the current target state and uses it to initialise a model. For an imperfect model, however, even the exact state can be a poor initial condition: under the wrong dynamics it is drawn toward the model's own attractor (Figure 2, right). This is the "true state, wrong model" baseline in Table 2. We instead forecast by conjugation, transporting y_0 into a dynamically corresponding model state, advancing the unchanged solver, and transporting back, $T(\mathcal{F}_t(T^{-1}y_0))$ (14). Both forecasts use the same observation and are scored against the same true future.

The conjugated Lorenz-63 forecast remains substantially closer to the target attractor: at $t = 0.5$, its normalized distance is 0.0066, against 0.0255 for the wrong model and 0.0038 for the sampling floor. The structural wake shows the same mechanism when the model misplaces the cylinder: iEDMD aligns the wake where Euclidean coupling fails (Figure 2, left). Relative to the direct wrong-model forecast, conjugation reduces the error by 77% at a quarter shedding period and by 19% at three periods.

Table 2 separates the benefit of conjugation from that of the coupling. On native Lorenz-63, conjugation lowers error by 42–68% with either OT coupling; under a nonlinear change of coordinates, iEDMD improves by 62% over geometric coupling. The same framework transports a full pinball state to a delay embedding of partial observations, lowering seven-snapshot error by 19% over geometric and 50% over random coupling. Thus conjugation can make an unchanged imperfect model more useful for forecasting, although the benefit remains problem- and horizon-dependent.

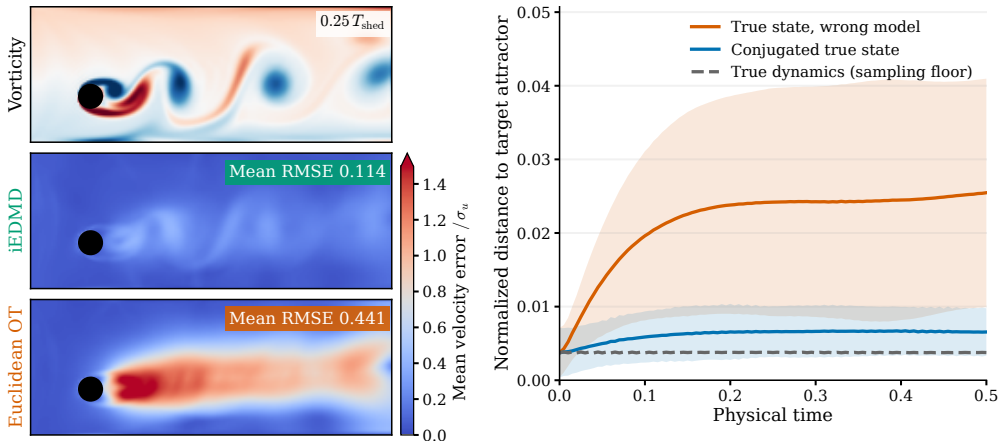


Figure 2: Dynamical alignment and initialisation. **Left:** structural wake at $0.25T_{\text{shed}}$: transported vorticity and mean velocity-conjugation error over 200 starts and three seeds. **Right:** distance to the target Lorenz-63 attractor ($\rho : 28 \to 30$); mean±std over 2,000 exact target starts; dashed: sampling floor.

Table 2: Forecast error from common observations (↓), 200 starts and three seeds. ODEs: RMSE at $t = 0.24$; wakes: normalized field RMSE at $0.25T_{\text{shed}}$; delay embeddings: observed-coordinate NRMSE after seven snapshots. Geometric denotes Euclidean coupling; bold: lowest per row.

	True state, wrong model	Conjugated true state		
ODE system		random	geometric	iEDMD
Lorenz-63, h, $\rho : 28 \to 34$	10.971	11.200±0.266	7.828±0.062	**2.993±0.143**
Lorenz-63, $\rho : 28 \to 30$	1.604	1.307±0.671	0.515±0.001	**0.507±0.005**
Lorenz-63, $\rho : 28 \to 34$	4.731	4.385±0.133	**2.713±0.013**	2.753±0.315
Lorenz-96, $F : 8 \to 10$	**0.488**	1.472±0.001	1.220±0.001	1.230±0.001

Cylinder wakes		
Pair	wrong model	iEDMD
Parametric	0.064	**0.062±0.001**
Structural	0.325	**0.074±0.002**

Full state → delay embedding			
Pair	random	geometric	iEDMD
Pinball	1.019±0.024	0.633±0.010	**0.511±0.036**
Lorenz-63	0.992±0.033	0.598±0.007	**0.343±0.008**

4.4 Complementary experiments

The appendices report complete horizons and marginal diagnostics (Appendix G), together with ablations of the cost, regularisation, geometric baselines, readouts, fluids and delay transport (Appendix H). Data and protocols are in Appendices F and C.

5 Limitations

Conjugation can degrade already accurate forecasts, as on the parametric wake and Lorenz-96. Different Koopman spectra also preclude exact measure-preserving conjugacy (Appendix A.6), so we target finite horizons; the wakes eventually drift as their shedding frequencies diverge. We assume noise-free, long records and known POD observation operators; noise, unknown operators and short records remain untested.

6 Conclusion

We learn a dynamics-aware ground cost for simulation-free transport of invariant measures. It outperforms geometric costs when geometry and dynamics disagree; otherwise, simpler maps remain preferable. A low discrepancy identifies alignable models despite poor direct forecasts, allowing an unchanged solver to start from a dynamically corresponding state.

AI USE STATEMENT

We used generative AI tools as assistants for parts of the code development, proof checking, experimental design, and manuscript editing. This interaction also led us to systematize the protocol and explore additional controls beyond our initial plan, contributing to the rigor and reproducibility of the experimental study.

The authors designed the method, selected and validated the experiments, checked the mathematical arguments, and interpreted the results. We take full responsibility for the scientific claims and final content of the paper.

Reproducibility. Systems, data, settings, protocols, and run manifests are documented in the appendices; code and instructions are in the anonymous repository linked in Section 4.

REFERENCES

Michael S Albergo and Eric Vanden-Eijnden. Building normalizing flows with stochastic interpolants. *arXiv preprint*, 2022. doi: 10.48550/arXiv.2209.15571. URL https://arxiv.org/abs/2209.15571.

Omri Azencot, N Benjamin Erichson, Vanessa Lin, and Michael Mahoney. Forecasting sequential data using consistent Koopman autoencoders. In *International Conference on Machine Learning*, 475–485, 2020. doi: 10.48550/arXiv.2003.02236. URL https://proceedings.mlr.press/v119/azencot20a.html.

Aayush Bansal, Shugao Ma, Deva Ramanan, and Yaser Sheikh. Recycle-GAN: Unsupervised video retargeting. In *European Conference on Computer Vision*, 2018. doi: 10.1007/978-3-030-01228-1_8. URL https://doi.org/10.1007/978-3-030-01228-1_8.

Peter Bauer, Alan Thorpe, and Gilbert Brunet. The quiet revolution of numerical weather prediction. In *Nature*, 525(7567):47–55, 2015. doi: 10.1038/nature14956. URL https://doi.org/10.1038/nature14956.

Petar Bevanda, Johannes Kirmayr, Stefan Sosnowski, and Sandra Hirche. Learning the Koopman eigendecomposition: A diffeomorphic approach. In *2022 American Control Conference (ACC)*, 2736–2741, 2022. doi: 10.23919/ACC53348.2022.9867829. URL https://ieeexplore.ieee.org/document/9867829.

Erik M Bollt, Qianxiao Li, Felix Dietrich, and Ioannis Kevrekidis. On matching, and even rectifying, dynamical systems through Koopman operator eigenfunctions. In *SIAM Journal on Applied Dynamical Systems*, 17(2):1925–1960, 2018. doi: 10.1137/17M116207X. URL https://doi.org/10.1137/17M116207X.

Jason J Bramburger, Steven L Brunton, and J Nathan Kutz. Deep learning of conjugate mappings. In *Physica D: Nonlinear Phenomena*, 427:133008, 2021. doi: 10.1016/j.physd.2021.133008. URL https://www.sciencedirect.com/science/article/pii/S0167278921001652.

Steven L. Brunton, Marko Budišić, Eurika Kaiser, and J. Nathan Kutz. Modern koopman theory for dynamical systems. *arXiv preprint*, 2021. doi: 10.48550/arXiv.2102.12086. URL https://arxiv.org/abs/2102.12086.

Marko Budišić, Ryan Mohr, and Igor Mezić. Applied koopmanism. In *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 22(4), December 2012. ISSN 1054-1500. doi: 10.1063/1.4772195. URL http://dx.doi.org/10.1063/1.4772195.

Eric P. Chassignet and David P.Marshall. Gulf stream separation in numerical ocean models. In *Ocean Modeling in an Eddying Regime*, 17739–61, 2008. doi: 10.1029/177GM05. URL https://doi.org/10.1029/177GM05.

Eric P. Chassignet and Xiaobiao Xu. Impact of horizontal resolution (1/12° to 1/50°) on Gulf Stream separation, penetration, and variability. In *Journal of Physical Oceanography*, 47(8):1999–2021, 2017. doi: 10.1175/JPO-D-17-0031.1. URL https://doi.org/10.1175/JPO-D-17-0031.1.

Ricky T.Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David K. Duvenaud. Neural ordinary differential equations. In *Advances in neural information processing systems*, 31, 2018. doi: 10.48550/arXiv.1806.07366. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/69386f6bb1dfed68692a24c8686939b9-Paper.pdf.

Matthew J. Colbrook. The multiverse of dynamic mode decomposition algorithms. In *Numerical Analysis Meets Machine Learning*, 25127–230, 2024. doi: 10.1016/bs.hna.2024.05.004. URL https://www.sciencedirect.com/science/article/pii/S1570865924000048.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in neural information processing systems*, 26, 2013. doi: 10.5555/2999792.2999868. URL https://proceedings.neurips.cc/paper_files/paper/2013/file/af21d0c97db2e27e13572cbf59eb343d-Paper.pdf.

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*, 34:17695–17709, 2021. doi: 10.48550/arXiv.2106.01357. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/940392f5f32a7ade1cc201767cf83e31-Paper.pdf.

Persi Diaconis and James A. Fill. Strong stationary times via a new form of duality. In *The Annals of Probability*, 18(4):1483–1522, 1990. doi: 10.1214/aop/1176990628. URL https://www.jstor.org/stable/2244330.

Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using Real NVP. *arXiv preprint*, 2016. doi: 10.48550/arXiv.1605.08803. URL https://arxiv.org/abs/1605.08803.

Alban Farchi, Patrick Laloyaux, Massimo Bonavita, and Marc Bocquet. Using machine learning to correct model error in data assimilation and forecast applications. In *Quarterly Journal of the Royal Meteorological Society*, 147(739):3067–3084, 2021. doi: 10.1002/qj.4116. URL https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/qj.4116.

Thibaut Germain, Rémi Flamary, Vladimir R. Kostic, and Karim Lounici. A Spectral-Grassmann Wasserstein metric for operator representations of dynamical systems. In *International Conference on Learning Representations*, 2026: 119652–119688, 2026. doi: 10.48550/arXiv.2509.24920. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/c256585cb41de955198b9268cdc480d4-Paper-Conference.pdf.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative Adversarial Networks. In *Commun. ACM*, 2014. doi: 10.1145/3422622. URL http://arxiv.org/abs/1406.2661.

Arthur Gretton, Karsten Borgwardt, Malte Rasch, Bernhard Scholkopf, and Alexander J. Smola. A kernel method for the two-sample problem. In *Advances in Neural Information Processing Systems*, 19: 513–520, 2006. doi: 10.5555/2976456.2976521. URL https://proceedings.neurips.cc/paper_files/paper/2006/file/e9fb2eda3d9c55a0d89c98d6c54b5b3e-Paper.pdf.

Jonathan Ho, Ajay N. Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in neural information processing systems*, 33:6840–6851, 2020. doi: 10.5555/3495724.3496298. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/4c5bcfec8584af0d967f1ab10179ca4b-Paper.pdf.

Samuel Howard, George Deligiannidis, Patrick Rebeschini, and James Thornton. Differentiable cost-parameterized Monge map estimators. *arXiv preprint*, 2024. doi: 10.48550/arXiv.2406.08399. URL https://arxiv.org/abs/2406.08399.

D. P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint*, 2013. doi: 10.48550/arXiv.1312.6114. URL https://arxiv.org/abs/1312.6114.

Paul J. Knopp and Richard D. Sinkhorn. Concerning nonnegative matrices and doubly stochastic matrices. In *Pacific Journal of Mathematics*, 21(2):343 – 348, 1967. doi: 10.2140/pjm.1967.21.343. URL https://msp.org/pjm/1967/21-2/p14.xhtml.

Bernard O. Koopman. Hamiltonian systems and transformation in hilbert space. In *Proceedings of the National Academy of Sciences*, 17(5):315–318, 1931. doi: 10.1073/pnas.17.5.315. URL https://www.pnas.org/doi/abs/10.1073/pnas.17.5.315.

Remi Lam, Alvaro Sanchez-Gonzalez, Matthew Willson, Peter Wirnsberger, Meire Fortunato, Ferran Alet, Suman Ravuri, Timo Ewalds, Zach Eaton-Rosen, Weihua Hu, et al. Learning skillful medium-range global weather forecasting. In *Science*, 382(6677):1416–1421, 2023. doi: 10.1126/science.adi2336. URL https://www.science.org/doi/abs/10.1126/science.adi2336.

Qianxiao Li, Felix Dietrich, Erik M. Bollt, and Ioannis G. Kevrekidis. Extended dynamic mode decomposition with dictionary learning: A data-driven adaptive spectral decomposition of the Koopman operator. In *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 27(10):103111, 2017. doi: 10.1063/1.4993854. URL https://doi.org/10.1063/1.4993854.

Yaron Lipman, Ricky T.Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint*, 2022. doi: 10.48550/arXiv.2210.02747. URL https://arxiv.org/abs/2210.02747.

Bethany Lusch, J. Nathan Kutz, and Steven L. Brunton. Deep learning for universal linear embeddings of nonlinear dynamics. In *Nature Communications*, 9(1):4950, 2018. doi: 10.1038/s41467-018-07210-0. URL https://doi.org/10.1038/s41467-018-07210-0.

Facundo Mémoli. Gromov–wasserstein distances and the metric approach to object matching. In *Foundations of computational mathematics*, 11(4):417–487, 2011. doi: 10.1007/s10208-011-9093-5. URL https://doi.org/10.1007/s10208-011-9093-5.

David P. Mulholland, Patrick Laloyaux, Keith Haines, and Magdalena Alonso Balmaseda. Origin and impact of initialization shocks in coupled atmosphere–ocean forecasts. In *Monthly Weather Review*, 143(11):4631–4644, 2015. doi: 10.1175/MWR-D-15-0076.1. URL https://journals.ametsoc.org/view/journals/mwre/143/11/mwr-d-15-0076.1.xml.

Samuel E. Otto and Clarence W. Rowley. Linearly recurrent autoencoder networks for learning dynamics. In *SIAM Journal on Applied Dynamical Systems*, 18(1):558–593, 2019. doi: 10.1137/18M1177846. URL https://doi.org/10.1137/18M1177846.

Jaideep Pathak, Shashank Subramanian, Peter Harrington, Sanjeev Raja, Ashesh Chattopadhyay, Morteza Mardani, Thorsten Kurth, David Hall, Zongyi Li, Kamyar Azizzadenesheli, et al. Fourcastnet: A global data-driven high-resolution weather model using adaptive fourier neural operators. *arXiv preprint*, 2022. doi: 10.48550/arXiv.2202.11214. URL https://arxiv.org/abs/2202.11214.

Gabriel Peyré and Marco Cuturi. Computational optimal transport with applications to data science. In *Foundations and Trends in Machine Learning*, 11:355–607, 2019. doi: 10.1561/2200000073. URL https://doi.org/10.1561/2200000073.

Gabriel Peyré, Marco Cuturi, and Justin Solomon. Gromov-Wasserstein averaging of kernel and distance matrices. In *Proceedings of The 33rd International Conference on Machine Learning*, 48:2664–2672, 2016. doi: 10.5555/3045390.3045671. URL https://proceedings.mlr.press/v48/peyre16.html.

Aldo Pratelli. On the equality between Monge's infimum and Kantorovich's minimum in optimal mass transportation. In *Annales de l'Institut Henri Poincaré (B) Probabilités et Statistiques*, 43 (1):1–13, 2007. doi: 10.1016/j.anihpb.2005.12.001. URL https://www.sciencedirect.com/science/article/pii/S0246020306000148.

Stephan Rasp, Stephan Hoyer, Alexander Merose, Ian Langmore, Peter Battaglia, Tyler Russell, Alvaro Sanchez-Gonzalez, Vivian Yang, Rob Carver, Shreya Agrawal, et al. Weatherbench 2: A benchmark for the next generation of data-driven global weather models. In *Journal of Advances in Modeling Earth Systems*, 16(6):e2023MS004019, 2024. doi: 10.1029/2023MS004019. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2023MS004019.

Danilo J. Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *Proceedings of the 32nd International Conference on Machine Learning*, 37:1530–1538, 2015. doi: 10.5555/3045118.3045281. URL https://proceedings.mlr.press/v37/rezende15.html.

Clarence W. Rowley, Igor Mezić, Shervin Bagheri, Philipp Schlatter, and Dan S. Henningson. Spectral analysis of nonlinear flows. In *Journal of Fluid Mechanics*, 641:115–127, 2009. doi: 10.1017/S0022112009992059. URL https://www.cambridge.org/core/journals/journal-of-fluid-mechanics/article/abs/spectral-analysis-of-nonlinear-flows/311041E1027AE7FEE7DDA36AC9AD4270.

Meyer Scetbon, Gabriel Peyré, and Marco Cuturi. Linear-time Gromov Wasserstein distances using low rank couplings and costs. In *Proceedings of the 39th International Conference on Machine Learning*, 162:19347–19365, 2022. doi: 10.48550/arXiv.2106.01128. URL https://proceedings.mlr.press/v162/scetbon22b.html.

Joseph D. Skufca and Erik M. Bollt. A concept of homeomorphic defect for defining mostly conjugate dynamical systems. In *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 18(1):013118, 2008. doi: 10.1063/1.2837397. URL https://doi.org/10.1063/1.2837397.

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. doi: 10.48550/arXiv.2011.13456. URL https://openreview.net/forum?id=PxTIG12RRHS.

Floris Takens. Detecting strange attractors in turbulence. In *Dynamical Systems and Turbulence, Warwick 1980*, volume 898 of *Lecture Notes in Mathematics*, pages 366–381. Springer, Berlin, Heidelberg, 1981. doi: 10.1007/BFb0091924. URL https://link.springer.com/chapter/10.1007/BFb0091924.

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*, 2023. doi: 10.48550/arXiv.2302.00482. URL https://arxiv.org/abs/2302.00482.

Erkan Turan, Aristotelis Siozopoulos, Louis Martinez, Julien Gaubil, Emery Pierson, and Maks Ovsjanikov. Unfolding generative flows with Koopman operators: trajectory-preserving linearization. *arXiv preprint*, 2025. doi: 10.48550/arXiv.2506.22304. URL https://arxiv.org/abs/2506.22304.

Titouan Vayer, Romain Tavenard, Nicolas Courty, Laetitia Chapel, and Rémi Flamary . Optimal transport for structured data with application on graphs. In *Proceedings of the 36th International Conference on Machine Learning*, 97:6275–6284, 2019. doi: 10.48550/arXiv.1805.09114. URL https://proceedings.mlr.press/v97/titouan19a.html.

Cédric Villani. Optimal transport: old and new *Grundlehren der mathematischen Wissenschaften*, 2009. doi: 10.1007/978-3-540-71050-9. URL https://link.springer.com/book/10.1007/978-3-540-71050-9.

Danila Volpi, Virna L. Meccia, Virginie Guemas, Pablo Ortega, Roberto Bilbao, Francisco J. Doblas-Reyes, Arthur Amaral, Pablo Echevarria, Rashed Mahmood, and Susanna Corti. A novel initialization technique for decadal climate predictions. In *Frontiers in Climate*, 3, 2021. doi: 10.3389/fclim.2021.681127. URL https://www.frontiersin.org/journals/climate/articles/10.3389/fclim.2021.681127.

Matthew O. Williams, Ioannis G. Kevrekidis, and Clarence W. Rowley. A data–driven approximation of the koopman operator: Extending dynamic mode decomposition. In *Journal of Nonlinear Science*, 25(6):1307–1346, June 2015. ISSN 1432-1467. doi: 10.1007/s00332-015-9258-5. URL http://dx.doi.org/10.1007/s00332-015-9258-5.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *2017 IEEE International Conference on Computer Vision (ICCV)*, 2223–2232, 2017. doi: 10.1109/ICCV.2017.244. URL https://ieeexplore.ieee.org/document/8237506.

A Proofs and properties of the alignment criterion

A.1 Exact conjugacy and zero discrepancy

Proposition 2 (Conjugacy as an equivalence relation)**.** *Define $\mathcal{F} \sim \mathcal{M}$ when there exists a diffeomorphism T with $T_{\#}\mu_X = \mu_Y$ such that, almost everywhere,*

$$T_{\#}(P^X \delta_x) = P^Y \delta_{T(x)}. \tag{15}$$

Then $\sim$ is an equivalence relation.

Proof. The identity establishes reflexivity. Applying $(T^{-1})_{\#}$ to (15) establishes symmetry, with $(T^{-1})_{\#}\mu_Y = \mu_X$. If T_1 intertwines P^X, P^Y and T_2 intertwines P^Y, P^Z, then $(T_2 \circ T_1)_{\#}P^X = (T_2)_{\#}P^Y(T_1)_{\#} = P^Z(T_2 \circ T_1)_{\#}$, and $(T_2 \circ T_1)_{\#}\mu_X = \mu_Z$. This proves transitivity. □

An exact conjugacy in Ω gives $\mathcal{D}_\Omega = 0$. Conversely, zero discrepancy implies an exact conjugacy only if the infimum is attained by an invertible admissible map and the discrepancy separates the transition measures. For deterministic dynamics, the anchored dictionary separates the relevant point masses. A finite feature-mean discrepancy need not separate general stochastic transition measures.

A.2 Proof of Proposition 1 and cross-moment certificate

For $\pi \in \Pi(\mu_X, \mu_Y)$, set $m_\pi(x) = \mathbb{E}[\Psi(Y) \mid X = x]$ and $q_\pi(x) = \mathbb{E}[\Psi(\mathcal{M}Y) \mid X = x]$. The coupling residual

$$\mathcal{L}_\Psi(\pi) := \mathbb{E}_{\mu_X} \|m_\pi(\mathcal{F}X) - q_\pi(X)\| \tag{16}$$

agrees with (3) on a Monge coupling. Define

$$\Sigma_\pi = \int \Psi(y)\Psi(x)^\top \, d\pi(x,y), \qquad \Sigma_\pi^+ = \int \Psi(\mathcal{M}y)\Psi(\mathcal{F}x)^\top \, d\pi(x,y),$$
$$\mathcal{S}_\Psi(\pi) = \|\Sigma_\pi - \Sigma_\pi^+\|_F. \tag{17}$$

Proposition 3 (Intertwining bounds)**.** *Assume $\mathcal{F}_{\#}\mu_X = \mu_X$ and $0 < R_X := \operatorname{ess\,sup}_{\mu_X} \|\Psi(X)\| < \infty$. For every coupling π and measurable g for which the terms are integrable,*

$$\mathcal{S}_\Psi(\pi)/R_X \le \mathcal{L}_\Psi(\pi) \le \int c^g_{\mathrm{succ}} \, d\pi.$$

Proof. Write $D_\pi(x) = m_\pi(\mathcal{F}x) - q_\pi(x)$. Disintegration and invariance give

$$\Sigma_\pi - \Sigma_\pi^+ = \mathbb{E}_{\mu_X}[D_\pi(X)\Psi(\mathcal{F}X)^\top].$$

Since $\|ab^\top\|_F = \|a\|\|b\|$, its norm is bounded by $R_X \mathbb{E}\|D_\pi(X)\|$. For the upper bound, insert $g\Psi(\mathcal{F}X)$ and use conditional Jensen and invariance:

$$\begin{aligned}\mathcal{L}_\Psi(\pi) &\le \mathbb{E}\|m_\pi(\mathcal{F}X) - g\Psi(\mathcal{F}X)\| + \mathbb{E}\|g\Psi(\mathcal{F}X) - q_\pi(X)\| \\ &\le \int \big(\|\Psi(y) - g\Psi(x)\| + \|\Psi(\mathcal{M}y) - g\Psi(\mathcal{F}x)\|\big) \, d\pi.\end{aligned}$$

□

For the Monge distribution $\pi_T = (\mathrm{Id}, T)_{\#}\mu_X$, $m_{\pi_T}(x) = \Psi(Tx)$ and $q_{\pi_T}(x) = \Psi(\mathcal{M}(Tx))$. Hence $\mathcal{L}_\Psi(\pi_T) = \mathcal{L}_\Psi(T)$, and the upper bound in Proposition 3 is exactly Proposition 1.

Role of the lower bound. The cross-moment defect is a certificate, not a pairing objective. For the product coupling $\mu_X \otimes \mu_Y$, stationarity gives $\Sigma_\pi = \Sigma_\pi^+$ and $m_\pi(\mathcal{F}x) = q_\pi(x) = \mathbb{E}\Psi(Y)$. Thus both $\mathcal{S}_\Psi$ and the extended residual vanish without a state correspondence. The successor cost generally remains positive on this coupling.

In a seed-42 diagnostic over the thirteen ODE pairs, the learned-map residual lay between the two bounds in every case; the successor-cost upper bound was within a factor 2.4–5.0 of the residual.

A.3 Operator proxy and prediction errors

Write the reduced prediction residuals as

$$r_U(x) = \Psi(\mathcal{F}x) - K_U\Psi(x), \qquad r_V(y) = \Psi(\mathcal{M}y) - K_V\Psi(y),$$

with $\epsilon_U = \mathbb{E}_{\mu_X}\|r_U\|$ and $\epsilon_V = \mathbb{E}_{\mu_Y}\|r_V\|$. For an L_g-Lipschitz alignment,

$$\|\Psi(\mathcal{M}y) - g\Psi(\mathcal{F}x)\| \le \|K_V\Psi(y) - g(K_U\Psi(x))\| + \|r_V(y)\| + L_g\|r_U(x)\|. \tag{18}$$

Consequently the reduced-predictor cost

$$c^g_{\mathrm{op}}(x,y) = \|\Psi(y) - g\Psi(x)\| + \|K_V\Psi(y) - g(K_U\Psi(x))\|$$

satisfies

$$\mathcal{L}_\Psi(\pi) \le \int c^g_{\mathrm{op}}\, d\pi + L_g\epsilon_U + \epsilon_V. \tag{19}$$

The prediction-error remainder depends only on the marginals.

For the simplified iEDMD proxy, insert $K_V g\Psi(x)$ in the second norm and define the reduced intertwining defect

$$\delta_g = \mathbb{E}_{\mu_X}\|g(K_U\Psi(X)) - K_V g\Psi(X)\|.$$

This gives, for $\pi_T = (\mathrm{Id}, T)_{\#}\mu_X$,

$$\mathcal{L}_\Psi(T) \le \int c^{\mathrm{op}}_{\mathrm{iEDMD}}\, d\pi_T + L_g\epsilon_U + \epsilon_V + \delta_g. \tag{20}$$

The same inequality holds for the coupling residual. iEDMD has $\delta_g = 0$ by construction. For a general linear $g(z) = Gz$, $L_g = \|G\|$ and $\delta_g \le R_X\|GK_U - K_VG\|_{\mathrm{op}}$.

Squared-norm implementations. For either realization, write $c_{\mathrm{iEDMD}} = a + b$ for its two norm terms. ODEs and full-state pinball use $a^2 + b^2$; the source-to-delay pinball uses $a + b$. Since π is a probability measure,

$$\int (a+b)\, d\pi \le \sqrt{2\int (a^2+b^2)\, d\pi}.$$

Thus the squared successor cost controls the residual through Proposition 1, and the squared operator cost through (20), with the same prediction-error remainder.

A.4 Exact and learned feature-space alignments

Remark 2 (Exactness in aligned feature coordinates)**.** *Since Ψ contains the state, every admissible T has a feature-space version G_T satisfying $G_T \circ \Psi = \Psi \circ T$ on the source support. For the discrepancy induced by d_Ψ,*

$$\begin{aligned} \inf_G \int c^G_{\mathrm{succ}}\, d\pi_T &= \int c^{G_T}_{\mathrm{succ}}\, d\pi_T = \mathcal{L}_\Psi(T), \\ \mathcal{D}_\Omega(\mathcal{F} \to \mathcal{M}) &= \inf_{\substack{T\in\Omega \\ T_{\#}\mu_X = \mu_Y}} \inf_G \int c^G_{\mathrm{succ}}\, d\pi_T, \end{aligned} \tag{21}$$

where the inner infimum is over measurable alignments.

Proof. Injectivity of Ψ defines G_T on its source image. On the graph of T, the endpoint term vanishes and the successor term equals $\|\Psi(\mathcal{M}TX) - \Psi(T\mathcal{F}X)\|$. The successor bound shows that no other G gives a smaller integral. Taking the infimum over T proves the second equality. □

The learned iEDMD family need not contain this ideal G_T. Its prediction accuracy can instead be related to the achievable OT cost. Let $\nu_G = G_{\#}\Psi_{\#}\mu_X$ and $\nu_Y = \Psi_{\#}\mu_Y$ have finite first moments. Under exact reduced intertwining, linear K_V, L_G-Lipschitz G and finite prediction errors,

$$\inf_{\pi\in\Pi(\mu_X,\mu_Y)} \int c^G_{\mathrm{succ}}\, d\pi \le (1 + \|K_V\|)W_1(\nu_G, \nu_Y) + L_G\epsilon_U + \epsilon_V. \tag{22}$$

Indeed, (18) bounds the integrand by $(1 + \|K_V\|)\|\Psi(y) - G\Psi(x)\| + L_G\|r_U(x)\| + \|r_V(y)\|$. Minimizing the first term over state couplings gives the feature-space W_1 distance: any feature coupling lifts to a state coupling by disintegration. This proves (22) without assuming that G exactly transports the feature marginals. The population iEDMD loss controls the prediction terms and uses MMD as a soft distributional penalty. For a characteristic kernel, zero population MMD means $\nu_G = \nu_Y$ (Gretton et al., 2006); a quantitative bound in terms of a nonzero MMD would require additional assumptions. Equation (22) concerns the population transport cost, not convergence of the fitted networks or recovery of a unique correspondence.

Proposition 4 (Alignment of the induced transport)**.** *Assume* $G(\Psi(\operatorname{supp}\mu_X)) \subseteq \Psi(Y)$ *and* $G_\#\Psi_\#\mu_X = \Psi_\#\mu_Y$. *Define* $T_G = \Psi^{-1} \circ G \circ \Psi$. *If* G *is* L_G*-Lipschitz, then* $(T_G)_\#\mu_X = \mu_Y$ *and*

$$\left|\mathcal{L}_\Psi(T_G) - \delta_G\right| \le L_G\epsilon_U + \epsilon_V. \tag{23}$$

Proof. Injectivity and the feature-marginal equality give $(T_G)_\#\mu_X = \Psi_\#^{-1}G_\#\Psi_\#\mu_X = \mu_Y$. The vectors defining the two residuals in (23) differ by

$$\big[G\Psi(\mathcal{F}X) - G(K_U\Psi(X))\big] - \big[\Psi(\mathcal{M}T_GX) - K_V\Psi(T_GX)\big].$$

The reverse triangle inequality and the marginal identity bound the difference of their expected norms by $L_G\epsilon_U + \epsilon_V$. □

For an exact conjugacy $T^\star$, writing $e_G(x) = \|\Psi(T^\star x) - G\Psi(x)\|$ gives $c^G_{\mathrm{succ}}(x, T^\star x) = e_G(x) + e_G(\mathcal{F}x)$. Invariance therefore yields $\int c^G_{\mathrm{succ}}\, d\pi_{T^\star} = 2\mathbb{E}_{\mu_X} e_G(X)$. This cost vanishes when $G\Psi = \Psi T^\star$.

A.5 Identity dictionary and Wasserstein transport

When $\Psi = \mathrm{Id}$ and $X, Y \subseteq \mathbb{R}^d$, the endpoint term gives

$$\inf_{\pi\in\Pi(\mu_X,\mu_Y)} \int \|y - g(x)\|_2\, d\pi = W_1(g_\#\mu_X, \mu_Y). \tag{24}$$

Squaring this endpoint norm gives W_2^2. The operator cost is

$$c^{\mathrm{op}}_{\mathrm{iEDMD}}(x, y) = \|y - g(x)\|_2 + \|K_V(y - g(x))\|_2 = \|y - g(x)\|_{K_V,+}, \tag{25}$$

where $\|u\|_{K_V,+} = \|u\|_2 + \|K_Vu\|_2$ is a norm, even if K_V is singular. Its OT objective is therefore $W_1(g_\#\mu_X, \mu_Y)$ in this anisotropic norm. The squared operator cost similarly uses the positive definite quadratic form $I + K_V^\top K_V$. The successor cost instead compares one-step trajectory segments:

$$c^g_{\mathrm{succ}}(x, y) = \|y - g(x)\|_2 + \|\mathcal{M}y - g(\mathcal{F}x)\|_2. \tag{26}$$

A.6 Spectral obstruction to exact conjugacy

An invertible measure-preserving conjugacy T induces the unitary composition operator $C_T : L^2(\mu_Y) \to L^2(\mu_X)$, $C_Th = h \circ T$. The relation $T\mathcal{F} = \mathcal{M}T$ gives $UC_T = C_TV$, hence $V = C_T^{-1}UC_T$. Thus the Koopman operators are unitarily equivalent and have the same spectrum (Budišić et al., 2012). Different spectral invariants rule out such a conjugacy. This statement concerns an invertible intertwiner on the corresponding observable spaces; noninvertible links may intertwine systems with different spectra. Differences between finite EDMD estimates can also reflect approximation error.

For linear invertible G, the analogous reduced statement is

$$K_V = GK_UG^{-1}, \qquad \sigma(K_U) = \sigma(K_V), \tag{27}$$

with equal Jordan structure. Right eigenvectors transform as $v_V = Gv_U$ and left eigenvectors as $w_V^\top = w_U^\top G^{-1}$. Consequently the associated eigenfunction coordinates agree when $\Psi(y) = G\Psi(x)$. For a nonzero Sylvester residual $R_{\mathrm{Syl}} = GK_U - K_VG$,

$$\|K_V - GK_UG^{-1}\| \le \|R_{\mathrm{Syl}}\|\,\|G^{-1}\|. \tag{28}$$

Interpretation therefore depends on conditioning. iEDMD enforces reduced intertwining exactly; incompatibility with the observed dynamics appears through prediction errors. For nonlinear G, K_U is a nonlinear reduced map and the matrix-similarity statement does not apply.

The wake pairs have rotation-number differences 0.014 and 0.024 (Appendix F.1), producing phase drift despite useful short-horizon alignment. Nonexistence of an exact conjugacy alone does not imply a strictly positive infimum of the discrepancy: that also requires attainment in an admissible class excluding singular limits.

B iEDMD AND TRAINING DETAILS

Parameterizations of the alignment stage. All core coupling comparisons follow the four-step procedure of Section 3. They differ in the parameterization fitted in the first step. In every iEDMD conjugator we set $K_U := G^{-1} \circ K_V \circ G$, so that $\delta_G = 0$ by construction. There is no separately trained source matrix or Sylvester penalty. The source prediction loss uses the explicit inverse of G, except on L96, where it is evaluated in aligned coordinates as $\|G(\Psi(\mathcal{F}X)) - K_V(G(\Psi(X)))\|^2$. For nonlinear G, this K_U is a nonlinear predictor rather than an EDMD matrix, so matrix-similarity conclusions do not apply. The successor bound remains valid in both cases. The identity dictionary, learned dictionaries, multistep losses, and conditioning terms are specified below, together with the per-system selections. Regularisation experiments are separate controls, not additional steps of the core coupling procedure.

Table 3: Hyperparameters of the primary ODE comparisons and regularised controls. Fluid settings are specified in Appendix F.1.

Flow matching	(Random, ℓ_2, iEDMD, and regularised controls)
Architecture	MLP, width 128, 4 blocks, SiLU
Normalization	One scalar mean and std over pooled source and target (preserves geometry)
Epochs / batch	600 / 2048
Optimizer	Adam, lr $= 10^{-3}$, weight decay 10^{-3} (selected on the validation split)
Seeds	42/43/44 in both primary ODE stages
Koopman operator	
Encoder ϕ	MLP (periodic convolution on L96), $C_\phi = 16$
Conjugator G	Linear or affine coupling flow; $K_U = G^{-1} K_V G$
Epochs	500 (300 on L96)
Batch size / lr	256 / 10^{-3}
Koopman regularization $\mathcal{L}_K$	
Strength	calibrated: $\lambda \cdot (\text{term}) = 3\,\mathcal{L}_{\text{FM}}$ at the first regularised step
Regularised states / frequency	256 generated states, every 4th step
Flow integration	10 RK4 steps, differentiated; gradient clipping (norm 5)
Sinkhorn OT coupling	
ε (entropic reg.)	relative: $0.02\times$ mean batch cost, every cost and pair
Iterations	log domain, run to convergence: $\max_i \lvert \sum_j P_{ij} - 1 \rvert < 10^{-3}$ (≤ 5000)
Batch	2048; one target sampled per source row

Primary ODE optimization details. The objective optimized in practice instantiates Eq. (9) with multistep prediction losses. For both linear and nonlinear g, the optimized minibatch loss is $\mathcal{L}_{\text{EDMD}} + \lambda_{\text{MMD}}\,\text{MMD}^2 + \lambda_{\text{id}}\,\|g(\Psi(x)) - \Psi(x)\|^2$ with an additional singular-value penalty on K_V once per epoch to control its conditioning. This penalty is not an intertwining loss. The identity anchor uses λ_{id}=0.5 on the base pairs and zero on the deformed pairs. With the deformed-pair setting (affine-coupling g, no anchor; three seeds), iEDMD improves on Euclidean coupling on four of the seven base pairs, remains within 0.006 on two, and is worse on widened Lorenz-63. The MMD uses an RBF kernel with the median-bandwidth heuristic recomputed per minibatch. Nonlinear g uses an affine coupling flow (Dinh et al., 2016) with hidden width 128; training uses Adam, learning rate 10^{-3}, 500 epochs. Flow-Matching transports use an MLP (width 128, 4 blocks, SiLU), batch 2048, weight decay 10^{-3} on the ODE pairs and the wakes and 0.01 on full-state pinball

(Appendix F), 600 epochs, learning rate 10^{-3}, a single scalar normalization shared by both sides; Sinkhorn couplings use the relative ε=0.02 and the convergence criterion above on the ODE pairs for both costs (Appendix E.1). The ODE weight decay is the value of $\{0, 10^{-4}, 10^{-3}, 10^{-2}\}$ with the lowest pushforward MMD^2 over its sampling floor on validation trajectories, averaged over four pairs.

C PROTOCOL

Primary ODE comparisons. The prediction lag is determined from the first $N = 30{,}000$ source training states, centred coordinatewise at their original relative scales. Define

$$r(k) = \frac{\sum_{t=0}^{N-k-1} \langle z_t, z_{t+k} \rangle / (N-k)}{\sum_{t=0}^{N-1} \|z_t\|^2 / N}, \qquad k_{\mathrm{ac}} = \min\{1 \le k \le N/10 : r(k) \le e^{-1}\}.$$

Set $\mathrm{stride} = \max(1, \lfloor 0.4 k_{\mathrm{ac}} + \frac{1}{2} \rfloor)$ and $\Delta t_A = \mathrm{stride}\, \delta t$, with the same stride for the two recorded endpoints of the successor cost. The resulting lags are 0.08 for Lorenz-63, 0.04 for Chen, 0.40 for Rössler and 0.11 for Lorenz-96. All pairs use $C_\varphi = 16$, five prediction steps and decay 0.5. The common coefficient 0.4 and the family-specific architectures reflect preliminary experiments; they are held fixed across this comparison. The dictionary network φ has width 128 and depth 4 on Chen a, Rössler and the deformed pairs, width 32 and depth 3 on the other Lorenz-63 pairs and Chen b, and is a periodic convolution of width 48 on Lorenz-96, which also uses a circulant K_V and the aligned-coordinate source loss.

Both stages read stored training pools; the conjugator uses the first 30,000 starting states and their required successors. Both stages are retrained with seeds 42, 43 and 44, retaining the final epoch without test-based selection. Test trajectories are continuations beyond the validation trajectories with a 2000-step gap. For the rotated pairs, test trajectories are inherited from the base pair and transformed by the declared target observation map. All methods use the same test states; input arrays and configurations are identified by the archived manifest.

Method specification. Table 4 collects the dictionaries, alignments and cost realizations for the full-state experiments. Pinball velocity features are regressed from finite differences (Appendix I.2); all families use $K_U = G^{-1} \circ K_V \circ G$ and a linear target predictor. The source-to-delay variant uses identity dictionaries in source POD and target delay coordinates, with the unsquared successor cost and newly trained G, K_V (Appendix I.3).

Table 4: Method specification for the full-state experimental families. Source-to-delay transport: Appendix I.3.

	chaotic ODEs	wakes	pinball
dictionary Ψ	$[x; \varphi(x)]$ (no velocity feature)	Id (POD coefs)	$[z; \hat{v}(z); \varphi(z)]$
alignment	diffeomorphic g or linear G	diffeomorphic g	diffeomorphic g
cost	squared successor	unsquared operator	squared operator
transport	FM on c-coupling	FM on c-coupling; g control	FM on c-coupling

Solver-free diagnostic and selection scope. The operator score $\mathbb{E}[\|u\| + \|K_V u\|]$, $u = \Psi(Tx) - g\Psi(x)$, can evaluate a fitted transport without querying a physical solver. It still requires applying the transport, hence integrating its ODE. Its use as a solver-free selection rule is tested prospectively in Appendix H.6 (Table 15).

Segment design and paired uncertainty. The n=200 evaluation segments start at indices evenly spaced over the test trajectory. Adjacent segments can overlap slightly and are temporally correlated. All methods are evaluated from the *same* starting states. We report per-segment iEDMD-minus-ℓ_2 differences at the first evaluation horizon, pooled over the three seeds, with 95% block-bootstrap intervals over blocks of 10 consecutive segments (escaped Rössler segments excluded). The interval excludes zero in favour of the dynamical coupling on seven of the thirteen ODE pairs — by a wide margin on the deformed and rotated pairs (-4.87 on h(L63), -3.16 and -1.03 on the rotated Chen pairs, -0.81 and -0.52 on the rotated Rössler pairs) and by 0.06–0.08 on Chen a and Chen b — includes zero on the two Lorenz-63 ρ pairs and on Rössler c, and excludes it against the dynamical coupling on Rössler a, Lorenz-63 σ and Lorenz-96 ($+0.012$ to $+0.057$).

Evaluation. The ODE conjugation error is $\mathrm{RMSE}(t) = d^{-1/2}\|T(F_s^t(x)) - F_t^t(T(x))\|_2$. The harness stores 100 snapshots and reports indices 24, 49, and 98: physical times 0.24, 0.49, 0.98 for sampling interval 0.01, and 1.20, 2.45, 4.90 for Rössler (interval 0.05). The table labels $0.25/0.50/0.99$ denote fractions of the evaluation window, rather than common physical times. Conjugation uses fixed-step RK4; the ODE forecast comparison uses the separate adaptive-solver harness. Table 5 gives the three-seed ODE comparison at the two later indices; the first index and marginal ratios are in Table 9. Table 16 reports wake seed variability.

Table 5: Later-horizon conjugation error for the three couplings under the same FM architecture and fixed family-specific protocol (Appendix C). Mean±sample standard deviation over three complete seeds; Rössler uses the segment median and other pairs the mean. Indices $49/98$ correspond to $t = 0.49/0.98$ and, on Rössler, $2.45/4.90$. Bold: best coupling. The first horizon is in Table 9.

index	Pair	iEDMD	ℓ_2	Random
49	h(L63), $28\to34$	**5.643**±.131	9.924±.075	12.503±.203
	Chen a, $h_{\mathrm{rot+nl}}$	**4.802**±.052	6.696±.083	9.323±.518
	Rössler a, $h_{\mathrm{rot+nl}}$	**0.344**±.064	1.024±.025	3.433±.220
	Chen a, h_{rot}	**3.948**±.058	4.647±.007	8.628±.175
	Rössler a, h_{rot}	**0.357**±.042	0.814±.026	3.579±.157
	L63 ρ, $28\to34$	5.866±.349	**5.559**±.020	7.244±.923
	Chen a	3.802±.019	**3.637**±.007	4.053±.238
	Rössler a	0.197±.003	**0.188**±.002	1.314±.149
	L63 ρ, $28\to30$	**1.080**±.012	1.088±.006	2.216±.725
	L63 σ	0.974±.016	**0.929**±.003	4.363±1.021
	Chen b	**1.620**±.006	1.659±.005	1.799±.298
	Rössler c	0.361±.031	**0.356**±.032	2.080±.196
	Lorenz-96 (d=40)	2.410±.002	**2.334**±.001	3.032±.002
98	h(L63), $28\to34$	**9.527**±.107	13.307±.087	13.652±.816
	Chen a, $h_{\mathrm{rot+nl}}$	**8.959**±.212	9.752±.087	10.636±.829
	Rössler a, $h_{\mathrm{rot+nl}}$	**0.476**±.054	1.031±.049	4.108±.126
	Chen a, h_{rot}	**8.166**±.179	8.290±.023	9.811±.170
	Rössler a, h_{rot}	**0.430**±.052	0.754±.101	4.575±.126
	L63 ρ, $28\to34$	8.996±.198	**8.888**±.017	10.143±.093
	Chen a	7.474±.030	7.694±.038	**7.314**±.142
	Rössler a	0.283±.003	**0.282**±.001	1.567±.186
	L63 ρ, $28\to30$	**2.610**±.029	2.653±.015	4.094±.791
	L63 σ	1.710±.025	**1.657**±.019	5.554±.818
	Chen b	**4.946**±.134	5.512±.015	5.156±.084
	Rössler c	**0.807**±.044	0.878±.020	2.366±.107
	Lorenz-96 (d=40)	4.307±.003	**4.252**±.002	4.886±.008

Compute. The primary ODE comparison comprises 39 runs on eight A100 GPUs (80 GB), completed in approximately 16 hours; each run fits one conjugator and three FM transports, then evaluates them. For n-dimensional features, constructing a dense $B \times B$ cost matrix requires $O(B^2 n)$ pairwise arithmetic and $O(B^2)$ storage; each Sinkhorn iteration costs $O(B^2)$. A single float32 matrix occupies 16 MiB at $B = 2048$ and 256 MiB at $B = 8192$, before solver workspaces. ODE plans are computed per minibatch; wake plans are precomputed on blocks. Thus plan construction scales with batch or block size rather than requiring a matrix over the full training set.

D CODE AVAILABILITY AND REPRODUCIBILITY

The anonymous companion repository contains the code and instructions for reproducing the experiments reported in this paper:

`https://anonymous.4open.science/r/dynamical-alignment-transport-33F6/README.md`

The repository is anonymized for double-blind review. Experimental systems, data splits, hyperparameters and evaluation protocols are additionally documented in Appendices F, B and C.

E COUPLING DIAGNOSTICS

E.1 SINKHORN CONVERGENCE

With 50 iterations and the absolute scale $\varepsilon = 0.005$, row-sum errors range from 0.2 to 22.3 on a batch of 2048, so the resulting matrix does not satisfy the target marginal. All ODE and full-state pinball comparisons instead use log-domain Sinkhorn with $\varepsilon = 0.02C$ until the maximum row-sum error is below 10^{-3}; convergence takes 190–930 iterations. Wake and source-to-delay settings are given in Appendices F.1 and I.3.

E.2 LORENZ-96 AT FOUR LEVELS OF CONJUGACY

Table 6 separates parameter mismatch from coordinate mismatch. For $F{=}8 \to 10$, Euclidean and iEDMD couplings differ by only 1.5%, while random pairing is 33% worse. For $F{=}8 \to 12$, random pairing is instead best, and iEDMD is 3.3% worse than Euclidean coupling: there is no consistent benefit when the coordinates are shared. Removing the identity anchor further worsens iEDMD from 2.638 to 2.743. Under the ring-local diffeomorphism (Appendix F), iEDMD becomes best by 6%. The ratio r_{Id}/r_g is reported to show how strongly the learned alignment departs from the identity; its magnitude alone does not predict whether the resulting coupling helps.

Table 6: Lorenz-96 ($d{=}40$), conjugation RMSE at the first horizon (↓, mean over three seeds) at four levels of conjugacy; weight decay 10^{-3}. $\|g - \mathrm{Id}\|/\|\psi\|$: distance of the learned conjugator from the identity in the lifted space; r_{Id}/r_g: intertwining residual with $g := \mathrm{Id}$ over the learned residual.

Lorenz-96 pair	$\|g - \mathrm{Id}\|/\|\psi\|$	r_{Id}/r_g	random	Euclidean	iEDMD	Eucl. vs iEDMD
$F{=}8 \to 10$, identity anchor	0.15	6.8	1.677	**1.258**	1.277	-1.5%
$F{=}8 \to 12$, identity anchor	0.11	3.2	**2.439**	2.551	2.638	-3.3%
$F{=}8 \to 12$, deformed-pair settings	0.86	10.9	**2.439**	2.551	2.743	-7.0%
$F{=}8 \to h(12)$, ring diffeomorphism	1.35	10.1	4.364	4.115	**3.883**	$+6.0\%$

F DATA

We consider six 3D chaotic systems with close parameter shifts, one high-dimensional system (Lorenz-96, $d = 40$), one widened Lorenz-63 pair, and pairs observed through a diffeomorphism h of the target coordinates (Table 7). All diffeomorphisms are of the same family, $h = C \circ (R \cdot + t)$ with R a rotation, t a translation of 10% of the attractor range and C the triangular coupling $(u_1, u_2, u_3) \mapsto (u_1,\ u_2 + a\tanh(u_1/s_1),\ u_3 e^{b\tanh(u_2/s_2)} + c\tanh(u_1/s_1))$, a global diffeomorphism with closed-form inverse and Jacobian, so that the target vector field $f_M(y) = Dh(x)f_F(x)$, $x = h^{-1}(y)$, is available to the solver-in-the-loop harness. The Lorenz-63 ladder uses a 20° rotation about the x-axis and $(a, b, c) \in \{(0, 0, 0), (2, 0.15, 1.5), (4, 0.3, 3), (8, 0.5, 6)\}$ (levels zero, small, medium, large; $(s_1, s_2) = (8, 9)$, one standard deviation), with target $\rho{=}34$ (misspecified) or $\rho{=}28$ (exact conjugacy h, suffix "same"). The Chen a and Rössler a pairs use a 30° rotation about the axis $(1, 1, 1)/\sqrt{3}$ — every coordinate mixes — either alone ("rot") or followed by the medium nonlinearity rescaled to the system's standard deviation ("rotnl": $(a, b, c) = (4, 0.3, 3)$, $(s_1, s_2) = (8.5, 9.1)$ on Chen; $(2.1, 0.3, 1.6)$, $(5.2, 4.9)$ on Rössler). The non-affinity δ — residual of the best affine fit of h on the attractor, relative to the attractor scale — is 0 for the pure rotations, 0.117 (L63 medium), 0.115 (Chen rotnl) and 0.085 (Rössler rotnl).

For each system, we generate long trajectories (100,000 points, $\mathrm{dt} = 0.01$ for Lorenz/Chen, $\mathrm{dt} = 0.05$ for Rössler) after a transient burn-in of 2,000 steps. The primary comparisons fit iEDMD on the first 30,000 starting states and their successors; FM samples from the full training pools. Evaluation uses the prospective test split of Appendix C: $n{=}200$ segments of 100 snapshots from temporal continuations.

Geometric and dynamical diagnostic sweeps. Figure 1 uses 1,500 sampled states per system and the normalised residual $\mathbb{E}\|T(\mathcal{F}X) - \mathcal{M}(TX)\|^2/s_Y^2$, where s_Y is the mean target coordinate standard deviation. Affine and residual-MLP maps are fitted by minimising this residual plus squared, normalised deviations of the transported coordinate means and standard deviations from the target;

Table 7: Dynamical systems and parameter perturbations.

System	Param	Source	Target	Dim
Lorenz-63 (ρ)	ρ	28.0	30.0	3
Lorenz-63 (σ)	σ	10.0	12.0	3
Chen (a)	a	35.0	36.0	3
Chen (b)	b	3.0	2.75	3
Rössler (a)	a	0.2	0.225	3
Rössler (c)	c	5.7	6.85	3
Lorenz-96	F	8.0	10.0	40
Lorenz-96	F	8.0	12.0	40
Lorenz-96, ring diffeomorphism	F, then h_{ring}	8.0	$h_{\mathrm{ring}}(\cdot)$, 12.0	40
Lorenz-63, widened	ρ	28.0	34.0	3
h(L63), ladder zero/small/medium/large	ρ, then h	28.0	$h(\cdot)$, 34.0	3
h(L63), exact conjugacy ("same")	h only	28.0	$h(\cdot)$, 28.0	3
Chen (a), rot / rotnl	a, then h	35.0	$h(\cdot)$, 36.0	3
Rössler (a), rot / rotnl	a, then h	0.2	$h(\cdot)$, 0.225	3

this soft moment constraint gives a diagnostic surrogate for the constrained discrepancy of Section 3.1. For $\Omega = \{\mathrm{Id}\}$ nothing is fitted: the residual is evaluated on target states, which makes it the direct forecast error of the source model over one lag (0.1). The target is $h_s(\mathrm{L63})$ at $\rho = 28$, where h_s interpolates the rotation ($s/2 \cdot 20^\circ$ about the x-axis) and the translation ($s/2 \cdot t$) together with the triangular nonlinearity $(a, b, c) = (4s, 0.25s, 3s)$, so that $h_0 = \mathrm{Id}$ and h_2 is the large deformation. The Wasserstein distance is the exact W_2 between source and target samples in units of the mean source standard deviation, corrected for its finite-sample floor $W_{2,0} = 0.17$ (two independent samples of the same measure) as $\sqrt{\max(W_2^2 - W_{2,0}^2, 0)}$. Changing the dynamics instead, with the target ρ swept from 28 to 40 at $s = 0$, the residual over residual MLPs, fitted at eight values of ρ, rises from $8 \cdot 10^{-4}$ to $2.6 \cdot 10^{-2}$.

Ring-local diffeomorphism for Lorenz-96. The additional d=40 deformation control observes the F=12 target through h_{ring}, an invertible coupling on the periodic ring of sites: $y_{2k} = x_{2k}$ and $y_{2k+1} = x_{2k+1} + a \tanh\big((x_{2k} + x_{2k+2})/2s\big)$ with $a = 8.0$ (1.6 target standard deviations) and $s = 5.0$. Even sites pass through, each odd site is shifted by a bounded function of its two even neighbours; the map is exactly invertible, strictly local, and preserves the translation structure used by the convolutional conjugator. The target vector field is the exact pushforward $Dh\, f \circ h^{-1}$.

F.1 Von Kármán wakes and pinball: data and protocol

Physical setup. All flows are simulated with a D2Q9–BGK lattice-Boltzmann solver (parabolic inlet with $U_{\max}$=0.1, no-slip walls, bounce-back cylinder of diameter D=16, snapshots every 20 LBM steps, 50,000 vorticity snapshots per system, unpaired). The *parametric* pair uses a 256×64 channel with the cylinder at $(W/6,\ H/2)$ for both systems and Re=100 (source) versus Re=200 (target). The *structural* pair uses a taller 256×96 channel; the source solver carries the cylinder at c_y=64 (Re=100) and the target at c_y=32 (Re=200), a displacement of two diameters. Horizons are quoted in units of the target shedding period, T_{shed}=40 snapshots; the measured rotation numbers per $T_{\mathrm{shed}}/4$ are 0.233/0.247 (parametric) and 0.199/0.223 (structural).

Representation. Each system is reduced with its *own* POD basis (SVD of 5,000 mean-centred snapshots; d=16 modes, $\geq 99.8\%$ variance), and the two coefficient spaces are never registered. Using separate bases preserves the coordinate mismatch between systems.

Conjugator and transports. The iEDMD conjugator is a learned diffeomorphism g (affine coupling flow (Dinh et al., 2016), 6 coupling layers, width 128) with a linear reduced operator K_V estimated at a lag of 2 snapshots and multistep consistency over 8 steps, trained with an MMD term on $g_\#$; since g is exactly invertible, $K_U := g^{-1} \circ K_V \circ g$ by construction. All wake transports share one Flow-Matching architecture (MLP, 3 residual blocks of width 256, input noise 0.4, weight decay

10^{-3}, 400 epochs). The primary iEDMD results use coupling alone, without a penalty differentiated through the flow: the unsquared operator cost $\|u\| + \|K_V u\|$, $u = y - g(x)$ in per-side normalized POD coordinates, defines a precomputed Sinkhorn plan on blocks of 8,192 states. We use $\varepsilon = 0.02$ times the mean block cost and iterate until the maximum row-sum deviation, after multiplying the unit-mass plan by the block size, is below 10^{-3} (cap 5,000 iterations). One target is sampled per source row; FM is trained on the resulting cached pairs, with seeds 0, 1, and 2.

The Euclidean wake baseline uses unsquared distance with $\varepsilon = 0.5$ times the median cost. Thus the wake comparison does not hold the entropic scale fixed across costs. On the structural pair, a coupling-only iEDMD control at this same median-relative scale reaches 0.161 ± 0.008 at a quarter period (three seeds), compared with 0.114 under the mean-relative protocol (three seeds) and 0.441 for the Euclidean baseline. The iEDMD transports with the generated-state penalty $0.05\, c_g(x, T(x))$ are retained as explicitly labelled ablations. They integrate and differentiate the flow during training, so they are not simulation-free.

Full-state pinball transport protocol. The displaced-cylinder pinball uses separate 32-mode POD representations and the velocity-augmented dictionary of Appendix I.2. Its coupling uses the squared operator cost $\|u\|^2 + \|K_V u\|^2$, with the same converged Sinkhorn settings as the ODE experiments ($\varepsilon = 0.02\bar{C}$, tolerance 10^{-3}, batch 2,048). FM uses width 128, 4 blocks, 600 epochs, learning rate 10^{-3}, weight decay 0.01, global normalization and seeds $42/43/44$ for every coupling, with the conjugator held fixed. The velocity conjugator has $C_\phi = 32$, operator lag 0.25, no identity anchor and 400 training epochs; the state-only controls are specified in Appendix I.2. The source-to-delay experiment uses the distinct protocol of Appendix I.3.

Simulator-in-the-loop evaluation. For each of 200 source starts, we transport and decode through the target POD basis into a velocity field, initialize the target solver with it (velocities inside the solver's own solid mask are zeroed and densities set to unity before equilibration), advance $k\, T_{\mathrm{shed}}/4$, and compare velocity fields against the decoded transported truth $T(x_{t+k})$, normalized by the standard deviation of target velocity fields. The wrong-model row initializes the *source* solver (wrong Re; and, structurally, wrong cylinder position — the state enters through the same decode-and-mask procedure, so the flow initially occupies the region where the source solver has no obstacle) from the exact target state and is compared to the true target future. Full-state pinball uses starts in the held-out block $[17500, 20000)$ and horizons $15/29/59$ snapshots; its conjugation error divides each velocity RMSE by the standard deviation of the corresponding decoded reference field. Source-to-delay scoring is defined separately in Appendix I.3.

G Full results

G.1 Forecasting: complete results

This extends the ODE rows of Table 2 to all thirteen pairs and three horizons, using the same archived evaluations. It is the full report of that forecast experiment, rather than a separate benchmark. Complete fluid forecasts appear in Table 16.

Transport improves forecasting on the deformed pairs and several shared-coordinate pairs, while it can degrade an already accurate direct forecast on Rössler and Lorenz-96. On Rössler c the two systems nearly coincide in native coordinates (direct error 0.009 at the first index), so any learned map adds its own approximation error; the solver-free selection rule keeps the identity on this pair (Table 15). These are comparisons at the reported horizons, not a validated rule for selecting a transport from short-horizon error alone.

G.2 Attractor proximity after initialisation

The Lorenz-63 panel of Figure 2 uses the native-coordinate parameter pair $\rho = 28 \to 30$, with no observation deformation. From the same 2,000 exact target states, we compare $F_t(y_0)$ and $T(F_t(T^{-1}y_0))$ over $t \in [0, 0.5]$ with output spacing 0.005. Proximity to the target attractor is

Table 8: Cross-model forecast RMSE (↓): three seeds and 200 common target starts. Each cell reports the three evaluation horizons $t_1/t_2/t_3 = 0.24/0.49/0.98$ (Rössler: $1.20/2.45/4.90$). Direct denotes $F_X^t(y_0)$; the other columns report $T(F_X^t(T^{-1}(y_0)))$. Values are seed means, aggregated over starts by the mean except for Rössler by the segment median.

Pair	direct	random	Euclidean	iEDMD
		$t_1/t_2/t_3$		
h(L63), $28 \to 34$	10.971/13.633/13.652	11.200/11.774/12.696	7.828/8.817/13.567	2.993/5.422/8.981
Chen a, $h_{\mathrm{rot+nl}}$	14.812/12.668/13.092	8.534/10.962/9.947	4.933/7.596/9.074	2.079/6.128/9.094
Rössler a, $h_{\mathrm{rot+nl}}$	3.665/4.303/4.176	3.860/3.051/3.792	0.773/1.081/1.258	0.267/0.340/0.464
Chen a, h_{rot}	16.734/12.024/12.380	7.702/9.707/8.795	1.884/5.229/8.316	1.268/4.597/8.213
Rössler a, h_{rot}	2.895/3.888/3.567	4.208/3.796/3.829	0.428/0.877/0.735	0.269/0.367/0.414
L63 ρ, $28 \to 34$	4.731/7.388/9.045	4.385/6.030/9.574	2.713/4.418/7.466	2.753/4.602/7.657
Chen a	1.923/6.037/9.032	1.226/4.701/7.647	1.036/4.185/7.681	1.006/4.263/7.687
Rössler a	0.070/0.112/0.273	2.013/1.888/2.321	0.131/0.177/0.255	0.131/0.190/0.255
L63 ρ, $28 \to 30$	1.604/2.986/5.013	1.307/2.168/4.066	0.515/1.042/2.459	0.507/1.060/2.380
L63 σ	0.573/1.033/1.749	2.949/4.721/5.601	0.492/0.946/1.680	0.500/0.950/1.698
Chen b	0.737/2.427/6.191	0.749/1.743/4.966	0.609/1.825/5.150	0.578/1.682/4.914
Rössler c	0.009/0.030/0.486	3.286/3.501/3.270	0.241/0.379/0.918	0.193/0.336/0.884
Lorenz-96 (d=40)	0.488/1.425/3.685	1.472/2.647/4.668	1.220/2.240/4.219	1.230/2.283/4.290

measured by

$$d_{\mathrm{ref}}(z) = \frac{\min_{a \in A_Y^{\mathrm{ref}}} \|z - a\|_2}{r_Y}, \qquad r_Y^2 = \frac{1}{|A_Y^{\mathrm{ref}}|} \sum_{a \in A_Y^{\mathrm{ref}}} \|a - \bar{a}\|_2^2.$$

The reference contains $200{,}000$ target states, spaced by 0.01 time units after a 100-unit burn-in. Initial states come from a separately initialised trajectory after the same burn-in, spaced by 2 time units. Neither the initial states nor their continuations enter the reference cloud.

The illustration uses the coupling-only iEDMD transports of Table 9 (seeds $42, 43, 44$). For each initial state we average distances over the three transports, then report the mean and sample std across states. Bands measure initial-state dispersion, not confidence intervals or training-seed uncertainty. Curves are unsmoothed. The dashed reference evolves the same states under the true target dynamics; its positive distance calibrates finite reference-cloud sampling.

At $t = 0.5$, mean normalized distances are 0.0255 for the direct forecast, 0.00656 for the transported forecast and 0.00376 for the true target continuation. The transport is approximate: its residual remains above the sampling floor. On the wider native-coordinate pair $\rho = 28 \to 34$, the corresponding values are 0.0646, 0.0289 and 0.00384. Thus a larger parameter mismatch gives a larger remaining deviation; neither example establishes exact attractor membership, correct phase or equality of invariant measures.

Both native-coordinate diagnostics save individual distances, checkpoint and code hashes, round-trip errors (at most 8×10^{-7} in normalized units) and checks with twice as many transport integration steps (changes below 10^{-5}).

G.3 ODE CONJUGATION AND MARGINAL FIDELITY

Table 9 reports the additional low-dimensional ODE pairs and the native-coordinate Lorenz-96 control omitted from Table 1, together with conjugation and marginal fidelity. The latter is measured by $\mathrm{MMD}^2(T_{\#}\mu_X, \mu_Y)$, normalized by the same statistic between two independent target subsamples of $4{,}000$ states; a ratio near one is comparable to this finite-sample reference, not evidence of equality. On the rotated Rössler pairs, random coupling reaches 158–171 times the floor, against 5.1–5.2 for Euclidean and 2.3–4.3 for iEDMD coupling; on deformed Lorenz, the two OT couplings are near 5, against 46 for random pairing. Regularised transports are reported in Appendix H.2.

Table 9: Conjugation RMSE at the first evaluation horizon (mean±std over seeds; all Rössler pairs in median) and pushforward MMD divided by the floor for additional ODE pairs and the native Lorenz-96 control, using the same Flow Matching model and coupling-only training. The ring-deformed Lorenz-96 pair is reported in Table 1.

	conjugation RMSE (↓)			MMD^2/floor (↓)		
Pair	random	Euclidean	iEDMD	random	Euclidean	iEDMD
h(L63), $28 \to 34$	13.607±0.408	7.862±0.016	**2.995±0.120**	46	5	5
Chen a, $h_{\mathrm{rot+nl}}$	9.192±0.494	5.653±0.058	**2.496±0.181**	13	3	2
Rössler a, $h_{\mathrm{rot+nl}}$	3.885±0.124	0.669±0.048	**0.284±0.079**	158	5	2
Chen a, h_{rot}	7.046±0.323	2.386±0.001	**1.351±0.097**	15	3	2
Rössler a, h_{rot}	3.812±0.133	0.442±0.055	**0.255±0.029**	171	5	4
L63 ρ, $28 \to 34$	4.310±0.456	**2.674±0.013**	2.801±0.320	11	12	10
Chen a	1.286±0.060	1.085±0.015	**1.005±0.001**	1	2	2
Rössler a	1.225±0.144	**0.120±0.001**	0.124±0.000	36	6	6
L63 ρ, $28 \to 30$	1.174±0.452	0.513±0.001	**0.511±0.002**	3	1	1
L63 σ	2.752±0.947	**0.522±0.001**	0.534±0.006	7	1	1
Chen b	0.834±0.267	0.672±0.009	**0.614±0.022**	1	2	2
Rössler c	1.921±0.030	**0.217±0.011**	0.218±0.004	63	1	1
Lorenz-96 (d=40)	1.677±0.001	**1.258±0.001**	1.277±0.000	2	1	1

H CONTROLS

H.1 SOLVER-IN-THE-LOOP REFERENCE

We compare iEDMD coupling and operator-cost penalties with a Flow Matching map regularised by the true conjugation residual,

$$\mathcal{R}_\tau(\theta) = \mathbb{E}_{X\sim\mu_X} \left\| T_\theta(F_\tau X) - M_\tau(T_\theta X) \right\|^2, \tag{29}$$

using RK4 with differentiation through the target solver and the transport. The comparison covers the deformed Lorenz pair, the rotated and deformed Chen a and Rössler a pairs, widened L63 ($\rho : 28 \to 34$) and Chen a, with three seeds and the protocol of Table 1. Regularisers are applied to 256 states every fourth training step with a 10-step differentiable transport. Reference lags are 0.1 on the Lorenz pairs and 0.05 on Chen a.

Calibrating the strength. Each regulariser has one absolute weight per pair, shared by the three couplings and seeds. It is set once, on the first training batch of the Euclidean coupling at initialisation (seed 42), so that the weighted term equals three times the Flow Matching loss. Penalised runs clip the gradient norm at 5.

The reference gives 1.91 and 1.36 on the two Lorenz pairs, against 2.16 and 1.89 for the operator penalty on the iEDMD coupling; on Chen a the two are equal (0.689 and 0.683; Table 11). On the widened pair the reference doubles the pushforward MMD^2 ratio (21.2, against 10.0 for the coupling alone), whereas the operator penalty leaves it unchanged (9.9).

H.2 THE iEDMD COST BEYOND THE COUPLING

The iEDMD cost $c(x, y)$ builds the coupling from observed pairs. Two further uses test what it encodes: extended along orbits, it links consecutive states; evaluated at generated pairs, $c(x, T(x))$, it becomes a penalty on the transport. All runs follow the protocol of Table 1: same data, conjugators, Flow Matching architecture, optimiser, budget, weight decay 10^{-3}, Sinkhorn rule, three seeds and prospective split.

Trajectory coupling. For each candidate source–target match, we form two trajectory segments of K states, sampled at the autocorrelation stride. The coupling criterion is evaluated jointly on these segments, so that Euclidean and iEDMD coupling have access to the same temporal window. Once two segments are matched, their K time-aligned state pairs are used as separate Flow Matching training examples. Thus, K=1 recovers statewise coupling, whereas $K > 1$ lets every method use

short-time evolution. On the ring-deformed Lorenz-96 pair (Table 10), Euclidean coupling improves from 4.12 at $K{=}1$ to 3.94 at $K{=}4$. The iEDMD coupling remains better for every K, reaching 3.82 at $K{=}4$, while random coupling is essentially unchanged. Comparing $K{=}1$ (3.892) with the statewise result in Table 1 (3.883) shows no measurable contribution from the successor term on this pair; the small gain comes from the learned alignment.

Table 10: Orbit coupling on Lorenz-96 ($d{=}40$, h_{ring}): conjugation RMSE at $t{=}0.24$, mean±std over three seeds. Candidate matches are scored over segments of K states (stride equal to the autocorrelation lag), and Flow Matching is trained on the resulting K time-aligned pairs. At $K{=}1$ the iEDMD cost has no successor term.

Linked states K	random	Euclidean	iEDMD
1	4.364±0.024	4.115±0.001	**3.892±0.136**
2	4.345±0.005	4.060±0.001	**3.888±0.042**
4	4.350±0.015	3.936±0.001	**3.820±0.035**
8	4.354±0.032	3.956±0.003	**3.893±0.036**

The cost as a penalty. Table 11 adds to Flow Matching the operator cost evaluated at transported states, $\lambda\,\mathbb{E}[\|G\Psi(x)-\Psi(Tx)\|^2+\|K_V(G\Psi(x)-\Psi(Tx))\|^2] = \lambda\,\mathbb{E}[c^{\mathrm{op}}_{\mathrm{iEDMD}}(x,T(x))]$ in its squared form, for the three couplings, and compares it with the solver-in-the-loop reference. The penalty is applied to 256 states every fourth step with one absolute λ per pair (Appendix H.1). It integrates the transport once per regularised batch and the reference twice, both with ten RK4 steps and backpropagation through the trajectories. Unlike these penalties, using the cost in the coupling requires no solver call, no query of learned dynamics at generated states, and no integration or differentiation through the transport during training.

Evaluated at generated pairs, the cost lowers the conjugation error of every coupling, including random pairing, which carries no dynamical information (13.6 → 2.06 on the deformed Lorenz pair). The coupling still matters on the hardest pairs: on Chen a and Rössler a with $h_{\mathrm{rot+nl}}$, the penalised iEDMD coupling reaches 1.92 and 0.142, against 2.93 and 0.306 for random pairing with the same penalty, close to the solver-in-the-loop reference (1.72 and 0.119); the iEDMD coupling alone, without penalty or integration of the transport, already matches penalised random pairing (2.50 and 0.284). On the native pairs the couplings coincide once penalised.

H.3 Gromov–Wasserstein control

Gromov–Wasserstein (GW) aligns within-domain distances and is therefore invariant to rigid changes of coordinates, but it neither handles general non-isometries nor enforces dynamical intertwining (Mémoli, 2011; Peyré et al., 2016). Our dense entropic implementation uses $O(N^2)$ memory and $O(N^3)$ arithmetic per outer iteration, whereas the iEDMD cost yields a linear OT problem with $O(N^2)$ work per Sinkhorn iteration after the alignment has been learned. Low-rank GW can reduce this cost (Scetbon et al., 2022).

Fixed-support comparison. We compare GW and Euclidean couplings on independently sampled supports of $N = 4096$ or 8192 states for rotated Chen ($a : 35 \to 36$) and nonlinearly deformed Lorenz-63 ($\rho : 28 \to 34$). Both use normalized squared distances, entropy coefficient 0.02, and the same FM architecture and training budget; GW retains the best converged solution from three initializations. At $N = 8192$, GW improves over Euclidean coupling on the rigid Chen transformation (1.488 versus 2.298) but degrades on the Lorenz deformation (10.953 versus 7.981). These fixed-support values are not directly comparable to the full-pool iEDMD references (1.351 and 2.995). On Chen, rigid Procrustes maps fitted to the GW plans reach 1.234 and 1.079 at the two support sizes, showing that FM can lose alignment already present in the geometric plan.

Exact-conjugacy control. On 200 fixed Lorenz states, recovery of the known correspondence by GW falls from 100% under a rigid map to 63%, 23.5%, and 14% under increasingly nonlinear triangular deformations. For the largest deformation, GW even attains lower geometric distortion than the known correspondence, so the exact conjugacy is not a geometric GW minimizer. We then train FM on 4096 states using either the GW plan or the known pairs, for a rigid map and a large

Table 11: Regularized transports, protocol of Table 1. Conjugation RMSE at the three fixed indices ($t_1, t_2, t_3 = 0.24, 0.49, 0.98$; Rössler: $1.20, 2.45, 4.90$, segment median), mean±std over three seeds; pushforward MMD^2 on the training pools divided by its sampling floor. Bold: lowest error at t_1.

		E_{conj} (↓)			marginal / floor
Pair	Map	t_1	t_2	t_3	MMD^2
h(L63), $28 \to 34$	random + pen.	2.062±0.131	4.687±0.043	8.093±0.233	9.51
	Euclidean + pen.	2.947±0.114	5.583±0.137	9.643±0.137	10.27
	iEDMD + pen.	2.157±0.072	4.388±0.158	8.299±0.155	6.38
	iEDMD + reference	**1.905±0.051**	4.719±0.037	8.514±0.081	6.82
Chen a, $h_{\mathrm{rot+nl}}$	random + pen.	2.928±0.109	7.525±0.244	10.257±0.618	10.04
	Euclidean + pen.	2.014±0.024	4.961±0.025	9.458±0.146	2.85
	iEDMD + pen.	1.916±0.109	5.267±0.249	9.219±0.260	1.31
	iEDMD + reference	**1.724±0.147**	4.399±0.082	8.692±0.165	2.13
Rössler a, $h_{\mathrm{rot+nl}}$	random + pen.	0.306±0.007	0.302±0.024	0.421±0.050	4.45
	Euclidean + pen.	0.565±0.059	0.597±0.081	0.771±0.044	8.54
	iEDMD + pen.	0.142±0.014	0.227±0.010	0.354±0.026	2.39
	iEDMD + reference	**0.119±0.011**	0.175±0.017	0.314±0.042	2.29
L63 ρ, $28 \to 34$	random + pen.	1.683±0.062	3.918±0.111	7.498±0.191	11.46
	Euclidean + pen.	1.610±0.115	3.720±0.013	7.466±0.222	15.38
	iEDMD + pen.	1.890±0.057	4.092±0.172	7.712±0.033	9.89
	iEDMD + reference	**1.361±0.001**	3.190±0.001	6.776±0.023	21.21
Chen a	random + pen.	0.674±0.004	2.982±0.014	7.567±0.003	2.86
	Euclidean + pen.	**0.667±0.004**	2.974±0.014	7.581±0.006	2.89
	iEDMD + pen.	0.683±0.006	3.000±0.022	7.615±0.018	2.85
	iEDMD + reference	0.689±0.001	3.068±0.005	8.124±0.008	3.97

deformation. The known-pair arm verifies that the architecture can represent the exact conjugacy; a further control rigidly prealigns the FM inputs while retaining the GW plan (Table 12).

FM pairing	ID (%)	Mass (%)	$E_{.24}$	$E_{.98}$	MMD^2/floor
Rotation and translation					
GW	99.35 ± 0.26	1.94 ± 0.02	0.144 ± 0.003	0.889 ± 0.007	1.4 ± 0.0
Oracle h	100.00 ± 0.00	100.00 ± 0.00	0.081 ± 0.002	0.566 ± 0.008	1.4 ± 0.0
Large deformation					
GW	0.12 ± 0.06	0.04 ± 0.00	9.022 ± 0.256	10.618 ± 0.509	25.1 ± 4.1
GW + prealignment	0.12 ± 0.06	0.04 ± 0.00	2.677 ± 0.093	5.192 ± 0.694	4.7 ± 0.3
Oracle h	100.00 ± 0.00	100.00 ± 0.00	0.871 ± 0.030	2.011 ± 0.103	1.5 ± 0.0

Table 12: Exact Lorenz conjugacies, mean±sample standard deviation over three seeds. ID and Mass measure recovery of the known pairs; E_t evaluates the learned FM map and allows any valid dynamical correspondence. Oracle rows use the known pairs, while prealignment applies a proper rigid transformation before FM. Marginal discrepancy is MMD^2 divided by its sampling floor.

Under the rigid transformation, GW recovers the correspondence and approaches the oracle. Under the large deformation, recovery falls to 0.12% and the conjugation error rises to 9.022, whereas oracle training gives 0.871. Rigid prealignment lowers the error to 2.677 without changing the GW plan, showing that map approximation explains part, but not all, of the gap. Thus GW is a useful geometric control under near-isometries, while the dynamical cost targets the non-isometric correspondences relevant here.

H.4 Marginal transport: Flow Matching and the conjugator readout

The conjugator supplies the cost, not the map. Its state readout $T_g = P_{\mathrm{state}} \circ g \circ \Psi$ serves as a control: can the transport step be skipped? Table 13 compares $W_2(T_{\#}\mu_X, \mu_Y)$ on the prospective test split

for the readout and the three couplings, by exact uniform assignment on $4{,}096$ states (five draws, seed 42) in native coordinates; "floor" is the same distance between two disjoint target samples, what a perfect map scores with this many points. Every transport uses the validation-selected weight decay 10^{-3} of Appendix B.

A transport is the best map on 9 of 12 pairs. On L63 σ it sits at the floor (0.93 against 0.96) where the readout is 35% above it; on Rössler c it reaches 0.77 against 1.37. The readout keeps a 1–4% edge on the widened L63 pair, Chen b and rotated Rössler a. Lorenz-96 is excluded because at $d = 40$ the assignment floor is 59% of the attractor radius and shrinks only as $n^{-1/40}$, so this diagnostic cannot rank maps reliably.

Table 13: Transport against the direct conjugator readout, $W_2(T(x), Y)$ in native units (↓), $4{,}096$ test states, seed 42. Bold: best map.

Pair	floor	G readout	iEDMD FM	ℓ_2 FM	random FM	FM vs G
h(L63), $28 \to 34$	1.20	5.32	**5.20**	5.30	6.96	-2.4%
Chen a, $h_{\mathrm{rot+nl}}$	1.10	1.77	**1.54**	1.94	2.59	-13.0%
Rössler a, $h_{\mathrm{rot+nl}}$	0.60	0.77	**0.71**	0.82	2.75	-7.2%
Chen a, h_{rot}	1.07	1.62	**1.47**	1.64	2.63	-9.2%
Rössler a, h_{rot}	0.61	**0.77**	0.80	0.89	2.87	$+3.6\%$
L63 ρ	1.04	2.14	1.81	**1.79**	2.16	-15.4%
L63 σ	0.96	1.30	0.93	**0.92**	2.58	-28.7%
L63 ρ (wide)	1.12	**2.50**	2.53	2.57	3.07	$+1.2\%$
Chen a	1.07	1.45	1.32	**1.32**	1.34	-9.0%
Chen b	1.17	**1.32**	1.35	1.36	1.34	$+2.4\%$
Rössler a	0.61	0.92	**0.87**	0.87	1.30	-5.3%
Rössler c	0.71	1.37	0.77	**0.76**	1.63	-44.0%

H.5 Does the learned map realize the optimized coupling?

The successor bound of Section 3.2 controls the residual of a map through the coupling $\pi_T = (\mathrm{Id}, T)_{\#}\mu_X$ it induces, while the algorithm optimizes a plan π_B and fits Flow Matching to it; conditional Flow Matching reproduces the marginals of π_B, not necessarily π_B itself. We measure how much of the plan's pairing the learned map keeps.

ODE pairs. Table 14 compares, on $2{,}048$ test states per pair, the learned maps with the exact assignment of each training cost: the iEDMD successor cost on physical states at the training lag, and the Euclidean cost in the normalized coordinates of Flow Matching. The floor is the distance between the partners that the same cost assigns in two independent target samples, the resolution of a finite-sample plan. On the deformed pairs the iEDMD map stays within 0.10–0.20 radii of its plan, 1.2–3.0 times the floor, while the Euclidean map lies 1.1–3.2 times farther from that plan. In shared coordinates both maps are within 0.07–0.15 radii of either plan, close to the floor. The final Flow Matching loss is 0.80–1.09 times that of the Euclidean coupling on every reported pair, so the iEDMD plans are not harder for Flow Matching to realize than the Euclidean ones. Inversion is not a factor: the round trip $\|T^{-1}(Tx) - x\|/\|x\|$ of the 300-step RK4 transport is at most 10^{-6} on every pair.

H.6 Does the surrogate cost predict downstream conjugation error?

The solver-free operator score of a transport is $\mathbb{E}[\|u\| + \|K_V u\|]$, $u = \Psi(Tx) - g\Psi(x)$, evaluated on training-side source states without querying the target dynamics. It uses the same learned g and K_V as the iEDMD coupling, so the experiment is prospective with respect to solver-based errors rather than an independent model-selection test.

Prospective selection. Table 15 uses the operator score as a selection rule. For every pair, the identity, the affine moment map and the ℓ_2- and iEDMD-coupled transports (three seeds) are scored on $2{,}048$ training-side source states with the pair's fixed conjugator; no target solver and no test state enters the score. The map with the lowest score is then read on the prospective test split. The rule selects the oracle map on 11 of 13 pairs and lands within 10% of the oracle error on 12; its median

Table 14: Learned map against its training plan, seed 42, prospective test split, 2,048 states; median distances in units of the target attractor RMS radius. Plans are exact assignments of the training costs; floor: the same cost against an independent target sample. FM loss: final training loss of the iEDMD coupling over that of the Euclidean coupling, mean over three seeds. Round trip: maximum relative $\|T^{-1}(Tx) - x\|$ over both couplings.

	iEDMD plan			Euclidean plan		FM loss	round trip
Pair	floor	iEDMD map	Eucl. map	floor	Eucl. map	iEDMD/Eucl.	max
h(L63), $28 \to 34$	0.067	0.200	0.373	0.064	0.282	1.02	5×10^{-7}
Chen a, $h_{\mathrm{rot+nl}}$	0.084	0.119	0.217	0.076	0.090	0.80	7×10^{-7}
Rössler a, $h_{\mathrm{rot+nl}}$	0.080	0.099	0.316	0.077	0.101	0.85	1×10^{-6}
Chen a, h_{rot}	0.093	0.132	0.145	0.084	0.093	0.84	3×10^{-7}
Rössler a, h_{rot}	0.083	0.109	0.192	0.081	0.104	0.96	5×10^{-7}
L63 ρ, $28 \to 34$	0.096	0.148	0.158	0.087	0.141	0.99	3×10^{-7}
Chen a	0.094	0.107	0.109	0.084	0.083	1.01	2×10^{-7}
Rössler a	0.083	0.114	0.117	0.082	0.111	1.01	1×10^{-7}
L63 ρ, $28 \to 30$	0.112	0.097	0.097	0.100	0.097	0.94	2×10^{-7}
L63 σ	0.075	0.080	0.080	0.070	0.074	0.96	1×10^{-7}
Chen b	0.091	0.112	0.111	0.083	0.097	0.96	1×10^{-7}
Rössler c	0.055	0.071	0.079	0.054	0.070	1.06	4×10^{-7}

test error is 0.559, against 0.522 for the oracle and 0.614 for always choosing iEDMD. It correctly declines to transport on the three registered pairs where the identity is best (Rössler a, Rössler c, Lorenz-96) and picks the affine map on Chen b. Its two misses are pairs where the candidates are nearly tied: the widened L63 pair, where the two OT couplings differ within the seed spread, and L63 σ, where it prefers the identity (0.166 against 0.172 for iEDMD) and pays 0.65 instead of 0.52.

Table 15: Prospective selection by the solver-free operator score (↓; training-side source states, fixed conjugator, FM scores averaged over three seeds; bold: selected). Right: conjugation error of the selected map on the prospective test split, against the oracle.

	operator score (training side)					test E_{conj}		
Pair	identity	affine	ℓ_2	iEDMD	selected	selected	oracle	oracle map
h(L63), $28 \to 34$	1.867	2.485	1.409	**0.516**	iEDMD	2.995	2.995	iEDMD
Chen a, $h_{\mathrm{rot+nl}}$	2.407	3.741	0.787	**0.390**	iEDMD	2.496	2.496	iEDMD
Rössler a, $h_{\mathrm{rot+nl}}$	2.377	2.110	1.333	**0.325**	iEDMD	0.284	0.284	iEDMD
Chen a, h_{rot}	2.025	1.544	0.432	**0.341**	iEDMD	1.351	1.351	iEDMD
Rössler a, h_{rot}	2.050	1.562	0.932	**0.318**	iEDMD	0.255	0.255	iEDMD
L63 ρ, $28 \to 34$	0.757	0.431	0.435	**0.400**	iEDMD	2.801	2.674	ℓ_2
Chen a	0.292	0.194	0.195	**0.183**	iEDMD	1.005	1.005	iEDMD
Rössler a	**0.129**	0.168	0.178	0.200	identity	0.072	0.072	identity
L63 ρ, $28 \to 30$	0.292	0.151	0.149	**0.141**	iEDMD	0.511	0.511	iEDMD
L63 σ	**0.166**	0.179	0.175	0.172	identity	0.652	0.522	ℓ_2
Chen b	0.143	**0.099**	0.109	0.104	affine	0.559	0.559	affine
Rössler c	**0.288**	1.309	0.831	0.699	identity	0.014	0.014	identity
Lorenz-96 (d=40)	**1.692**	2.943	3.896	3.957	identity	0.469	0.469	identity

I FLUID BENCHMARKS: ADDITIONAL MATERIAL

I.1 VON KÁRMÁN WAKES

Table 16 separates matched forecasts from conjugation errors. The complete conjugation curves appear in Figure 3. At longer horizons, rotation-number mismatch produces phase drift; geometric and random couplings have oscillatory errors and can attain lower averages at some horizons. Marginal diagnostics also show amplitude contraction for the Euclidean transport (Figure 4).

Using precomputed 8,192-state blocks for the Euclidean coupling gives conjugation errors $0.438/0.319/0.433$, compared with $0.441/0.330/0.446$ at $0.25/0.5/1T$ under its minibatch pro-

tocol (three seeds). For matched structural forecasts, nearest-neighbour initialisation from $4{,}000$ decoded source fields gives 0.69 at a quarter period. Translating the target field by the known cylinder displacement gives 0.481. Both exceed the direct wrong-model error (0.325) and the learned round-trip error (0.074).

Coupling-cost ablation. On the structural pair, adding raw observed successors to the Euclidean cost changes essentially nothing ($0.441 \to 0.429$): across unregistered bases, successor coordinates are as arbitrary as state coordinates, so temporal information alone is not the ingredient. A learned *linear* intertwiner over EDMD features recovers about a third of the gap (0.305), and the nonlinear iEDMD cost accounts for the remaining short-horizon gain (0.114 without a generated-state penalty). The random baseline reaches 0.301, essentially the linear-alignment score; the nonlinear iEDMD coupling provides the substantial improvement.

In the coupling comparison on a separate common draw, coupling-only FM has coefficient-space RBF MMD^2 0.00098, versus 0.00670 for Euclidean and 0.00859 for random coupling (sampling floor 0.00017). Squared sliced-W_2 gives the same ordering (0.00827, 0.18075, 0.28372); Laplacian MMD, energy distance and Kolmogorov–Smirnov statistics agree. Figure 4 illustrates the penalised variant, whose per-mode variance ratios are 0.96–1.04, compared with 0.46–0.73 for the Euclidean and random flows. Direct-readout comparisons with learned dictionaries are reported in Appendix H.4.

Matched-initial-state protocol. Every forecast starts from the same observed target states $y_0 \sim \mu_Y$ and is scored against the true target future $\mathcal{M}^k(y_0)$. On the structural pair we compare the wrong-model forecast $F^k(y_0)$ with the FM round-trip $T(F^k(T^{-1}y_0))$, using backward ODE integration for the inverse. F is the misspecified source solver and all decoding/reprojection is done with the appropriate per-side bases (200 starts, velocity-field metric). Coupling-only results use three FM seeds (mean$\pm$sample std); the penalized forecast uses seed 0. The horizons up to $1T$ are reported in Table 16. At $2T/3T$, coupling-only FM gives $0.456 \pm 0.002/0.591 \pm 0.001$, against $0.664/0.732$ for the direct forecast; the penalized variant gives $0.465/0.526$. Thus coupling-only FM is best through $2T$, while the penalized variant is lower at $3T$.

Table 16: Both wake regimes at all horizons (velocity-field RMSE, true solver in the loop). **Top:** matched forecasts from the same observed y_0. **Bottom:** conjugation error of the transports (different quantity). $\pm$ = sample std ($n-1$ denominator) across 3 FM training seeds where available. On the parametric pair, coupling-only FM is close to the direct model at $0.25T$ and has higher error from $0.5T$ onward. Matched FM forecasts use 200 target starts and three seeds on both pairs.

	Parametric			Structural		
Matched forecast	0.25	0.5	1	0.25	0.5	1
wrong-model (exact state)	0.064	**0.082**	**0.118**	0.325	0.422	0.547
FM round-trip (coupling only)	**0.062**$\pm$**0.001**	0.113$\pm$0.001	0.204$\pm$0.001	**0.074**$\pm$**0.002**	**0.146**$\pm$**0.007**	**0.257**$\pm$**0.004**
Conjugation error	0.25	0.5	1	0.25	0.5	1
FM, random coupling	0.238$\pm$0.055	0.273$\pm$0.038	0.259$\pm$0.016	0.301$\pm$0.020	0.331$\pm$0.033	0.328$\pm$0.009
FM, Sinkhorn ℓ_2	0.280$\pm$0.032	0.298$\pm$0.013	0.240$\pm$0.025	0.441$\pm$0.038	0.330$\pm$0.016	0.446$\pm$0.030
FM, iEDMD coupling only	**0.083**$\pm$**0.0001**	**0.109**$\pm$**0.0002**	**0.142**$\pm$**0.0005**	**0.114**$\pm$**0.004**	**0.155**$\pm$**0.002**	**0.247**$\pm$**0.002**
FM, iEDMD + penalty	0.076$\pm$0.00004	0.095$\pm$0.0001	0.140$\pm$0.0001	0.100$\pm$0.002	0.144$\pm$0.002	0.242$\pm$0.004

I.2 Fluidic pinball: what makes the coupling cost informative

This subsection concerns full-state POD transport. The experiment with a full source state and a target delay embedding is reported separately in Appendix I.3. The dictionary and marginal controls below use seed 42; Table 1 reports the three-FM-seed comparison.

Dictionary ablation. Across two state-only settings, the learned and Euclidean costs remain strongly correlated (Pearson 0.90–0.987) and the conjugation error stays between 0.359 and 0.376 at $0.25T$. Adding a velocity feature $\hat{v}(z)$, regressed from finite differences of the coefficient time series ($R^2 = 0.96$–0.99), reduces the cost correlation to 0.46 and the error to 0.256, compared with 0.370

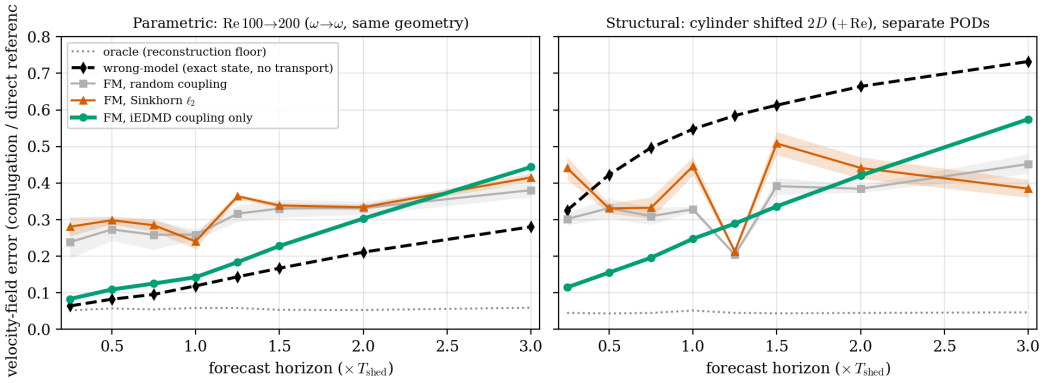


Figure 3: Velocity-field conjugation error versus horizon for the three coupling methods (bands: standard deviation across three training seeds). The direct wrong-model and reconstruction errors are shown as references and use a different initialisation protocol. Matched forecast comparisons are reported in Table 16.

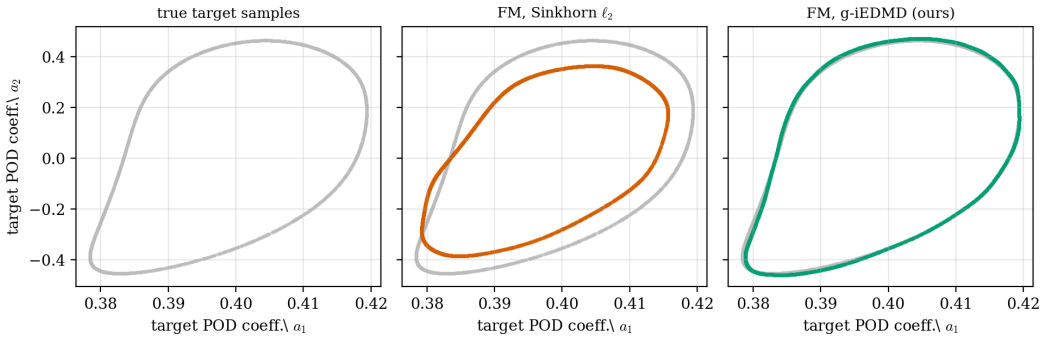


Figure 4: Pushforward of 6000 source snapshots, projected onto the leading target-POD plane of the structural wake. Target samples are gray. The Euclidean transport contracts inside the target cycle (mean radius 17% smaller). The green panel shows the iEDMD transport with a generated-state penalty (mean radius 2% larger); numerical marginal diagnostics for coupling-only FM are reported in the text.

for Euclidean coupling. The velocity-augmented dictionary also improves on Euclidean coupling over the tested horizons: by 4% at 3 snapshots, 11% at 7, 31% at 15 and 46% at 59 snapshots, approximately one characteristic shedding time.

Marginal fidelity and forecasts. For the flow, the matched round-trip variance ratio is 1.04–1.31 over 1–3 characteristic shedding times. Its forward coefficient-space marginal discrepancy is 17.3× the sampling floor against 30.3× for the ℓ_2 coupling, 31.5× for the random coupling and 47.7× for the identity (5 draws; the affine moment map reaches 9.5×). The matched forecast through the flow reaches 0.371/0.371/0.449 at 15/29/59 snapshots (seed 42), approximately 1.9 times lower than the wrong-model error. The separate round-trip marginal diagnostic gives MMD ≈ 3.2× the floor at 1, 2 and 3 characteristic shedding times.

I.3 PINBALL: FULL SOURCE STATE TO TARGET DELAY EMBEDDING

At Re = 130, we transport the displaced pinball's full POD32 state x_t to partial observations of the standard pinball. The target representation stacks its first eight POD coefficients at four causal times, $d_t = [o_t, o_{t-2}, o_{t-4}, o_{t-6}] \in \mathbb{R}^{32}$, following the delay coordinate idea of Takens (1981). One snapshot corresponds to 20 LBM steps. The POD observation operators are assumed known, and no unobserved target coefficient is used.

We retrain a six-layer affine-coupling alignment G (width 128) and a linear K_V with identity dictionaries. The successor cost compares the present states and their successors two snapshots later; iEDMD is trained for 200 epochs with eight-step prediction. The four FM transports use the same

three-block MLP (width 256), initialization and sampled supports within each seed, and are trained for 300 epochs. Sinkhorn plans are computed on three blocks of 4096 states with $\varepsilon = 0.1\bar{C}$. Besides random and iEDMD couplings, the controls use Euclidean distance either in POD/delay coordinates or between the corresponding vorticity reconstructions; the latter requires the known observation modes but no hidden target state.

The train, validation and test intervals are $[0, 14000)$, $[14500, 17000)$ and $[17500, 20000)$, with complete delay and forecast windows inside each split. Seven snapshots were selected on validation. At test time, we invert the transport, advance the unchanged source LBM, and transport back. We report coordinate-normalized RMSE on the eight current observations over 200 common starts; uncertainty is the sample standard deviation across three complete training seeds. Forward and inverse transports use 100 RK4 steps.

Table 17: Full source POD state to target delay embedding on pinball: observed-coordinate forecast NRMSE (↓), mean±sample std across three complete training seeds, 200 common test starts. Horizons are in snapshots. These errors differ from the full-field conjugation and forecast metrics of the full-state experiment.

Coupling	1	3	7	15	29
Random	0.194±0.008	0.518±0.014	1.019±0.024	1.153±0.011	1.104±0.027
ℓ_2 POD	0.112±0.005	0.283±0.016	0.633±0.010	0.993±0.008	1.203±0.008
ℓ_2 grid	0.158±0.003	0.389±0.007	0.769±0.025	0.984±0.016	**0.949±0.028**
iEDMD succ.	**0.102±0.003**	**0.261±0.012**	**0.511±0.036**	**0.816±0.132**	1.154±0.247