LEARNING ACROSS FEATURE DIMENSIONS VIA DIMENSION-CONDITIONED LATENT SCHRÖDINGER TRANSPORT

Anonymous authors
Paper under double-blind review

ABSTRACT

We study supervised learning when observations of the same type are represented by different numbers of features. The training data are grouped by feature dimension, and cross-dimensional sample identities are not used. The objective is a single predictor that operates on all training dimensions and transfers to dimensions not used for fitting. We introduce a dimension-conditioned encoder followed by a conditional latent transport to a common reference distribution. The transport is formulated as a soft-constrained multi-marginal Schrödinger bridge and approximated by conditional flow matching. A single task head is then trained on the transported representations. We give a uniform score-error decomposition, a dimension-coverage bound under an assumed per-dimension estimation rate, and a penalty-convergence result for the ideal soft bridge. Synthetic experiments examine shared prediction, dimension coverage, and finite-network penalty behavior. In an independent PBMC3k benchmark retraining, the final configuration attains mean accuracy 0.811, macro-F1 0.683, and macro-AUROC 0.947 across six feature dimensions. Dimension-specific classifiers remain stronger in accuracy and F1, which identifies the present cost of using one model across dimensions.

Keywords. Cross-dimensional learning, domain generalization, latent transport, Schrödinger bridge, single-cell classification

1 INTRODUCTION

Many learning pipelines assume a fixed input dimension. This assumption is restrictive when the available feature set varies across instruments, sites, acquisition budgets, or preprocessing choices. A separate model can be trained for each dimension, but this discards information shared across dimensions and provides no direct mechanism for a dimension that was absent during training.

We consider a collection of datasets indexed by their feature dimension. Each dataset contains task labels, but correspondences between samples in different dimensional views are not supplied to the learning algorithm. The goal is not to reconstruct an unobserved high-dimensional sample. Instead, it is to learn one task predictor by mapping every input to a common latent reference space. This narrower objective matches the classification experiments studied in this paper and avoids relying on an unidentifiable cross-dimensional sample correspondence.

The proposed construction has two components. A dimension-conditioned encoder embeds zero-padded inputs while preserving within-dimension geometry. A conditional latent flow then transports the encoded distribution of each dimension toward a reference distribution. The ideal transport is expressed as a soft-constrained Schrödinger bridge; the implementation uses a conditional flow-matching surrogate. The task head is shared across dimensions and is trained only after the encoder and bridge have acquired stable representations.

In preliminary experiments, simultaneously applying classification, distribution alignment, contrastive, and adversarial objectives caused latent collapse. Staged optimization, in which geometry and transport are fitted before task-specific regularization, was substantially more reliable. The resulting predictor on dimension d is simply the reference task head composed with the dimension-

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.

conditioned representation map. We refer to this composition as a pullback predictor; no high-dimensional data decoder is required.

The contributions of this paper are as follows.

1. We formulate classification across feature dimensions under a protocol that does not use cross-dimensional sample identities. The distinction between training dimensions and held-out dimensions is explicit in the learning objective.
2. We propose a staged method that combines a dimension-conditioned encoder, soft-constrained latent transport, and a shared task head. The method maps all inputs to a common reference distribution without training one predictor per dimension.
3. We establish three limited but transparent theoretical results: a score-composition error bound, an expected dimension-coverage bound conditional on a per-dimension estimation guarantee, and convergence of the ideal soft bridge as its marginal penalties grow. The assumptions and the scope of each result are stated explicitly.
4. We evaluate the method on synthetic regression and on PBMC3k cell-type classification. The experiments report both the benefit of staged training and the remaining gap to dimension-specific models.

The remainder of the paper is organized as follows. Section 2 reviews the relevant transport and flow literature. Section 3 defines the learning problem. Section 4 presents the method, and Section 5 gives the analysis. The experiments appear in Section 6.

2 Related Work

Optimal transport across representations. Optimal transport compares probability measures through a prescribed transport cost (Villani, 2003). When feature coordinates have a known common meaning, a Euclidean cost can be used directly. Across heterogeneous representations, however, specifying a meaningful cross-space cost is itself part of the problem. Gromov–Wasserstein (GW) distances address this issue by comparing within-space metric relations (Mémoli, 2011). We use this idea only as a within-batch geometry regularizer; it does not identify sample pairs across dimensions. Entropically regularized optimal transport and Sinkhorn iterations provide practical minibatch computations (Cuturi, 2013). Debiased Sinkhorn divergences remove the entropic self-bias and provide a discrepancy with zero self-distance (Feydy et al., 2019).

Schrödinger bridges and flows. The Schrödinger bridge problem seeks a path measure closest in relative entropy to a reference process while satisfying prescribed marginals (Schrödinger, 1931). Modern diffusion-based solvers connect this problem with score-based generative modeling (De Bortoli et al., 2021). In our setting, intermediate dimensions provide additional marginal information, but enforcing all empirical marginals exactly is neither necessary nor numerically convenient. We therefore use penalties for intermediate marginal discrepancies.

Continuous normalizing flows parameterize transformations through neural ordinary differential equations (Chen et al., 2018); normalizing flows more broadly provide tractable invertible transformations (Rezende and Mohamed, 2015). Flow matching fits a vector field by regression along conditional probability paths (Lipman et al., 2023). We use flow matching as a numerical surrogate for the latent bridge. This surrogate should not be confused with an exact solution of the entropy-minimization problem.

3 Problem Formulation

Dimension-indexed data. Let $\mathcal{D} \subset \mathbb{N}$ be a set of admissible feature dimensions. For each $d \in \mathcal{D}$, let $\mathcal{X}_d \subset \mathbb{R}^d$ be an input space and let μ_d be a distribution on $\mathcal{X}_d \times \{1, \dots, C\}$. A sample from μ_d is denoted by (x, y), where y is a task label whose semantic meaning is shared across dimensions.

The set of training dimensions is $\mathcal{D}_{\mathrm{tr}} = \{d_1, \dots, d_N\}$. For each d_i, the learner receives $S_{d_i} = \{(x_{i,j}, y_{i,j})\}_{j=1}^{m_i}$. The datasets are treated as separate collections. In particular, the training objective

does not use an identity linking $x_{i,j}$ to an observation in another dimension. A held-out dimension belongs to $\mathcal{D}_{\mathrm{te}} \subset \mathcal{D} \setminus \mathcal{D}_{\mathrm{tr}}$.

This protocol is weaker than paired multi-view supervision but stronger than fully unsupervised distribution matching because class labels are available within each dimension. It also differs from missing-feature imputation: the output of interest is the task label, not a completed feature vector.

Shared predictor and evaluation risk. Choose a reference latent space $\mathcal{Z} \subset \mathbb{R}^p$. A dimension-conditioned representation map is denoted by $\Phi_d : \mathcal{X}_d \longrightarrow \mathcal{Z}$. A single task head $g : \mathcal{Z} \to \Delta^{C-1}$ induces the predictor

$$f_d(x) = g(\Phi_d(x)), \tag{1}$$

where $\Delta^{C-1} := \{q \in \mathbb{R}^C_{\geq 0} : \sum_{c=1}^C q_c = 1\}$ is the probability simplex over the C classes. We use the multiclass cross-entropy loss $\ell : \Delta^{C-1} \times \{1, \ldots, C\} \to [0, \infty]$, $\ell(q, y) = -\log q_y$, where q_y is the predicted probability of the true class y. The population risk on dimension d is

$$R_d(f) = \mathbb{E}_{(x,y)\sim\mu_d}[\ell(f_d(x), y)]. \tag{2}$$

We report both the average risk over all evaluated dimensions and the risk on dimensions excluded from fitting.

The term "pullback" in this paper refers only to the composition in (1): a task head defined on the reference latent space induces a predictor on every input space. It does not imply that a physically meaningful high-dimensional sample has been generated.

Identifiability and scope. Unpaired marginal distributions do not determine a unique pointwise map between two feature spaces. Our method therefore targets a weaker object: a representation sufficient for the shared label and regularized to preserve within-dimension geometry. The analysis assumes that such a representation exists and varies smoothly with the dimension index. These assumptions are stated in Section 5; they cannot be deduced from unpaired data alone.

4 Dimension-Conditioned Latent Transport

Fix a reference dimension d_r, selected from the training dimensions. The representation map is decomposed as

$$\Phi_d = T_{d\to r} \circ E_d, \tag{3}$$

where E_d is a dimension-conditioned encoder and $T_{d\to r}$ is a latent transport to the reference distribution. The task predictor is therefore

$$f_d = g \circ T_{d\to r} \circ E_d. \tag{4}$$

4.1 Dimension-Conditioned Encoder

Let $D_{\max}$ be the largest supported input dimension, and let $P_d : \mathbb{R}^d \to \mathbb{R}^{D_{\max}}$ denote zero padding. We first apply an orthogonal mixing matrix Q and then a dimension-conditioned network:

$$E_d(x) = E_\theta(QP_d x, e(d)), \tag{5}$$

where $e(d)$ is a sinusoidal encoding of the scalar dimension. The same parameters θ are used for every d.

During encoder pretraining, two perturbed copies of the mixed input are used:

$$h^{(a)} = QP_d x + \epsilon\xi^{(a)}, \qquad \xi^{(a)} \sim \mathcal{N}(0, I), \quad a \in \{1, 2\}. \tag{6}$$

Let $z^{(a)} = E_\theta(h^{(a)}, e(d))$ and let D_d be the associated decoder. The encoder loss is

$$\mathcal{L}_{\mathrm{enc}} = \mathcal{L}_{\mathrm{rec}} + \alpha_{\mathrm{cons}}\mathcal{L}_{\mathrm{cons}} + \alpha_{\mathrm{geo}}\mathcal{L}_{\mathrm{geo}}, \tag{7}$$

with

$$\mathcal{L}_{\mathrm{rec}} = \frac{1}{2d}\sum_{a=1}^{2} \|D_d(z^{(a)}) - x\|^2, \qquad \mathcal{L}_{\mathrm{cons}} = \|z^{(1)} - z^{(2)}\|^2. \tag{8}$$

The geometry term compares pairwise distances in an input minibatch and its latent representation using a GW-type discrepancy. It preserves relative geometry within a dimension; it is not used to create cross-dimensional pairs.

4.2 SOFT-CONSTRAINED LATENT BRIDGE

Let $\nu_d = (E_d)_\sharp \mu_d^x$ be the encoded input distribution, where μ_d^x is the input marginal. Suppose dimensions $d_0, \ldots, d_K, d_{K+1}$ are assigned increasing times $0 = t_0 < \cdots < t_{K+1} = 1$. Let $\mathbb{Q}$ be a reference path measure on $C([0,1], \mathcal{Z})$. The ideal soft-constrained bridge solves

$$\min_{\mathbb{P} \ll \mathbb{Q}} \mathrm{KL}(\mathbb{P} \| \mathbb{Q}) + \sum_{k=0}^{K+1} \lambda_k \mathcal{D}(\mathbb{P}_{t_k}, \nu_{d_k}), \tag{9}$$

Here $\mathbb{P} \ll \mathbb{Q}$ means that $\mathbb{P}$ is absolutely continuous with respect to $\mathbb{Q}$: for every measurable set A in the path space, $\mathbb{Q}(A) = 0$ implies $\mathbb{P}(A) = 0$. The function $\mathcal{D}$ is a nonnegative discrepancy between probability measures; the convergence analysis additionally requires it to vanish only when its two arguments agree. In the implementation, marginal discrepancies are estimated by the debiased Sinkhorn divergence.

We parameterize a conditional vector field $v_\phi(z, t; d_s, d_t)$ and use its ODE flow to define $T_{d_s \to d_t}$. For a minibatch from d_s and a minibatch from d_t, an entropically regularized optimal transport plan provides soft endpoint pairings (z_0, z_1). We sample a conditional path

$$z_t = (1-t)z_0 + t z_1 + \sigma(t)\xi, \quad \xi \sim \mathcal{N}(0, I), \tag{10}$$

and regress the vector field against the corresponding conditional velocity u_t:

$$\mathcal{L}_{\mathrm{FM}} = \mathbb{E} \| v_\phi(z_t, t; d_s, d_t) - u_t(z_t | z_0, z_1) \|^2. \tag{11}$$

For an intermediate dimension d_k, samples transported to time t_k define an empirical marginal $\widehat{\nu}_{t_k}$. The additional penalty is

$$\mathcal{L}_{\mathrm{mid}} = \sum_{k=1}^{K} \lambda_k \mathcal{D}_\varepsilon(\widehat{\nu}_{t_k}, \widehat{\nu}_{d_k}), \tag{12}$$

where $\mathcal{D}_\varepsilon$ is the Sinkhorn divergence with entropic regularization parameter ε. The bridge-training objective is

$$\mathcal{L}_{\mathrm{bridge}} = \mathcal{L}_{\mathrm{FM}} + \mathcal{L}_{\mathrm{mid}}. \tag{13}$$

Equation (11) is a tractable approximation to the transport component of (9). The convergence result in Section 5.3 concerns the ideal measure problem, not finite neural networks or minibatch optimization.

4.3 SHARED TASK HEAD, STABLE TRAINING, AND INFERENCE

The method is trained in three stages. First, the encoder is fitted with reconstruction and geometry-preservation objectives. Second, the latent transport is fitted while the encoder is frozen or updated with a small learning rate. Third, a shared classifier is optimized on transported representations. This order prevents task gradients from destroying the geometry before a common latent coordinate system has formed.

In the third stage, every training sample is mapped to the reference representation $\widetilde{z} = T_{d \to r}(E_d(x))$. The shared classifier g_ω is trained with cross entropy

$$\mathcal{L}_{\mathrm{cls}} = -\mathbb{E} \log [g_\omega(\widetilde{z})]_y. \tag{14}$$

Three task-stage regularizers are used in the stable configuration. A center loss keeps dimension-specific batch means near an exponential moving average $\bar{z}$:

$$\mathcal{L}_{\mathrm{center}} = \sum_{d \in \mathcal{D}_{\mathrm{tr}}} \| \bar{z}_d - \bar{z} \|^2. \tag{15}$$

Here $\bar{z}_d \in \mathbb{R}^p$ is the current dimension-d minibatch mean of transported representations, $\bar{z} \in \mathbb{R}^p$ is the shared exponential-moving-average center, and $\| \cdot \|$ is the Euclidean norm on the reference latent space. For a task-stage minibatch $\{(\widetilde{z}_i, y_i)\}_{i=1}^B$ with $\widetilde{z}_i \in \mathbb{R}^p$, define its mean and sample covariance by

$$m_B = \frac{1}{B} \sum_{i=1}^{B} \widetilde{z}_i, \qquad \Sigma_B = \frac{1}{B-1} \sum_{i=1}^{B} (\widetilde{z}_i - m_B)(\widetilde{z}_i - m_B)^\top.$$

Let $s_j = \sqrt{(\Sigma_B)_{jj} + \eta}$, where $\eta > 0$ is a numerical stability constant, and let $\gamma > 0$ be the target minimum standard deviation. The variance and covariance penalties are

$$\mathcal{L}_{\text{var}} = \frac{1}{p}\sum_{j=1}^{p}\max(0, \gamma - s_j), \qquad \mathcal{L}_{\text{cov}} = \frac{1}{p(p-1)}\sum_{\substack{j,k=1 \\ j\neq k}}^{p}(\Sigma_B)_{jk}^2.$$

Thus $\mathcal{L}_{\text{var}}$ prevents individual coordinates from collapsing, while $\mathcal{L}_{\text{cov}}$ discourages redundant correlations between distinct coordinates.

The supervised contrastive term uses the normalized representations $h_i = \widetilde{z}_i/\|\widetilde{z}_i\|$. For anchor i, let $P(i) = \{r \neq i : y_r = y_i\}$ be its same-label positive set and let $I = \{i : P(i) \neq \varnothing\}$. With temperature $\tau > 0$, the loss is

$$\mathcal{L}_{\text{supcon}} = -\frac{1}{|I|}\sum_{i\in I}\frac{1}{|P(i)|}\sum_{r\in P(i)}\log\frac{\exp(h_i^\top h_r/\tau)}{\sum_{\substack{a=1 \\ a\neq i}}^{B}\exp(h_i^\top h_a/\tau)}.$$

It is activated after a warmup period so that class-based attraction is applied only after the transported representation has stabilized.

The task-stage objective has the form

$$\mathcal{L}_{\text{task}} = \mathcal{L}_{\text{cls}} + \beta_c\mathcal{L}_{\text{center}} + \beta_v\mathcal{L}_{\text{var}} + \beta_q\mathcal{L}_{\text{cov}} + \beta_s\mathcal{L}_{\text{supcon}}. \tag{16}$$

The coefficients are introduced gradually. The experiments show that placing all alignment terms in the encoder and bridge stages is less stable than this staged schedule.

Inference. At inference time, a sample $x \in \mathcal{X}_d$ is encoded and transported to the reference dimension before classification:

$$\widehat{y} = \arg\max_c [g_\omega(T_{d\to r}(E_d(x)))]_c. \tag{17}$$

No retraining of g_ω is required when d changes. For a held-out dimension, the continuous dimension encoding and conditional vector field are evaluated at the new value of d.

5 Theoretical Analysis

The analysis addresses three aspects of the method. First, it decomposes the score approximation error into representation and task components. Second, it quantifies the effect of dimension coverage on the expected risk at a new dimension. Third, it establishes conditions under which the ideal soft bridge approaches the corresponding hard-constrained bridge. These results do not establish identifiability from arbitrary unpaired marginals.

5.1 Composition Error

Let $\Phi_d^* : \mathcal{X}_d \to \mathcal{Z}$ be an ideal representation. For a fixed class coordinate, write $g, g^* : \mathcal{Z} \to \mathbb{R}$ for the learned and ideal scalar score functions, and set $f_d^* = g^* \circ \Phi_d^*$. The result applies componentwise to the multiclass score vector.

Assumption 5.1 (Lipschitz task)**.** The ideal task function g^* is L_g-Lipschitz on $\mathcal{Z}$.

Proposition 5.2 (Pullback composition error)**.** *Define*

$$\varepsilon_g := \sup_{z\in\mathcal{Z}}|g(z) - g^*(z)| < \infty, \qquad \varepsilon_{\Phi,d} := \sup_{x\in\mathcal{X}_d}\|\Phi_d(x) - \Phi_d^*(x)\| < \infty. \tag{18}$$

Then

$$\sup_{x\in\mathcal{X}_d}|g(\Phi_d(x)) - f_d^*(x)| \le \varepsilon_g + L_g\varepsilon_{\Phi,d}. \tag{19}$$

If, in addition, $\Phi_d = T_{d\to r}\circ E_d$ and $\Phi_d^* = T_{d\to r}^*\circ E_d^*$, suppose that $T_{d\to r}^*$ is L_T-Lipschitz on a set containing $E_d(\mathcal{X}_d)\cup E_d^*(\mathcal{X}_d)$ and that

$$\sup_{x\in\mathcal{X}_d}\|E_d(x) - E_d^*(x)\| \le \varepsilon_{E,d}, \qquad \sup_{z\in E_d(\mathcal{X}_d)}\|T_{d\to r}(z) - T_{d\to r}^*(z)\| \le \varepsilon_{T,d}.$$

Then

$$\varepsilon_{\Phi,d} \le \varepsilon_{T,d} + L_T\varepsilon_{E,d}. \tag{20}$$

The proposition is a score-error decomposition, not an estimation or risk guarantee. In particular, a cross-entropy risk bound would require an additional condition keeping the predicted true-class probabilities away from zero, or another regularity assumption on the loss. The result only identifies the score-level quantities controlled by encoder pretraining, bridge fitting, and task learning.

5.2 Coverage of Unseen Dimensions

For this subsection, d denotes a normalized continuous dimension index in $[a, b]$, rather than the integer feature count itself. This idealized model describes interpolation along an ordered family of feature spaces. The learned predictor may depend on the sampled training dimensions, their data, and algorithmic randomness. The assumptions below are required to hold almost surely with deterministic constants.

Assumption 5.3 (Dimension regularity)**.** There is a constant $L_{\dim} \geq 0$ such that, simultaneously for all $d, d' \in [a, b]$, the learned predictor satisfies

$$|R_d(f) - R_{d'}(f)| \leq L_{\dim}|d - d'|. \tag{21}$$

Assumption 5.4 (Per-dimension estimation)**.** Let $R^* \geq 0$ and $A \geq 0$ be fixed constants. If every training dimension provides m samples, the learned predictor satisfies, simultaneously for $i = 1, \ldots, N$,

$$R_{d_i}(f) \leq R^* + \frac{A}{\sqrt{m}}. \tag{22}$$

Theorem 5.5 (Expected dimension-coverage bound)**.** *Let $d_1, \ldots, d_N$ and an independent test dimension d be drawn from a distribution ν on $[a, b]$ whose density is bounded below by $c_\nu > 0$. Under Assumptions 5.3 and 5.4, the expectation over the training and test dimensions, the training data, and any algorithmic randomness satisfies*

$$\mathbb{E}R_d(f) \leq R^* + \frac{A}{\sqrt{m}} + \frac{L_{\dim}}{c_\nu N}. \tag{23}$$

The $A/\sqrt{m}$ term in (23) is inherited from Assumption 5.4; the theorem does not derive this sample-size rate. Its new conclusion is the N^{-1} expected coverage term for a one-dimensional continuous index. This is not a uniform maximum-gap rate, which generally introduces an additional logarithmic factor. For a finite discrete set of feature dimensions, the same argument replaces the last term by $L_{\dim}\mathbb{E}\min_i |d - d_i|$ under a chosen numerical embedding of the discrete indices.

5.3 Convergence of the Ideal Soft Bridge

Assume that $\mathcal{Z}$ is a closed subset of $\mathbb{R}^p$, and let $\Omega = C([0, 1], \mathcal{Z})$ with the uniform topology. Then Ω is a Polish space. Write $\mathcal{P}(\Omega)$ for its Borel probability measures and let $\mathbb{Q} \in \mathcal{P}(\Omega)$ be the reference path measure. Define $\mathcal{P}_{\mathbb{Q}}(\Omega) := \{\mathbb{P} \in \mathcal{P}(\Omega) : \mathbb{P} \ll \mathbb{Q}, \mathrm{KL}(\mathbb{P}\|\mathbb{Q}) < \infty\}$. For $0 = t_0 < \cdots < t_{K+1} = 1$, let $e_{t_k} : \Omega \to \mathcal{Z}$ be the evaluation map and $\mathbb{P}_{t_k} = (e_{t_k})_\sharp\mathbb{P}$. Given target marginals $\mu_k \in \mathcal{P}(\mathcal{Z})$, with $\mu_k = \nu_{d_k}$ in the idealized method, define the hard-constraint set $\mathcal{H} := \{\mathbb{P} \in \mathcal{P}_{\mathbb{Q}}(\Omega) : \mathbb{P}_{t_k} = \mu_k,\ k = 0, \ldots, K + 1\}$. The hard bridge minimizes $\mathrm{KL}(\mathbb{P}\|\mathbb{Q})$ over $\mathbb{P} \in \mathcal{H}$.

Assumption 5.6 (Bridge well-posedness)**.** The set $\mathcal{H}$ is nonempty, and the relative-entropy sublevel sets in $\mathcal{P}_{\mathbb{Q}}(\Omega)$ are weakly compact. The discrepancy $\mathcal{D} : \mathcal{P}(\mathcal{Z}) \times \mathcal{P}(\mathcal{Z}) \to [0, \infty]$ is weakly lower semicontinuous and satisfies $\mathcal{D}(\nu, \mu) = 0$ if and only if $\nu = \mu$. The hard bridge has a unique minimizer $\mathbb{P}^* \in \mathcal{H}$.

Theorem 5.7 (Penalty convergence)**.** *For positive weights $\lambda = (\lambda_0, \ldots, \lambda_{K+1})$, let $\mathbb{P}^\lambda$ be any minimizer of*

$$\min_{\mathbb{P} \in \mathcal{P}_{\mathbb{Q}}(\Omega)} \mathrm{KL}(\mathbb{P}\|\mathbb{Q}) + \sum_{k=0}^{K+1} \lambda_k \mathcal{D}(\mathbb{P}_{t_k}, \mu_k). \tag{24}$$

If $\min_k \lambda_k \to \infty$, then $\mathbb{P}^\lambda \Rightarrow \mathbb{P}^*$ weakly.

Theorem 5.7 justifies the penalty formulation at the level of path measures when $\mathcal{D}$ is a proper discrepancy, such as the Sinkhorn divergence used above. It does not quantify neural approximation, minibatch error, or optimization error.

6 EXPERIMENTS

The experiments provide four complementary evaluations:

1. A shared transported predictor is evaluated on both training and held-out dimensions.
2. Error trends are measured as dimension coverage and sample size increase.
3. The effect of intermediate penalties on marginal matching is assessed.
4. Training schedules are compared for stability on PBMC3k.

Results produced under different checkpoint protocols are reported separately. In particular, Stage-2 selection results, independent benchmark retraining, and the per-dimension oracle comparison are not combined into a single metric tuple.

6.1 SYNTHETIC REGRESSION

Setup. The ambient dimension is $D = 16$. For a fixed pair of vectors $a, b \in \mathbb{R}^{16}$, the target is

$$f^*(x) = \sin(\langle a, x\rangle) + 0.5\cos(\langle b, x\rangle), \tag{25}$$

where the entries of a and b are sampled from $\mathcal{N}(0, 0.25)$. A d-dimensional observation is drawn from a Gaussian mixture on active coordinates and padded to D coordinates when evaluating f^*.

Training dimensions are $\{2, 4, 6, 8, 10, 12\}$ and held-out dimensions are $\{3, 5, 7, 9, 11\}$. Unless otherwise stated, each dimension has 500 samples. The encoder has six coupling layers and latent dimension 16. Because the padded input and latent spaces are both 16-dimensional, the affine-coupling encoder is invertible in this synthetic setting. The conditional vector field has four fully connected layers, and the task head is a three-layer multilayer perceptron. Encoder, bridge, and task stages are trained for 200, 200, and 100 epochs, respectively, with Adam and learning rate 10^{-3}.

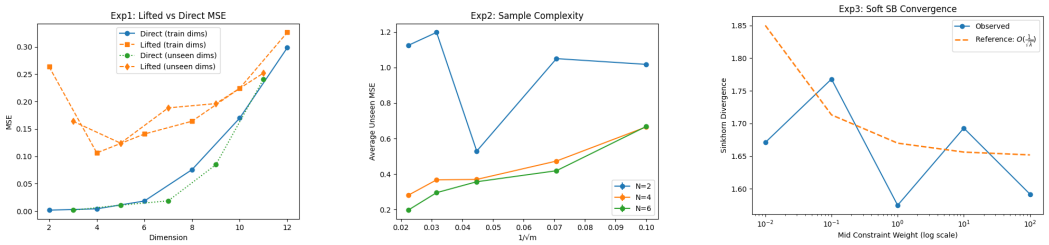
Shared predictor across dimensions. Table 1 compares a dimension-specific regressor with the shared transported predictor. The shared predictor is less accurate, especially at small dimensions, but its error remains bounded on every held-out dimension. The table should therefore be read as evidence of feasibility, not parity with dimension-specific training.

Table 1: Synthetic regression. "Specific" trains one regressor per dimension; "shared" uses one transported predictor.

Training dimensions			Held-out dimensions		
Dimension	Specific MSE	Shared MSE	Dimension	Specific MSE	Shared MSE
2	0.0018	0.2634	3	0.0017	0.1640
4	0.0042	0.1063	5	0.0109	0.1240
6	0.0185	0.1407	7	0.0188	0.1884
8	0.0758	0.1641	9	0.0847	0.1962
10	0.1698	0.2240	11	0.2411	0.2519
12	0.2983	0.3262			

Dimension coverage and sample size. We vary the number of training dimensions over $N \in \{2, 4, 6\}$ and the per-dimension sample size over $m \in \{100, 200, 500, 1000, 2000\}$. Each configuration is repeated three times. Figure 1b plots held-out MSE against $1/\sqrt{m}$. The observed trend decreases with m, and using more training dimensions generally lowers the curve. This behavior is consistent with the two terms in Theorem 5.5, but the experiment is not a proof of the rate.

Intermediate marginal penalty. We transport from dimension 2 to dimension 6 and use dimension 4 as an intermediate marginal at time $1/2$. The penalty λ_{mid} ranges from 0.01 to 100. Figure 1c reports the Sinkhorn divergence between the predicted and empirical intermediate marginals. The divergence is not monotone over the five settings: the largest penalty improves on the smallest one,



(a) Synthetic MSE of dimension-specific and shared transported predictors

(b) Held-out synthetic MSE versus $1/\sqrt{m}$ for different numbers of training dimensions.

(c) Intermediate Sinkhorn divergence across soft-constraint penalty values.

Figure 1: Results of Synthetic Experiments

but an intermediate value gives a slightly smaller value. This scan therefore does not establish numerical convergence. Instead, it shows that finite-network optimization can depart from the ideal limit in Theorem 5.7.

6.2 PBMC3K CLASSIFICATION

Dataset and protocol. We use the PBMC single-cell RNA-sequencing data introduced by Zheng et al. (2017). Different numbers of highly variable genes define the feature dimensions $\mathcal{D} = \{32, 64, 128, 256, 512, 1024\}$. Louvain cluster labels are used as the classification targets. Training dimensions are $\{32, 128, 512\}$; dimensions $\{64, 256, 1024\}$ are held out from dimension-specific fitting. Although all views are derived from the same source cells, cell identities are not supplied to the transport loss. This protocol evaluates learning without using cross-dimensional pairs; it is not equivalent to collecting independent cohorts at each dimension.

Inputs are padded to 1024 coordinates and encoded through six dimension-conditioned coupling blocks followed by a 512-dimensional bottleneck for computational reasons. The PBMC encoder is therefore not bijective, and none of our claims require exact invertibility in this setting. The Stage-2 runs use $12/12/15$ epochs for encoder, bridge, and task stages, with batch size 64 and 800 cells per dimension. The independent benchmark uses the same latent dimension unless otherwise specified. We report mean accuracy, macro-F1, macro-AUROC, cross-dimensional MMD, Wasserstein distance, and the effective rank of the latent covariance.

Staged training ablation. Table 2 compares the proposed staged schedule with the original joint-alignment variants. The jointly stacked alignment losses reduce MMD but substantially harm classification. The proposed schedule fits the encoder and bridge before introducing task-stage regularizers and gives the best accuracy and macro-F1 in this protocol.

Table 2: PBMC3k Stage-2 ablation. All rows use the same data and epoch protocol. Lower MMD and Wasserstein values indicate closer latent marginals.

Configuration	Accuracy	Macro-F1	MMD	Wasserstein
Proposed staged training	**0.826**	**0.697**	0.052	0.011
Shared encoder baseline	0.786	0.643	0.047	0.013
Dimension conditioning	0.781	0.618	0.037	0.011
Center alignment	0.787	0.620	0.035	0.011
Center + weak contrastive	0.743	0.509	**0.027**	**0.009**
Center + weak adversarial	0.794	0.630	0.037	0.011
Joint alignment stack	0.718	0.512	**0.027**	**0.009**

The main observation is that marginal alignment alone is not a sufficient selection criterion. The two rows with the smallest MMD have among the lowest F1 scores. Preserving label information and avoiding collapse are more important than forcing all empirical marginals to coincide.

Independent benchmark retraining. Eight configurations were retrained independently under a common benchmark protocol. The proposed Stage-2 configuration obtained mean accuracy 0.811, macro-F1 0.683, and macro-AUROC 0.947. A covariance regularization ablation obtained the highest accuracy, 0.833, but a lower AUROC of 0.940. Full results are given in Appendix D. We use the independently retrained tuple $0.811/0.683/0.947$ in the abstract because its three metrics come from the same run protocol.

Comparison with dimension-specific classifiers. Table 3 reports a separate checkpoint protocol in which the shared predictor is compared with a classifier trained independently on each dimension. The shared model uses one task head; the dimension-specific models form an oracle-style reference that does not satisfy the train-once requirement.

Table 3: Shared transported predictor and dimension-specific classifiers under the pullback checkpoint protocol.

	Accuracy		Macro-F1		Macro-AUROC	
Dim.	Shared	Specific	Shared	Specific	Shared	Specific
32	0.662	0.765	0.375	0.613	0.846	0.910
64	0.751	0.837	0.488	0.785	0.933	0.967
128	0.864	0.910	0.638	0.824	0.960	0.981
256	0.860	0.922	0.640	0.842	0.952	0.983
512	0.919	0.955	0.707	0.884	0.969	0.992
1024	0.870	0.951	0.625	0.919	0.976	0.997
Mean	0.821	0.890	0.579	0.811	0.939	0.973

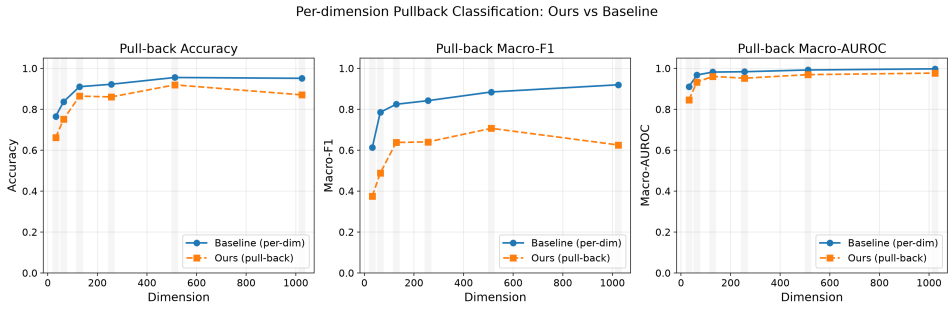


Figure 2: Per-dimension accuracy, macro-F1, and macro-AUROC for the shared predictor and dimension-specific classifiers.

The shared model trails the dimension-specific reference on every aggregate metric. Its AUROC gap is smaller than its macro-F1 gap, which suggests that the learned scores retain ranking information but are less well calibrated for minority-class decisions. The largest accuracy gaps occur at the smallest dimension and at $d = 1024$, outside the interior of the training-dimension range.

7 Conclusion

We presented a method for classification across changing feature dimensions without using cross-dimensional sample identities. A dimension-conditioned encoder preserves within-dimension geometry, a conditional latent flow aligns representations to a common reference, and one task head defines predictors on all dimensions by composition.

The analysis separates representation error, dimension coverage, and soft-constraint convergence. The experiments show that staged training is more reliable than jointly stacking alignment objectives and that a single predictor can operate on held-out dimensions.

AI USE STATEMENT

During the research and preparation of the original manuscript, generative AI tools were used solely to assist with code writing and language editing, including improvements to grammar, clarity, and readability. The research questions, methodological design, theoretical analysis, and scientific conclusions were developed by the authors. The authors have reviewed the AI-assisted code and text and retain full responsibility for the correctness of the implementation and the accuracy and integrity of the analyses, results, and final manuscript.

REPRODUCIBILITY STATEMENT

The model formulation and training objectives are presented in Section 4, and the experimental setups are summarized in Section 6. The assumptions underlying the theoretical results are stated in Section 5, with proofs provided in Appendices A–C. Additional PBMC3k benchmark results are reported in Appendix D. To facilitate reproduction of the synthetic regression and PBMC3k classification experiments, we will release the source code, experiment configurations, and data preparation, training, and evaluation scripts in supplementary materials. The accompanying documentation will specify the software environment, random seeds, preprocessing procedures, data splits, hyperparameters, and checkpoint-selection rules. It will also distinguish the Stage-2 ablation, independent benchmark retraining, and pullback checkpoint protocols, with separate configurations and evaluation instructions for each protocol.

REFERENCES

Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David Duvenaud. Neural ordinary differential equations. In *Advances in Neural Information Processing Systems*, volume 31, pages 6571–6583, 2018.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, volume 26, pages 2292–2300, 2013.

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

Jean Feydy, Thibault Séjourné, François-Xavier Vialard, Shun-ichi Amari, Alain Trouvé, and Gabriel Peyré. Interpolating between optimal transport and MMD using sinkhorn divergences. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89, pages 2681–2690. PMLR, 2019.

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

Facundo Mémoli. Gromov–Wasserstein distances and the metric approach to object matching. *Foundations of Computational Mathematics*, 11:417–487, 2011.

Danilo Jimenez Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37, pages 1530–1538, 2015.

Erwin Schrödinger. Über die Umkehrung der Naturgesetze. *Sitzungsberichte der Preussischen Akademie der Wissenschaften. Physikalisch-Mathematische Klasse*, pages 144–153, 1931.

Cédric Villani. *Topics in Optimal Transportation*. American Mathematical Society, Providence, RI, USA, 2003.

Grace X. Y. Zheng et al. Massively parallel digital transcriptional profiling of single cells. *Nature Communications*, 8:14049, 2017.

A Proof of Proposition 5.2

For any $x \in \mathcal{X}_d$, insert and subtract $g^*(\Phi_d(x))$:

$$|g(\Phi_d(x)) - g^*(\Phi_d^*(x))| \le |g(\Phi_d(x)) - g^*(\Phi_d(x))| \tag{26}$$
$$+ |g^*(\Phi_d(x)) - g^*(\Phi_d^*(x))| \tag{27}$$
$$\le \varepsilon_g + L_g \varepsilon_{\Phi,d}. \tag{28}$$

Taking the supremum gives (19).

For the second statement, write E_d^* and $T_{d\to r}^*$ for the ideal maps and use the two factorizations stated in the proposition. Then, for every $x \in \mathcal{X}_d$,

$$\|T_{d\to r}(E_d(x)) - T_{d\to r}^*(E_d^*(x))\| \tag{29}$$
$$\le \|T_{d\to r}(E_d(x)) - T_{d\to r}^*(E_d(x))\| \tag{30}$$
$$+ \|T_{d\to r}^*(E_d(x)) - T_{d\to r}^*(E_d^*(x))\| \tag{31}$$
$$\le \varepsilon_{T,d} + L_T \varepsilon_{E,d}. \tag{32}$$

Taking the supremum over x and using the definition of $\varepsilon_{\Phi,d}$ proves (20).

B Proof of Theorem 5.5

Let $d_{(1)}$ be the closest training dimension to an independent test dimension d. By Assumptions 5.3 and 5.4,

$$R_d(f) \le R_{d_{(1)}}(f) + L_{\dim}|d - d_{(1)}| \le R^* + \frac{A}{\sqrt{m}} + L_{\dim}|d - d_{(1)}|. \tag{33}$$

For a fixed d and radius $r \le b - a$, the interval $[d - r, d + r] \cap [a, b]$ has length at least r. Since the density of ν is at least c_ν,

$$\Pr(|d - d_{(1)}| > r \mid d) \le (1 - c_\nu r)^N \tag{34}$$
$$\le \exp(-c_\nu N r). \tag{35}$$

For $r > b - a$, the event $|d - d_{(1)}| > r$ is empty. The tail-integral formula therefore gives

$$\begin{aligned} \mathbb{E}[|d - d_{(1)}| \mid d] &= \int_0^{b-a} \Pr(|d - d_{(1)}| > r \mid d)\, \mathrm{d}r \\ &\le \int_0^{\infty} \exp(-c_\nu N r)\, \mathrm{d}r = \frac{1}{c_\nu N}. \end{aligned} \tag{36}$$

Taking expectations over all stated sources of randomness in (33) proves (23).

C Proof of Theorem 5.7

Let J_λ denote the objective in (24), and set $C := \mathrm{KL}(\mathbb{P}^*\|\mathbb{Q}) < \infty$. We first verify that a soft minimizer exists for each positive λ. Since $\mathbb{P}^*$ is an admissible competitor and all penalty terms vanish at $\mathbb{P}^*$, the infimum of J_λ is at most C. A minimizing sequence can therefore be chosen inside the relative-entropy sublevel set $\{\mathbb{P} : \mathrm{KL}(\mathbb{P}\|\mathbb{Q}) \le C + 1\}$. This set is weakly compact by Assumption 5.6. The evaluation maps e_{t_k} are continuous, so weak convergence of path measures implies weak convergence of their time marginals. Relative entropy and each discrepancy term are weakly lower semicontinuous; hence any limit point of the minimizing sequence attains the infimum.

Now fix an arbitrary sequence of positive weight vectors $\{\lambda^{(n)}\}_{n\ge1}$ such that $\min_k \lambda_k^{(n)} \to \infty$, and write $\mathbb{P}^n := \mathbb{P}^{\lambda^{(n)}}$. Optimality and feasibility of $\mathbb{P}^*$ give

$$J_{\lambda^{(n)}}(\mathbb{P}^n) \le J_{\lambda^{(n)}}(\mathbb{P}^*) = C. \tag{37}$$

Because relative entropy and the discrepancy terms are nonnegative, $\mathrm{KL}(\mathbb{P}^n\|\mathbb{Q}) \le C$. Compactness of this sublevel set gives a subsequence, denoted $\mathbb{P}^{n_j}$, and a measure $\overline{\mathbb{P}}$ such that $\mathbb{P}^{n_j} \Rightarrow \overline{\mathbb{P}}$.

For every k, equation (37) also gives

$$\mathcal{D}(\mathbb{P}_{t_k}^{n_j}, \mu_k) \leq \frac{C}{\lambda_k^{(n_j)}} \longrightarrow 0. \tag{38}$$

Continuity of e_{t_k} implies $\mathbb{P}_{t_k}^{n_j} \Rightarrow \overline{\mathbb{P}}_{t_k}$. Consequently, weak lower semicontinuity and (38) yield

$$0 \leq \mathcal{D}(\overline{\mathbb{P}}_{t_k}, \mu_k) \leq \liminf_{j\to\infty} \mathcal{D}(\mathbb{P}_{t_k}^{n_j}, \mu_k) = 0. \tag{39}$$

The identity property of $\mathcal{D}$ gives $\overline{\mathbb{P}}_{t_k} = \mu_k$ for every k, and hence $\overline{\mathbb{P}} \in \mathcal{H}$.

Since $\mathbb{P}^*$ minimizes relative entropy over $\mathcal{H}$, weak lower semicontinuity of relative entropy and (37) give the full optimality chain

$$\begin{aligned} \mathrm{KL}(\mathbb{P}^*\|\mathbb{Q}) &\leq \mathrm{KL}(\overline{\mathbb{P}}\|\mathbb{Q}) \\ &\leq \liminf_{j\to\infty} \mathrm{KL}(\mathbb{P}^{n_j}\|\mathbb{Q}) \\ &\leq \mathrm{KL}(\mathbb{P}^*\|\mathbb{Q}). \end{aligned} \tag{40}$$

Thus $\overline{\mathbb{P}}$ is a hard-constrained minimizer, and uniqueness gives $\overline{\mathbb{P}} = \mathbb{P}^*$. The original sequence is relatively compact, and every one of its cluster points is $\mathbb{P}^*$. Therefore $\mathbb{P}^{\lambda^{(n)}} \Rightarrow \mathbb{P}^*$ for every sequence with $\min_k \lambda_k^{(n)} \to \infty$, which proves the stated convergence.

D Additional PBMC Results

Table 4 contains the complete independent retraining benchmark referenced in the main text. Configuration names describe the recorded training variants. The table is reported in full to avoid selecting only the best metric from each run.

Table 4: Independent PBMC benchmark retraining.

Configuration	Acc.	F1	AUROC	MMD	Wass.	Eff. rank
S2-Stable-Final	0.811	**0.683**	**0.947**	0.118	0.441	10.2
v6-BestAcc (−CovReg)	**0.833**	0.683	0.940	0.123	0.557	13.2
Contrast-λ0.05	0.802	0.668	0.940	0.169	0.396	11.5
Ablation-CCA (legacy)	0.800	0.659	0.934	0.213	0.352	85.3
Combined-Optimal	0.787	0.592	0.929	0.152	0.268	34.1
v6-AlignFocus	0.766	0.553	0.931	0.142	0.526	64.5
v6-Full	0.797	0.535	0.919	0.114	0.403	73.7
v6-BestF1 (−VarReg)	0.805	0.531	0.921	0.180	0.303	39.2

The benchmark illustrates metric trade-offs. The configuration with the highest accuracy is not the one with the highest AUROC, and the lowest MMD does not imply the highest F1. For this reason, the main text reports the complete metric tuple for one fixed configuration.