ENTROPIC RIEMANNIAN NEURAL OPTIMAL TRANSPORT

Anonymous authors
Paper under double-blind review

ABSTRACT

Many machine learning problems involve data supported on curved spaces such as spheres, rotation groups, hyperbolic spaces, and general Riemannian manifolds, where Euclidean geometry can distort distances, averages, and the resulting optimal transport (OT) problem. Existing manifold OT methods have pursued amortized out-of-sample maps, while entropic regularization has made discrete OT more scalable, but these advantages have remained largely disjoint. We propose Entropic Riemannian Neural Optimal Transport (Entropic RNOT), a unified framework that combines intrinsic entropic OT with amortized out-of-sample evaluation on Riemannian manifolds. Our method learns a single target-side Schrödinger potential through a neural pullback parameterization, recovers the induced Gibbs coupling and uses the resulting conditional laws to construct intrinsic transport surrogates. These include barycentric projections on Cartan–Hadamard manifolds and heat-smoothed conditional surrogates on complete stochastically complete manifolds, the latter turning possibly atomic target laws into absolutely continuous ones. For fixed regularization $\varepsilon > 0$, we prove that the proposed hypothesis class recovers the entropic optimal coupling in strong probabilistic metrics. As consequences, barycentric surrogates converge in L^2, while heat-smoothed surrogates are stable at fixed heat time and asymptotically unbiased as the heat time vanishes. The guarantees hold for compactly supported data on possibly noncompact manifolds. Empirically, our method matches or improves over Euclidean, tangent-space, and log-Euclidean baselines on benchmarks over $\mathbb{S}^2$, $\mathrm{SO}(3)$, $\mathrm{SPD}(3)$, $\mathrm{SE}(3)$, and $\mathbb{H}^2$, scales favorably relative to discrete manifold Sinkhorn, and in a protein–ligand docking application, refines poses on $\mathrm{SE}(3)$ without retraining or per-instance optimization.

1 INTRODUCTION

Optimal transport (OT) provides a principled way to compare and transform probability distributions, and has become a powerful tool in modern machine learning. In many applications, however, the data are not naturally Euclidean: directions lie on spheres, orientations on rotation groups, rigid poses on $\mathrm{SE}(3)$, and covariance descriptors on SPD manifolds. In such settings, Euclidean approximations can distort both distances and transport summaries, motivating OT methods defined intrinsically on Riemannian manifolds.

Recent work has begun to extend neural OT to this setting. A central motivation is *amortization*: after training, a learned transport mechanism should generalize to unseen data without requiring a new OT solve at test time. Recent manifold OT models (Micheli et al., 2026) pursue intrinsic neural parameterizations of transport maps with out-of-sample generalization, including constructions based on implicit cost-concave potentials (Rezende & Racanière, 2021). However, these approaches also highlight a key computational difficulty: intrinsic manifold OT may still require iterative inner optimization, since essential quantities are defined through optimization subproblems rather than closed-form evaluation.

A complementary route to tractability is *entropic regularization*, which replaces OT by a smoother problem that is much easier to solve, notably via Sinkhorn iterations (Cuturi, 2013). This greatly improves scalability and numerical stability, but does not by itself provide an amortized model: each new instance still typically requires solving a new regularized OT problem. Amortization and entropic

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.

regularization therefore address distinct challenges, namely reusability at test time and per-instance computational cost.

We combine both ideas in an intrinsic geometric setting. Specifically, we introduce *Entropic Riemannian Neural Optimal Transport (Entropic RNOT)*, an intrinsic neural framework for the *static* entropic OT problem on possibly noncompact Riemannian manifolds that aims to learn a reusable manifold-aware transport model while retaining the smoothing and scalability benefits of entropic regularization. To the best of our knowledge, ours is the first intrinsic neural framework for this problem with amortized out-of-sample evaluation.

Our method is based on the semidual formulation of entropic OT: we learn a target-side Schrödinger potential through a neural pullback parameterization and recover the induced Gibbs coupling. We then study intrinsic transport surrogates extracted from the conditional laws of this coupling. In the Cartan–Hadamard setting, these conditionals define a deterministic transport surrogate by barycentric projection. More generally, on complete stochastically complete manifolds, heat smoothing provides a canonical Riemannian way to turn possibly atomic conditional target laws, such as those arising from finite samples or finite target supports, into absolutely continuous conditional distributions.

For fixed $\varepsilon > 0$, we show that the proposed hypothesis class admits approximating sequences whose induced Gibbs plans recover the entropic optimal coupling. As stable consequences of this plan recovery, barycentric projections recover the associated deterministic surrogate in $L^2(\mu)$ on Cartan–Hadamard manifolds, while heat-smoothed conditional laws are stable at every fixed heat time and have vanishing population-level smoothing bias as the heat time tends to zero.

Empirically, we evaluate our framework on synthetic and real manifold-valued transport tasks. On benchmarks over the compact and noncompact manifolds $\mathbb{S}^2$, $\mathrm{SO}(3)$, $\mathrm{SPD}(3)$, $\mathrm{SE}(3)$, and $\mathbb{H}^2$, our method is consistently competitive with or superior to ambient Euclidean, tangent-space, and log-Euclidean baselines. Scalability experiments show substantially better wall-clock, memory, and inference scaling than discrete manifold Sinkhorn in the large-support regime. On a real-world, rigid-pose refinement task on $\mathrm{SE}(3)$ for protein–ligand docking in CrossDocked2020, a single model trained on pooled complexes refines held-out docking poses toward the near-native region without crystal supervision during training or inference and without per-instance optimization. This reduces top-1 RMSD from 11.24 Å to 3.47 Å, improves success at 2 Å from 10.3% to 75.9%, and substantially outperforms physics-based minimization.

Our contributions are threefold. First, we introduce Entropic RNOT, an intrinsic neural framework for the *static* entropic OT problem on Riemannian manifolds that combines the semidual formulation with amortized out-of-sample evaluation. Second, for fixed $\varepsilon > 0$, we prove recovery of the entropic optimal coupling in strong probabilistic metrics, and derive stable recovery results for intrinsic transport surrogates: barycentric projections converge in $L^2(\mu)$ in the Cartan–Hadamard setting, while heat-smoothed conditional laws are stable at fixed heat time and asymptotically unbiased as the heat time vanishes. Third, we demonstrate empirically strong transport quality across diverse manifold geometries, more favorable scaling than discrete manifold Sinkhorn, and effective real-world $\mathrm{SE}(3)$ post-docking pose refinement.

2 Background

We briefly introduce the notation and constructions used throughout the paper; a more detailed review of entropic optimal transport on Riemannian supports is provided in Appendix A.

Throughout, $(\mathcal{M}, \mathfrak{g})$ denotes a complete p-dimensional Riemannian manifold with geodesic distance d, and $\mathcal{P}(\mathcal{M})$ denotes the set of Borel probability measures on $\mathcal{M}$. For $r \geq 1$, we write $\mathcal{P}_r(\mathcal{M})$ for the space of probability measures on $\mathcal{M}$ with finite rth moment. We fix an entropic regularization parameter $\varepsilon > 0$ and consider the quadratic geodesic cost

$$c(x, y) := \frac{1}{2} d(x, y)^2, \qquad x, y \in \mathcal{M}, \tag{1}$$

Entropic OT and the semidual formulation. For $\mu, \nu \in \mathcal{P}(\mathcal{M})$ with $c \in L^1(\mu \otimes \nu)$, the *entropically regularized optimal transport problem* is

$$\mathrm{OT}_\varepsilon(\mu, \nu) := \inf_{\pi \in \Pi(\mu, \nu)} \left\{ \int_{\mathcal{M} \times \mathcal{M}} c(x, y)\, d\pi(x, y) + \varepsilon\, \mathrm{KL}(\pi \mid \mu \otimes \nu) \right\}. \tag{2}$$

Given a measurable function $g : \mathcal{M} \to \mathbb{R}$, define its *soft c-transform relative to* ν by

$$(\mathcal{T}_\nu^\varepsilon g)(x) := -\varepsilon \log\left(\int_{\mathcal{M}} \exp\left(\frac{g(y) - c(x,y)}{\varepsilon}\right) d\nu(y)\right), \qquad x \in \mathcal{M}, \tag{3}$$

whenever the normalizing integral belongs to $(0, \infty)$. For admissible g, in the sense made precise in Proposition A.2, set

$$\mathcal{J}_\varepsilon(g) := \int_{\mathcal{M}} g \, d\nu + \int_{\mathcal{M}} \mathcal{T}_\nu^\varepsilon g \, d\mu. \tag{4}$$

Under the standing assumption $c \in L^1(\mu \otimes \nu)$, entropic OT admits Schrödinger potentials $(f_\varepsilon^\star, g_\varepsilon^\star)$, unique up to an additive constant, satisfying

$$f_\varepsilon^\star = \mathcal{T}_\nu^\varepsilon g_\varepsilon^\star$$

μ-almost surely, or everywhere on the relevant support after choosing the canonical pointwise representatives described in Proposition A.1. The unique optimizer $\pi_\varepsilon^\star$ has Gibbs density

$$\frac{d\pi_\varepsilon^\star}{d(\mu \otimes \nu)}(x,y) = \exp\left(\frac{f_\varepsilon^\star(x) + g_\varepsilon^\star(y) - c(x,y)}{\varepsilon}\right). \tag{5}$$

Moreover, the problem admits the following one-potential semidual representation, where the supremum is over admissible g:

$$\mathrm{OT}_\varepsilon(\mu, \nu) = \sup_g \mathcal{J}_\varepsilon(g),$$

Transport surrogates extracted from Gibbs conditionals. Let π be a probability measure on $\mathcal{M} \times \mathcal{M}$ with first marginal μ, and write its disintegration as

$$\pi(dx, dy) = \mu(dx)\, \pi_x(dy).$$

In general, the conditional law π_x may be viewed as a distributional transport output associated with the source point x. The two following intrinsic approaches to extract transport surrogates from these conditionals will be relevant.

First, when $(\mathcal{M}, g)$ is Cartan–Hadamard and the conditional laws have finite second moments, the barycentric projection is defined by

$$T_\pi(x) := \mathrm{bar}(\pi_x) := \operatorname*{arg\,min}_{z \in \mathcal{M}} \frac{1}{2} \int_{\mathcal{M}} d(z,y)^2 \, \pi_x(dy).$$

The finite-second-moment condition is automatic in the compact-support setting: if π is supported on $\mathcal{M} \times K_\nu$ for a compact set $K_\nu \subset \mathcal{M}$, then $\pi_x(K_\nu) = 1$ for μ-almost every x, and hence $\pi_x \in \mathcal{P}_2(\mathcal{M})$ for μ-almost every x.

Second, if $(\mathcal{M}, g)$ is complete and stochastically complete, with heat kernel $p_t(y,z)$, the conditional law π_x in the second variable may be smoothed by the heat semigroup. For $t > 0$, define

$$Q_{\pi,t}(x, dz) := \Big(\int_{\mathcal{M}} p_t(y,z)\, \pi_x(dy)\Big) \, \mathrm{vol}_g(dz).$$

Heat smoothing provides a canonical Riemannian way to turn possibly atomic conditional target laws, such as those obtained from finite samples of a learned plan, into absolutely continuous conditional distributions. This is useful when one wants a density-based transport surrogate rather than only a discrete conditional measure. Depending on the application, the smoothed law $Q_{\pi,t}(x, \cdot)$ may be used directly or a point prediction may be derived from it, for instance by selecting a mode of the heat-smoothed density. As shown in Proposition 4.3, the heat-smoothed surrogate is stable at every fixed heat time $t > 0$, and its population-level smoothing bias vanishes as $t \downarrow 0$.

Feature-induced hypothesis classes. We learn functions on $D \subset \mathcal{M}$ by pulling back Euclidean models through a feature map; see Appendix B for the compact-domain transfer principle and intrinsic examples of admissible feature maps. Let D be compact, and let $\varphi : D \to \mathbb{R}^n$ be a continuous feature map, and let $\mathcal{F} \subset C(\mathbb{R}^n, \mathbb{R})$ be a Euclidean function class, for example a neural network class. The induced pullback class is

$$\varphi^* \mathcal{F} := \{a \circ \varphi : \ a \in \mathcal{F}\}. \tag{6}$$

The approximation power of this class depends jointly on $\mathcal{F}$ and φ. Since D is compact, uniform convergence on compact sets reduces on D to ordinary uniform convergence. Hence, if $\mathcal{F}$ is dense in $C(\mathbb{R}^n, \mathbb{R})$ under the ucc topology and φ is continuous and injective, then $\varphi^*\mathcal{F}$ is dense in $C(D, \mathbb{R})$ under the uniform norm. This applies, for instance, to standard universal neural network architectures (Leshno et al., 1993; Lu et al., 2017; Zhou, 2020) and to posterior means of Gaussian processes with universal kernels (Micchelli et al., 2006).

Since Schrödinger potentials are identifiable only up to additive constants, we work with the centered pullback class

$$\mathsf{C}_\nu(\varphi^*\mathcal{F}) := \left\{ a \circ \varphi - \int a \circ \varphi \, d\nu : \; a \in \mathcal{F} \right\}, \tag{7}$$

where $\mathsf{C}_\nu h := h - \int h \, d\nu$ for ν-integrable h.

3 Entropic Riemannian neural optimal transport

We now present *Entropic RNOT*, the learning framework analyzed in our work. The model is obtained by restricting the semidual objective to the centered pullback class introduced in (7).

Parameterized semidual model. Let $\varphi : K_\nu \to \mathbb{R}^n$ be the target-side feature map introduced above. Let $a_\theta \in \mathcal{F}$ be a Euclidean parametric model, and define

$$h_\theta := a_\theta \circ \varphi \in \varphi^*\mathcal{F}, \qquad g_\theta := \mathsf{C}_\nu h_\theta.$$

We use g_θ as the target-side Schrödinger potential and train the model by solving

$$\max_\theta \mathcal{J}_\varepsilon(g_\theta). \tag{8}$$

Once g_θ is given, the corresponding source-side potential is recovered through the soft c-transform,

$$f^\varepsilon_\theta := \mathcal{T}^\varepsilon_\nu g_\theta.$$

Stochastic optimization. In practice, we optimize $\mathcal{J}_\varepsilon(g_\theta)$ using minibatch approximations. Given samples $x_1, \dots, x_B \sim \mu$ and $y_1, \dots, y_B \sim \nu$, we use the empirical centering

$$g_\theta(y_j) = h_\theta(y_j) - \frac{1}{B}\sum_{\ell=1}^{B} h_\theta(y_\ell).$$

The companion source-side potential is approximated by

$$f^\varepsilon_\theta(x_i) = -\varepsilon \log\left(\frac{1}{B}\sum_{j=1}^{B} \exp\left(\frac{g_\theta(y_j) - c(x_i, y_j)}{\varepsilon} \right) \right)$$

and evaluated numerically with a stable log-sum-exp implementation. This yields the following minibatch objective, which we maximize by stochastic gradient ascent:

$$\widehat{\mathcal{J}}_\varepsilon(\theta) = \frac{1}{B}\sum_{j=1}^{B} g_\theta(y_j) + \frac{1}{B}\sum_{i=1}^{B} f^\varepsilon_\theta(x_i).$$

Transport objects induced by entropic RNOT. The learned potential g_θ induces the Gibbs plan

$$d\pi^\varepsilon_\theta(x, y) = \exp\left(\frac{f^\varepsilon_\theta(x) + g_\theta(y) - c(x, y)}{\varepsilon} \right) d\mu(x) \, d\nu(y). \tag{9}$$

Equivalently, for each source point x, the model defines the conditional law

$$\pi^\varepsilon_{\theta,x}(dy) = \exp\left(\frac{f^\varepsilon_\theta(x) + g_\theta(y) - c(x, y)}{\varepsilon} \right) d\nu(y), \qquad \pi^\varepsilon_\theta(dx, dy) = \mu(dx) \, \pi^\varepsilon_{\theta,x}(dy).$$

Thus, the learned model assigns to each source point a regularized conditional output law $\pi^\varepsilon_{\theta,x}$. When $(\mathcal{M}, g)$ is Cartan–Hadamard and $\pi^\varepsilon_{\theta,x} \in \mathcal{P}_2(\mathcal{M})$ for μ-almost every x, we denote the induced barycentric surrogate by $\widehat{T}^\varepsilon_\theta := T_{\pi^\varepsilon_\theta}$. If $(\mathcal{M}, g)$ is complete and stochastically complete, we denote the heat-smoothed surrogate by $Q^\varepsilon_{\theta,t} := Q_{\pi^\varepsilon_\theta, t}$ for $t > 0$.

4 THEORETICAL GUARANTEES

We now show that Entropic RNOT is sufficiently expressive to recover the entropic coupling and the transport surrogates extracted from its Gibbs conditionals. To keep the argument concrete, we work in the compact-support regime, which provides a convenient sufficient setting for dual attainment, uniform approximation of Schrödinger potentials, and stability of the induced Gibbs plans. Proofs from this section are given in Appendix G.1, Appendix G.2, and Appendix G.3.

Accordingly, throughout this section, we assume that $\mu, \nu \in \mathcal{P}(\mathcal{M})$ have compact supports

$$K_\mu := \mathrm{spt}(\mu), \qquad K_\nu := \mathrm{spt}(\nu),$$

that $\mathcal{F} \subset C(\mathbb{R}^n, \mathbb{R})$ is dense under the ucc topology, and that the feature map $\varphi : K_\nu \to \mathbb{R}^n$ is continuous and injective. Since K_ν is compact, $\varphi(K_\nu)$ is compact as well, so ucc-density of $\mathcal{F}$ implies uniform density on $\varphi(K_\nu)$. Consequently, the centered pullback class $\mathsf{C}_\nu(\varphi^*\mathcal{F})$ is dense in the centered subspace of $C(K_\nu)$ under the uniform norm (see Proposition G.3).

Recovering the entropic coupling. Our first result below shows that for each fixed $\varepsilon > 0$, the class $\mathsf{C}_\nu(\varphi^*\mathcal{F})$ is rich enough to recover the entropic optimal coupling through the semidual objective.

Theorem 4.1 (Recovery of the Entropic Coupling)**.** *For fixed $\varepsilon > 0$, there exists a sequence $(g_m)_{m\in\mathbb{N}} \subset \mathsf{C}_\nu(\varphi^*\mathcal{F})$ such that*

$$\mathcal{J}_\varepsilon(g_m) \xrightarrow[m\to\infty]{} \mathrm{OT}_\varepsilon.$$

Let π^ε_m denote the Gibbs plan induced by g_m. Then, as $m \to \infty$,

$$\mathrm{KL}(\pi^\star_\varepsilon \,\|\, \pi^\varepsilon_m) \to 0, \qquad \|\pi^\varepsilon_m - \pi^\star_\varepsilon\|_{\mathrm{TV}} \to 0, \qquad \pi^\varepsilon_m \rightharpoonup \pi^\star_\varepsilon.$$

Recovering and stabilizing transport surrogates. Beyond recovery of the entropic coupling itself, the same approximation mechanism also controls the transport surrogates induced by Entropic RNOT and introduced in Section 2. These surrogate results should be understood as stable consequences of plan recovery. In the Cartan–Hadamard setting, the entropic barycentric projection is recovered in $L^2(\mu)$.

Proposition 4.2 (Recovery of the Entropic Barycentric Surrogate)**.** *Assume in addition that $(\mathcal{M}, g)$ is Cartan–Hadamard, and fix $\varepsilon > 0$. Then there exists a sequence $(g_m)_{m\in\mathbb{N}} \subset \mathsf{C}_\nu(\varphi^*\mathcal{F})$ as in Theorem 4.1. Let π^ε_m denote the Gibbs plan induced by g_m. If*

$$T_\varepsilon := T_{\pi^\star_\varepsilon}, \qquad \widehat{T}^\varepsilon_m := T_{\pi^\varepsilon_m},$$

then

$$\widehat{T}^\varepsilon_m \to T_\varepsilon \qquad \text{in } L^2(\mu).$$

Hence, $\widehat{T}^\varepsilon_m \to T_\varepsilon$ also in $L^1(\mu)$.

On complete stochastically complete manifolds, heat smoothing provides a canonical Riemannian way to turn possibly atomic conditional target laws, such as those obtained from finite samples or finite target supports of a learned plan, into absolutely continuous conditional distributions. This is useful when one wants to work with a density-based transport surrogate rather than only a discrete conditional law. At the level of the induced smoothed laws, the heat-smoothed surrogate is stable at every fixed heat time $t > 0$, and its population-level smoothing bias vanishes as $t \downarrow 0$. For any probability measure π on $\mathcal{M} \times \mathcal{M}$ with first marginal μ, written as $\pi(dx, dy) = \mu(dx)\pi_x(dy)$, we define for the corresponding heat-smoothed joint measure:

$$\Pi_{\pi,t}(dx, dz) := \mu(dx)\, Q_{\pi,t}(x, dz)$$

Proposition 4.3 (Stability of the Heat-Smoothed Surrogate)**.** *Assume in addition that $(\mathcal{M}, g)$ is complete and stochastically complete, and fix $\varepsilon > 0$. Let $(g_m)_{m\in\mathbb{N}} \subset \mathsf{C}_\nu(\varphi^*\mathcal{F})$ be the approximating sequence from Theorem 4.1, with induced Gibbs plans π^ε_m. For $t > 0$, define the heat-smoothed joint measures*

$$\Pi^\varepsilon_{m,t} := \Pi_{\pi^\varepsilon_m, t}, \qquad \Pi^\star_{\varepsilon,t} := \Pi_{\pi^\star_\varepsilon, t}.$$

Then $Q_{\pi^\varepsilon_m,t}(x, \cdot)$ and $Q_{\pi^\star_\varepsilon,t}(x, \cdot)$ are absolutely continuous with respect to vol_g for μ-almost every x, and

$$\|\Pi^\varepsilon_{m,t} - \Pi^\star_{\varepsilon,t}\|_{\mathrm{TV}} \le \|\pi^\varepsilon_m - \pi^\star_\varepsilon\|_{\mathrm{TV}} \xrightarrow[m\to\infty]{} 0.$$

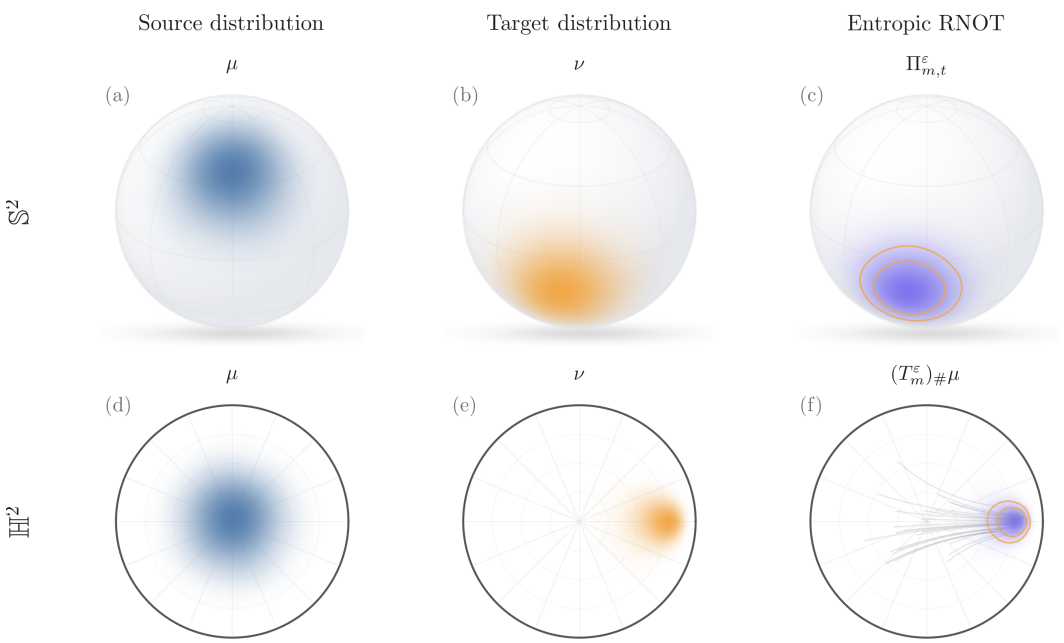


Figure 1: Qualitative transport on $\mathbb{S}^2$ (top) and $\mathbb{H}^2$ (bottom). Each row shows the source distribution μ, the target distribution ν, and the corresponding method output. On $\mathbb{S}^2$, the right panel shows the heat-smoothed output distribution $\hat{\nu}_{\mathrm{heat}}$. On $\mathbb{H}^2$, the right panel shows the pushforward $(\hat{T}_{\mathrm{bar}})_{\#}\mu$ induced by the barycentric transport map. Orange contours indicate the target distribution, and sparse geodesic curves are shown only in the $\mathbb{H}^2$ output panel to visualize representative barycentric map displacements. In both settings, the method output aligns closely with the target.

Consequently, for every fixed $t > 0$,

$$\Pi^{\varepsilon}_{m,t} \rightharpoonup \Pi^{\star}_{\varepsilon,t} \qquad (m \to \infty).$$

Moreover,

$$\Pi^{\star}_{\varepsilon,t} \rightharpoonup \pi^{\star}_{\varepsilon} \qquad (t \downarrow 0),$$

and therefore, for any sequence $t_m \downarrow 0$,

$$\Pi^{\varepsilon}_{m,t_m} \rightharpoonup \pi^{\star}_{\varepsilon}.$$

In particular, Proposition 4.3 applies to complete Riemannian manifolds with Ricci curvature bounded from below; see Appendix A.3 for further details. This includes compact Riemannian manifolds, Euclidean spaces, hyperbolic spaces, and products such as $\mathrm{SE}(3) \simeq \mathrm{SO}(3) \times \mathbb{R}^3$ equipped with the product metric used in our experiments. The same appendix also explains why the criterion applies to the SPD cone endowed with its standard affine-invariant Riemannian metric.

5 Experiments

We evaluate Entropic RNOT in three complementary settings. First, we test transport fidelity on controlled intrinsic-geometry benchmarks, measuring how accurately the learned plan and map recover a discrete manifold Sinkhorn reference. Second, we study computational scaling in time, memory, and inference throughput as the support size increases. Third, we demonstrate amortized out-of-sample inference on a real $\mathrm{SE}(3)$ task: crystal-free refinement of protein–ligand docking poses from CrossDocked2020 (Francoeur et al., 2020). Full implementation details, dataset construction, and supplementary qualitative examples are deferred to Appendix C. The code will be released publicly and is currently **submitted as supplementary material for double-blind review**.

Table 1: Intrinsic-geometry transport benchmarks. Plan KL and endpoint geodesic error are reported relative to a discrete manifold Sinkhorn reference. Lower is better.

Method	$\mathbb{S}^2$	$\mathrm{SO}(3)$	$\mathrm{SPD}(3)$	$\mathrm{SE}(3)$	$\mathbb{H}^2$
Plan KL $\downarrow$					
Ambient Euclidean	0.72 ± 0.040	1.40 ± 0.034	1.86 ± 0.032	1.68 ± 0.036	0.97 ± 0.021
Tangent-space	0.49 ± 0.016	0.15 ± 0.003	1.24 ± 0.018	1.29 ± 0.023	0.10 ± 0.005
Entropic RNOT	$\mathbf{0.04 \pm 0.003}$	$\mathbf{0.03 \pm 0.002}$	$\mathbf{0.01 \pm 0.001}$	$\mathbf{0.05 \pm 0.002}$	$\mathbf{0.01 \pm 0.001}$
Endpoint error $\downarrow$					
Ambient Euclidean	0.21 ± 0.000	0.77 ± 0.008	0.61 ± 0.005	0.64 ± 0.003	0.25 ± 0.001
Tangent-space	0.21 ± 0.001	0.25 ± 0.004	0.43 ± 0.002	0.50 ± 0.002	0.11 ± 0.002
Entropic RNOT	$\mathbf{0.07 \pm 0.002}$	$\mathbf{0.16 \pm 0.005}$	$\mathbf{0.04 \pm 0.003}$	$\mathbf{0.12 \pm 0.001}$	$\mathbf{0.04 \pm 0.003}$

5.1 Intrinsic-geometry benchmarks

We first evaluate Entropic RNOT on controlled synthetic benchmarks across five common geometries: positive-curvature manifolds ($\mathbb{S}^2$, $\mathrm{SO}(3)$), non-positively curved Cartan–Hadamard manifolds ($\mathrm{SPD}(3)$, $\mathbb{H}^2$), and a structured product manifold ($\mathrm{SE}(3)$). Source and target distributions are wrapped normals, with centers chosen to induce realistic geodesic separations in regimes where ambient approximations introduce substantial distortion. The full construction is described in Appendix C.

We compare against two established baselines: ambient Euclidean OT, which ignores the manifold structure entirely, and tangent-space OT, which linearizes the geometry around a single reference point. A discrete manifold Sinkhorn solution, computed directly on the sampled support, serves as the numerical reference. We report plan-level discrepancies from this reference, using KL and conditional W_1, together with map-level errors, using ambient L^2 and endpoint geodesic error.

Table 1 summarizes the results; full metrics are reported in Table 3. Our method consistently recovers the reference plan more accurately than all baselines, with the largest gains on $\mathrm{SPD}(3)$, $\mathrm{SE}(3)$, and $\mathbb{H}^2$, where intrinsic geometry plays the strongest role. Appendix 5 shows that Entropic RNOT also matches non-entropic RNOT in approximation quality on $\mathbb{S}^2$ while being substantially faster and more memory efficient.

5.2 Scalability

A primary motivation for our semidual formulation is computational. Discrete manifold Sinkhorn forms and requires a $N \times N$ geodesic cost matrix, incurring $O(N^2)$ memory and super-linear solve time. Our method replaces this with minibatch optimization of a neural parametric potential and batched inference at test time, both independent of the total support size. We compare wall-clock training time, peak GPU memory, and inference throughput as N grows from 128 to 32,768 across all five geometries; the full description is given in Appendix C.

Figure 2 demonstrates the expected computational benefits. Neural training time and memory remain constant in N ($\approx$24s and $\approx$11MB on $\mathbb{S}^2$), driven only by the batch size. Sinkhorn solve time benefits from GPU parallelism(Cuturi et al., 2022) and grows sub-quadratically in practice, but its $O(N^2)$ memory footprint is the binding constraint, reaching 34 GB at N=32,768 and exceeding the 36 GB device budget on several manifolds. Inference throughput of the learned potential scales linearly with N (up to a few$\times 10^6$ samples/s), reflecting GPU parallelism over a fixed-cost forward pass. These throughput rates make amortized inference practical: a single trained potential can be evaluated on new source samples at negligible marginal cost, without resolving a discrete OT problem for each query. These results show that our method yields a reusable transport model whose cost is decoupled from support size.

5.3 Real-world pose refinement on $\mathrm{SE}(3)$ for protein–ligand docking

We evaluate Entropic RNOT on protein–ligand docking poses derived from CrossDocked2020 (Francoeur et al., 2020), a benchmark of 3,765 complexes across 1,302 pocket-similarity clusters. Our goal is to study crystal-free refinement of docking pose ensembles in a rigid-pose setting. We learn

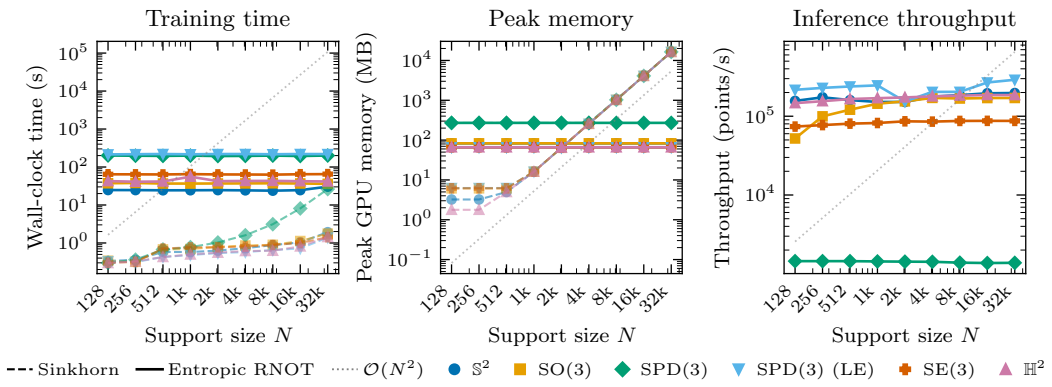


Figure 2: Scalability with increasing support size. Left: training wall-clock time. Center: peak GPU memory. Right: inference throughput. Entropic RNOT scales more favorably than discrete manifold Sinkhorn, especially in the large-support regime.

Table 2: Post-docking pose refinement on $\mathrm{SE}(3)$ (CrossDocked2020, 29 held-out complexes). Per-complex top-1 metrics with 95% bootstrap CIs. Crystal pose used for evaluation only.

Method	RMSD (Å) ↓	Median (Å) ↓	@2Å ↑	@5Å ↑
No refinement	$11.24_{[8.37,14.14]}$	$10.39_{[6.95,12.39]}$	$10.3\%_{[1.6,20.7]}$	$24.1\%_{[10.3,43.2]}$
GNINA –minimize	$11.03_{[8.22,14.89]}$	$8.68_{[4.51,15.27]}$	$0.0\%_{[0.0,0.0]}$	$38.9\%_{[22.2,64.0]}$
Sinkhorn SE(3)	$16.04_{[12.15,20.30]}$	$11.91_{[4.29,26.28]}$	$17.2\%_{[6.9,32.8]}$	$34.5\%_{[13.6,50.1]}$
Entropic RNOT	$\mathbf{3.47}_{[1.26,6.58]}$	$\mathbf{1.42}_{[1.30,1.87]}$	$\mathbf{75.9\%}_{[60.3,89.7]}$	$\mathbf{93.1\%}_{[86.2,100.0]}$

an Entropic RNOT refinement map on $\mathrm{SE}(3)$ that moves geometric outliers toward the docking engine's own top-ranked binding basin, thereby acting as a docking-pose refinement or denoising procedure using only quantities available at inference time. Crystallographic poses are never used during training or inference, and are reserved exclusively for held-out evaluation. This setting should be understood as docking-pose refinement rather than full de novo docking or general flexible pose prediction.

For each complex, we generate a docking-pose ensemble with GNINA (McNutt et al., 2025) and convert poses to rigid-body transforms $g_{k,m} \in \mathrm{SE}(3)$ by alignment to a canonical ligand conformer. Source and target pose sets are defined using only the docking engine's own top-ranked pose, and train/test complexes are split by pocket-similarity cluster. We train Entropic RNOT on pooled poses and extract deterministic refinement maps by heat-smoothed mode finding. Preprocessing choices, thresholds, and hyperparameters are reported in Appendix C.4.

Table 2 reports per-complex top-1 metrics on 29 held-out test complexes, evaluated against the crystal pose. Entropic RNOT reduces mean top-1 RMSD from 11.24 Å to 3.47 Å and improves the 2 Å success rate from 10.3% to 75.9%; the median top-1 RMSD is 1.42 Å, indicating near-native refinement for most complexes. A per-complex discrete Sinkhorn baseline on $\mathrm{SE}(3)$, which solves entropic OT separately using each test complex's own source and target poses, performs poorly because the per-complex target sets are too small to identify a stable transport plan. Our problem is support-limited rather than fundamentally ill-posed, held-out complexes contain on average only ≈ 3.8 target poses, and in the extreme case of a single target pose every feasible coupling must transport all sources to it, independently of ε. No solver preserving such sparse per-complex marginals can recover a pose outside their support, whereas Entropic RNOT's amortised approach learns a single potential from a richer set of 559 target poses pooled across training complexes. This supports the need for amortized cross-complex learning. GNINA's built-in energy minimization (McNutt et al., 2025) yields negligible improvement, with mean RMSD remaining 11.03 Å, suggesting that local scoring-surface refinement does not provide the global pose redistribution captured by manifold-aware optimal transport. Although this benchmark is restricted to near-rigid poses, we observe similar or better performance on the full unfiltered docking ensembles (Table 4).

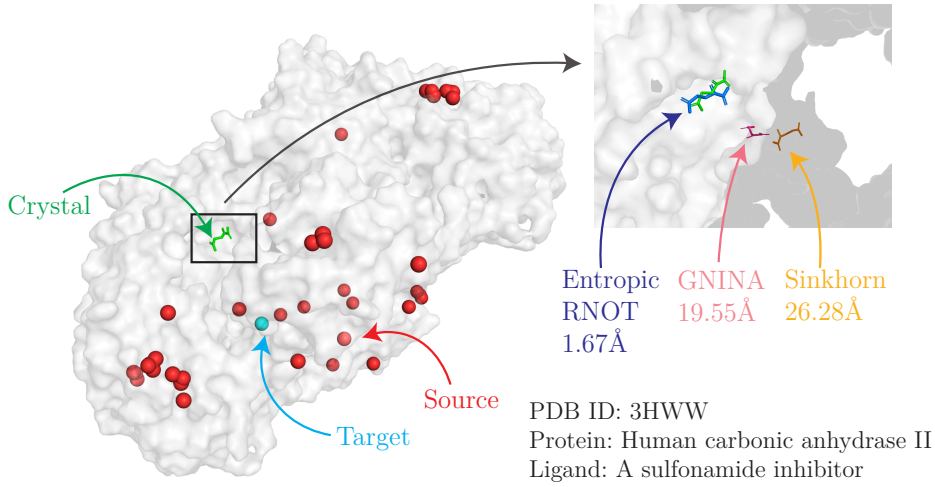


Figure 3: Post-docking pose refinement on complex 3HWW (human carbonic anhydrase II). **(a)** Ligand centroids from 40 GNINA docked poses: source (red, 39 poses) and target (cyan, 1 pose), with crystal ligand (green). **(b)** Pocket zoom: the Entropic RNOT refined pose (blue, 1.67 Å) overlaps the crystal, while GNINA (pink, 19.6 Å) and per-complex Sinkhorn SE(3) (orange, 26.3 Å) land far from the binding site.

6 Discussion and Limitations

We have introduced a manifold-aware neural framework for entropic optimal transport based on a semidual formulation and pullback parameterizations of Schrödinger potentials. Our method learns a target-side potential, recovers the induced Gibbs coupling intrinsically on the manifold, and extracts transport surrogates from its conditional laws: barycentric projections in the Cartan–Hadamard setting and heat-smoothed conditional laws on both geodesically complete and stochastically complete manifolds. For fixed $\varepsilon > 0$, our theory proves the ability to recover the entropic coupling, as well as L^2-recovery of barycentric surrogates when barycenters are uniquely defined, and stability with vanishing population-level bias for heat-smoothed surrogates. Empirically, the framework performs well across synthetic manifold benchmarks, scales better than discrete manifold Sinkhorn, and is shown to be effective in practice for real $\mathrm{SE}(3)$ protein–ligand pose refinement.

At positive temperature, the canonical transport object is a *coupling*, not a deterministic map. Barycentric projections and heat-smoothed modes are only point-valued summaries of the learned conditional law. In the docking experiment, this means that our method is a refinement or denoising procedure for an existing ensemble of docked poses, not an end-to-end docking or generative pose-prediction model. Specifically, it requires candidate poses from an external docking tool and refines them toward a target basin defined by docking-generated geometry and scores. Notably, this target construction does not require crystallographic supervision.

Several limitations remain. The theory treats fixed $\varepsilon > 0$, compact supports, and approximation assumptions on the feature map and Euclidean hypothesis class; it does not address the vanishing-regularization regime. Barycentric map guarantees require the Cartan–Hadamard setting, while outside it barycenters may be nonunique or unstable. Computationally, performance depends on efficient geodesic-distance evaluation, stable log-sum-exp computations, and expressive features. In docking, the current unconditional model ignores receptor pocket context and the sequence and represents poses only as rigid motions on $\mathrm{SE}(3)$, omitting torsional flexibility. Pocket-conditioned transport, product-manifold pose representations with torsions, and broader evaluation across docking benchmarks are natural directions for future work.

REFERENCES

Dimitri P Bertsekas and Steven E Shreve. *Stochastic Optimal Control: The Discrete-Time Case*, volume 139 of *Mathematics in Science and Engineering*. Academic Press, San Diego, CA, 1978.

Gabriele Corso, Hannes Stärk, Bowen Jing, Regina Barzilay, and Tommi Jaakkola. DiffDock: Diffusion steps, twists, and turns for molecular docking. *arXiv [q-bio.BM]*, October 2022.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. https://proceedings.neurips.cc/paper_files/paper/2013/file/af21d0c97db2e27e13572cbf59eb343d-Paper.pdf, 2013. Accessed: 2026-5-5.

Marco Cuturi and Gabriel Peyré. Semidual regularized optimal transport. *SIAM Rev. Soc. Ind. Appl. Math.*, 60(4):941–965, January 2018.

Marco Cuturi, Laetitia Meng-Papaxanthos, Yingtao Tian, Charlotte Bunne, Geoff Davis, and Olivier Teboul. Optimal transport tools (OTT): A JAX toolbox for all things wasserstein. *arXiv [cs.LG]*, January 2022.

Paul Francoeur, Tomohide Masuda, and David Koes. 3D convolutional neural networks and a CrossDocked dataset for structure-based drug design. *ChemRxiv*, March 2020.

Alexander Grigor'yan. Analytic and geometric background of recurrence and non-explosion of the brownian motion on riemannian manifolds. *Bull. New Ser. Am. Math. Soc.*, 36(2):135–249, February 1999.

Alexander Grigoryan. *Heat kernel and analysis on manifolds*. AMS/IP Studies in Advanced Mathematics. American Mathematical Society, Providence, RI, January 2013.

Mikhael Gromov. Filling Riemannian manifolds. *J. Differential Geom.*, 18(1):1–147, January 1983.

Olav Kallenberg. *Foundations of modern probability*. Probability and Its Applications. Springer, New York, NY, 2 edition, January 2002.

Anastasis Kratsios and Eugene Bilokopytov. Non-euclidean universal approximation. *arXiv [cs.LG]*, pp. 10635–10646, June 2020.

John Lee. *Introduction to Smooth Manifolds*. Graduate Texts in Mathematics. Springer, New York, NY, 2 edition, May 2012.

John M Lee. *Introduction to Riemannian manifolds*. Graduate Texts in Mathematics. Springer International Publishing, Cham, Switzerland, 2 edition, January 2019.

Moshe Leshno, Vladimir Ya Lin, Allan Pinkus, and Shimon Schocken. Multilayer feedforward networks with a nonpolynomial activation function can approximate any function. *Neural Netw.*, 6 (6):861–867, January 1993.

Zhou Lu, Hongming Pu, Feicheng Wang, Zhiqiang Hu, and Liwei Wang. The expressive power of neural networks: A view from the width. *arXiv [cs.LG]*, September 2017.

Andrew T McNutt, Yanjing Li, Rocco Meli, Rishal Aggarwal, and David Ryan Koes. GNINA 1.3: the next increment in molecular docking with deep learning. *J. Cheminform.*, 17(1):28, March 2025.

C Micchelli, Yuesheng Xu, and Haizhang Zhang. Universal kernels. *J. Mach. Learn. Res.*, 7: 2651–2667, December 2006.

Alessandro Micheli, Yueqi Cao, Anthea Monod, and Samir Bhatt. Riemannian neural optimal transport. *arXiv [cs.LG]*, February 2026.

Marcel Nutz. Introduction to entropic optimal transport. 2021.

Gabriel Peyré and Marco Cuturi. Computational optimal transport. *arXiv [stat.ML]*, March 2018.

Yury Polyanskiy and Yihong Wu. *Information theory: From coding to learning*. Cambridge University Press, Cambridge, England, January 2025.

Rodrigo Quiroga and Marcos A Villarreal. Vinardo: A scoring function based on autodock vina improves scoring, docking, and virtual screening. *PLoS One*, 11(5):e0155183, May 2016.

Danilo J Rezende and Sébastien Racanière. Implicit Riemannian Concave Potential Maps. *arXiv [stat.ML]*, October 2021.

Karl-Theodor Sturm. Probability measures on metric spaces of nonpositive curvature, 2003.

Yann Thanwerdas and Xavier Pennec. O(n)-invariant riemannian metrics on SPD matrices. *Linear Algebra Appl.*, 661:163–201, March 2023.

Oleg Trott and Arthur J Olson. AutoDock vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J. Comput. Chem.*, 31(2): 455–461, January 2010.

Cédric Villani. *Optimal Transport: Old and New*, volume 338 of *Grundlehren der mathematischen Wissenschaften*. Springer, Berlin, 2009. ISBN 978-3-540-71049-3. doi: 10.1007/978-3-540-71050-9.

Ding-Xuan Zhou. Universality of deep convolutional neural networks. *Appl. Comput. Harmon. Anal.*, 48(2):787–794, March 2020.

A Review on entropic Riemannian optimal transport

Throughout, $(\mathcal{M}, \mathfrak{g})$ denotes a complete, possibly noncompact, p-dimensional Riemannian manifold with geodesic distance d. We write $\mathcal{P}(\mathcal{M})$ for the Borel probability measures on $\mathcal{M}$ and $\mathrm{vol}_{\mathcal{M}}$ for the Riemannian volume measure. Let $\mu, \nu \in \mathcal{P}(\mathcal{M})$ be Borel probability measures and define the quadratic geodesic cost

$$c(x, y) := \frac{1}{2} d(x, y)^2, \qquad x, y \in \mathcal{M}.$$

For the entropic duality results recalled below, the standing assumption is

$$c \in L^1(\mu \otimes \nu).$$

This is the integrability condition used in the general theory of entropic optimal transport; see (Nutz, 2021, Theorem 3.2). In the compact-support setting, this condition is automatic. Indeed, if

$$K_\mu := \mathrm{spt}(\mu), \qquad K_\nu := \mathrm{spt}(\nu)$$

are compact, then $K_\mu \times K_\nu$ is compact and c is continuous, hence bounded, on $K_\mu \times K_\nu$.

Whenever compact supports are assumed, every coupling $\pi \in \Pi(\mu, \nu)$ is supported on $K_\mu \times K_\nu$, so the transport problem may be localized to this compact product set. Since the geodesic distance d is continuous on $\mathcal{M} \times \mathcal{M}$, the quadratic cost $c(x, y) = \frac{1}{2} d(x, y)^2$ is continuous. Hence, by the Heine–Cantor theorem, c is bounded and uniformly continuous on $K_\mu \times K_\nu$. Boundedness gives $c \in L^1(\mu \otimes \nu)$, while uniform continuity is the regularity assumption used below to obtain bounded uniformly continuous Schrödinger potentials.

A.1 Entropic primal, dual, and Schrödinger potentials

Fix $\varepsilon > 0$. The entropically regularized transport problem is

$$\mathrm{OT}_\varepsilon := \inf_{\pi \in \Pi(\mu,\nu)} \left\{ \int_{\mathcal{M}\times\mathcal{M}} c(x, y)\, d\pi(x, y) + \varepsilon\, \mathrm{KL}(\pi \mid \mu \otimes \nu) \right\}. \tag{10}$$

If μ and ν have compact supports K_μ and K_ν, then this is equivalently

$$\mathrm{OT}_\varepsilon = \inf_{\pi \in \Pi(\mu,\nu)} \left\{ \int_{K_\mu\times K_\nu} c(x, y)\, d\pi(x, y) + \varepsilon\, \mathrm{KL}(\pi \mid \mu \otimes \nu) \right\}, \tag{11}$$

because every coupling is concentrated on $K_\mu \times K_\nu$.

For measurable $g : \mathcal{M} \to \mathbb{R}$, define the soft c-transform

$$(\mathcal{T}^\varepsilon_\nu g)(x) := -\varepsilon \log \left(\int_{\mathcal{M}} \exp\left(\frac{g(y) - c(x, y)}{\varepsilon} \right) d\nu(y) \right), \qquad x \in \mathcal{M}. \tag{12}$$

Similarly, for measurable $f : \mathcal{M} \to \mathbb{R}$, define

$$(\mathcal{T}^\varepsilon_\mu f)(y) := -\varepsilon \log \left(\int_{\mathcal{M}} \exp\left(\frac{f(x) - c(x, y)}{\varepsilon} \right) d\mu(x) \right), \qquad y \in \mathcal{M}. \tag{13}$$

These transforms characterize the optimal Schrödinger potentials through the Schrödinger system

$$f^\star_\varepsilon = \mathcal{T}^\varepsilon_\nu g^\star_\varepsilon, \qquad g^\star_\varepsilon = \mathcal{T}^\varepsilon_\mu f^\star_\varepsilon,$$

stated precisely in Proposition A.1; see also (Nutz, 2021, Section 4.1). When μ and ν are compactly supported, only the restrictions of these functions to K_μ and K_ν are relevant. In that case, the transforms can be written equivalently as integrals over K_ν and K_μ, respectively.

We use the standard pointwise representatives of the Schrödinger potentials. The dual maximizers are initially defined only up to μ- and ν-almost-sure equivalence. However, the Schrödinger system provides canonical pointwise versions: each potential can be redefined by the right-hand side of its soft c-transform identity, after which the system holds everywhere on the relevant supports; see (Nutz, 2021, Section 4.1). This pointwise choice is important below because our approximation arguments take place in $C(K_\nu)$ with the uniform norm, rather than only in an almost-sure equivalence class.

In the compact-support setting, these pointwise representatives are regular. Since c is bounded and uniformly continuous on $K_\mu \times K_\nu$, the regularity estimates for entropic OT potentials imply that the ε-scaled Schrödinger potentials may be chosen bounded and uniformly continuous on K_μ and K_ν; see (Nutz, 2021, Lemmas 4.9 and 4.11, Remark 4.12). In particular, the normalized target-side potential used below may be chosen in $C(K_\nu)$.

The next statement recalls the standard entropic OT duality, the associated Schrödinger system, and the Gibbs form of the optimal coupling.

Proposition A.1 (Entropic duality, Schrödinger system, and Gibbs form)**.** *Assume that $c \in L^1(\mu\otimes\nu)$. Then for every $\varepsilon > 0$, the problem* (10) *admits a unique minimizer $\pi^\star_\varepsilon \in \Pi(\mu,\nu)$. Moreover, there exist measurable functions*

$$f^\star_\varepsilon : \mathcal{M} \to \mathbb{R}, \qquad g^\star_\varepsilon : \mathcal{M} \to \mathbb{R},$$

unique up to the transformation

$$(f^\star_\varepsilon, g^\star_\varepsilon) \mapsto (f^\star_\varepsilon + a, g^\star_\varepsilon - a), \qquad a \in \mathbb{R},$$

such that

$$f^\star_\varepsilon = \mathcal{T}^\varepsilon_\nu g^\star_\varepsilon \quad \mu\text{-a.s.}, \qquad g^\star_\varepsilon = \mathcal{T}^\varepsilon_\mu f^\star_\varepsilon \quad \nu\text{-a.s.}, \tag{14}$$

and

$$\frac{d\pi^\star_\varepsilon}{d(\mu\otimes\nu)}(x,y) = \exp\left(\frac{f^\star_\varepsilon(x) + g^\star_\varepsilon(y) - c(x,y)}{\varepsilon}\right) \qquad (\mu\otimes\nu)\text{-a.s.} \tag{15}$$

Equivalently,

$$\mathrm{OT}_\varepsilon = \sup_{f,g}\left\{\int f\,d\mu + \int g\,d\nu - \varepsilon\iint \exp\left(\frac{f(x)+g(y)-c(x,y)}{\varepsilon}\right)\,d\mu(x)d\nu(y) + \varepsilon\right\}, \tag{16}$$

where the supremum is taken over admissible measurable pairs and is attained at $(f^\star_\varepsilon, g^\star_\varepsilon)$.

If, in addition, μ and ν have compact supports K_μ and K_ν, then the optimal potentials admit bounded uniformly continuous representatives on K_μ and K_ν. In particular, after fixing the additive constant by a normalization such as

$$\int_{K_\nu} g^\star_\varepsilon\,d\nu = 0,$$

the normalized target-side potential may be chosen in $C(K_\nu)$.

Proof. Define the static Schrödinger reference measure

$$R_\varepsilon(dx,dy) := Z_\varepsilon^{-1}\exp\left(-\frac{c(x,y)}{\varepsilon}\right)\mu(dx)\nu(dy), \qquad Z_\varepsilon := \iint \exp\left(-\frac{c(x,y)}{\varepsilon}\right)\,d\mu(x)d\nu(y).$$

Then, for every $\pi \in \Pi(\mu,\nu)$,

$$\int c\,d\pi + \varepsilon\,\mathrm{KL}(\pi \mid \mu\otimes\nu) = \varepsilon\,\mathrm{KL}(\pi \mid R_\varepsilon) - \varepsilon\log Z_\varepsilon,$$

with both sides interpreted as $+\infty$ when $\pi \not\ll \mu\otimes\nu$. Thus (10) is equivalent, up to an additive constant, to the static Schrödinger problem with reference measure R_ε.

The existence, uniqueness, dual attainment, Schrödinger system, and Gibbs factorization then follow from the entropic optimal transport duality theorem (Nutz, 2021, Theorem 4.7), applied to the cost c/ε or, equivalently, to the reference measure R_ε. In the compact-support case, c is bounded and uniformly continuous on $K_\mu \times K_\nu$. The boundedness and uniform-continuity estimates for the ε-scaled potentials follow from (Nutz, 2021, Lemmas 4.9 and 4.11, Remark 4.12). □

A.2 One-potential semidual formulation

Define, for measurable $g : \mathcal{M} \to \mathbb{R}$,

$$\mathcal{J}_\varepsilon(g) := \int_{\mathcal{M}} g\,d\nu + \int_{\mathcal{M}} \mathcal{T}^\varepsilon_\nu g\,d\mu. \tag{17}$$

In the compact-support setting, this reduces to

$$\mathcal{J}_\varepsilon(g) = \int_{K_\nu} g \, d\nu + \int_{K_\mu} \mathcal{T}_\nu^\varepsilon g \, d\mu.$$

The regularized semidual formulation is standard in entropic optimal transport; see (Cuturi & Peyré, 2018, Section 2.2) and (Peyré & Cuturi, 2018, Section 5.3). We now include a short derivation to fix notation and to make explicit the normalization used in (16).

Proposition A.2 (Semidual formulation)**.** *Assume that $c \in L^1(\mu \otimes \nu)$. For every measurable $g : \mathcal{M} \to \mathbb{R}$ such that*

$$g \in L^1(\nu), \qquad 0 < Z_g(x) < \infty \quad \textit{for } \mu\textit{-a.e. } x, \qquad \mathcal{T}_\nu^\varepsilon g \in L^1(\mu),$$

where

$$Z_g(x) := \int_{\mathcal{M}} \exp\left(\frac{g(y) - c(x,y)}{\varepsilon}\right) d\nu(y),$$

one has

$$\mathcal{J}_\varepsilon(g) \le \mathrm{OT}_\varepsilon.$$

Moreover,

$$\mathrm{OT}_\varepsilon = \sup_g \mathcal{J}_\varepsilon(g),$$

where the supremum is taken over such admissible g. If $(f_\varepsilon^\star, g_\varepsilon^\star)$ is a maximizing dual pair from Proposition A.1, then

$$f_\varepsilon^\star = \mathcal{T}_\nu^\varepsilon g_\varepsilon^\star \qquad \mu\textit{-a.s.}$$

For the canonical pointwise representatives on compact supports, this identity holds everywhere on K_μ.

Proof. Let $g : \mathcal{M} \to \mathbb{R}$ be admissible. Define

$$Z_g(x) := \int_{\mathcal{M}} \exp\left(\frac{g(y) - c(x,y)}{\varepsilon}\right) d\nu(y).$$

By admissibility,

$$0 < Z_g(x) < \infty \qquad \mu\text{-a.e.}$$

and by definition of the soft c-transform,

$$(\mathcal{T}_\nu^\varepsilon g)(x) = -\varepsilon \log Z_g(x) \qquad \mu\text{-a.e.}$$

We first identify the best choice of f for this fixed g. For a measurable $f : \mathcal{M} \to \mathbb{R}$, define the two-potential dual objective from (16) by

$$D_\varepsilon(f,g) := \int_{\mathcal{M}} f \, d\mu + \int_{\mathcal{M}} g \, d\nu - \varepsilon \iint \exp\left(\frac{f(x) + g(y) - c(x,y)}{\varepsilon}\right) d\mu(x) d\nu(y) + \varepsilon.$$

For fixed g, this can be written as

$$D_\varepsilon(f,g) = \int_{\mathcal{M}} g \, d\nu + \int_{\mathcal{M}} \Psi_x(f(x)) \, d\mu(x),$$

where, for each fixed x,

$$\Psi_x(a) := a - \varepsilon \int_{\mathcal{M}} \exp\left(\frac{a + g(y) - c(x,y)}{\varepsilon}\right) d\nu(y) + \varepsilon.$$

Using the definition of $Z_g(x)$, we have

$$\int_{\mathcal{M}} \exp\left(\frac{a + g(y) - c(x,y)}{\varepsilon}\right) d\nu(y) = e^{a/\varepsilon} Z_g(x).$$

Therefore

$$\Psi_x(a) = a - \varepsilon e^{a/\varepsilon} Z_g(x) + \varepsilon.$$

For μ-a.e. x, this is a strictly concave function of a, because

$$\Psi_x''(a) = -\frac{1}{\varepsilon} e^{a/\varepsilon} Z_g(x) < 0.$$

Its derivative is

$$\Psi_x'(a) = 1 - e^{a/\varepsilon} Z_g(x).$$

Hence the unique maximizer is characterized by

$$e^{a/\varepsilon} Z_g(x) = 1.$$

Equivalently, the unique maximizer is

$$a^\star := -\varepsilon \log Z_g(x) = (\mathcal{T}_\nu^\varepsilon g)(x).$$

Thus, for μ-a.e. x,

$$\Psi_x(a) \le \Psi_x(a^\star) \qquad \text{for all } a \in \mathbb{R}.$$

It remains to compute the value of Ψ_x at this maximizer. Since

$$e^{a^\star/\varepsilon} Z_g(x) = 1,$$

we have

$$\int_{\mathcal{M}} \exp\left(\frac{a^\star + g(y) - c(x,y)}{\varepsilon}\right) d\nu(y) = 1.$$

Therefore

$$\begin{aligned}
\Psi_x(a^\star) &= a^\star - \varepsilon \int_{\mathcal{M}} \exp\left(\frac{a^\star + g(y) - c(x,y)}{\varepsilon}\right) d\nu(y) + \varepsilon \\
&= a^\star - \varepsilon + \varepsilon \\
&= a^\star \\
&= (\mathcal{T}_\nu^\varepsilon g)(x).
\end{aligned}$$

Consequently, for every measurable f for which the dual objective is well defined,

$$D_\varepsilon(f,g) \le \int_{\mathcal{M}} g \, d\nu + \int_{\mathcal{M}} \mathcal{T}_\nu^\varepsilon g \, d\mu = \mathcal{J}_\varepsilon(g),$$

and equality is attained by choosing

$$f = \mathcal{T}_\nu^\varepsilon g.$$

Since $D_\varepsilon(f,g) \le \mathrm{OT}_\varepsilon$ for every admissible dual pair (f,g), we may choose $f = \mathcal{T}_\nu^\varepsilon g$ and obtain

$$\mathcal{J}_\varepsilon(g) = D_\varepsilon(\mathcal{T}_\nu^\varepsilon g, g) \le \mathrm{OT}_\varepsilon.$$

Taking the supremum over admissible g gives

$$\sup_g \mathcal{J}_\varepsilon(g) \le \mathrm{OT}_\varepsilon.$$

Conversely, let $(f_\varepsilon^\star, g_\varepsilon^\star)$ be a maximizing dual pair from Proposition A.1. By the Schrödinger system (14),

$$f_\varepsilon^\star = \mathcal{T}_\nu^\varepsilon g_\varepsilon^\star \qquad \mu\text{-a.s.}$$

Moreover, this identity implies

$$\int_{\mathcal{M}} \exp\left(\frac{f_\varepsilon^\star(x) + g_\varepsilon^\star(y) - c(x,y)}{\varepsilon}\right) d\nu(y) = 1 \qquad \mu\text{-a.e. } x.$$

Substituting the optimal pair into the dual formula (16), the exponential term becomes

$$-\varepsilon \iint \exp\left(\frac{f_\varepsilon^\star(x) + g_\varepsilon^\star(y) - c(x,y)}{\varepsilon}\right) d\nu(y) d\mu(x).$$

By the preceding normalization, the inner integral equals 1 for μ-a.e. x. Therefore

$$-\varepsilon \iint \exp\left(\frac{f_\varepsilon^\star(x) + g_\varepsilon^\star(y) - c(x,y)}{\varepsilon}\right) d\nu(y) d\mu(x) = -\varepsilon \int_{\mathcal{M}} 1 \, d\mu(x) = -\varepsilon,$$

because μ is a probability measure. This cancels with the $+\varepsilon$ in the dual normalization. Hence

$$\begin{aligned}\mathrm{OT}_\varepsilon &= D_\varepsilon(f_\varepsilon^\star, g_\varepsilon^\star)\\ &= \int_{\mathcal{M}} f_\varepsilon^\star \, d\mu + \int_{\mathcal{M}} g_\varepsilon^\star \, d\nu\\ &= \int_{\mathcal{M}} \mathcal{T}_\nu^\varepsilon g_\varepsilon^\star \, d\mu + \int_{\mathcal{M}} g_\varepsilon^\star \, d\nu\\ &= \mathcal{J}_\varepsilon(g_\varepsilon^\star).\end{aligned}$$

Thus

$$\mathrm{OT}_\varepsilon \le \sup_g \mathcal{J}_\varepsilon(g).$$

Together with the opposite inequality proved above, this gives

$$\mathrm{OT}_\varepsilon = \sup_g \mathcal{J}_\varepsilon(g).$$

Finally,

$$f_\varepsilon^\star = \mathcal{T}_\nu^\varepsilon g_\varepsilon^\star \qquad \mu\text{-a.s.}$$

by the Schrödinger system. For the canonical pointwise representatives on compact supports, this identity holds everywhere on K_μ. □

A.3 Heat kernels and stochastic completeness

We recall the heat-kernel facts used in Proposition 4.3. Grigor'yan treats the more general case of a weighted manifold (M, g, μ). In our setting we take $\mu = \mathrm{vol}_{\mathcal{M}}$, the Riemannian volume measure.

Heat semigroup and heat kernel. Let $(P_t)_{t>0}$ denote the minimal heat semigroup associated with the Laplace–Beltrami operator on $(\mathcal{M}, g)$. We use the heat kernel $p_t(x, y)$ in the sense of (Grigoryan, 2013, Definition 7.12). The properties needed below are those collected in (Grigoryan, 2013, Theorem 7.13). Namely, for every $f \in L^2(\mathcal{M})$, every $x \in \mathcal{M}$, and every $t > 0$,

$$P_t f(x) = \int_{\mathcal{M}} p_t(x, y) f(y) \, \mathrm{vol}_{\mathcal{M}}(dy).$$

Moreover,

$$p_t(x, y) = p_t(y, x), \qquad p_t(x, y) \ge 0,$$

and

$$\int_{\mathcal{M}} p_t(x, y) \, \mathrm{vol}_{\mathcal{M}}(dy) \le 1.$$

Thus the minimal heat kernel is sub-Markovian in general. The same theorem also gives the semigroup identity

$$p_{t+s}(x, y) = \int_{\mathcal{M}} p_t(x, z) p_s(z, y) \, \mathrm{vol}_{\mathcal{M}}(dz), \qquad s, t > 0,$$

and the heat-equation regularity property: for fixed $y \in \mathcal{M}$, the function

$$u(t, x) := p_t(x, y)$$

is smooth on $(0, \infty) \times \mathcal{M}$ and satisfies the heat equation.

Finally, (Grigoryan, 2013, Theorem 7.13) states the small-time initial-condition property for compactly supported smooth functions: for every $f \in C_c^\infty(\mathcal{M})$,

$$\int_{\mathcal{M}} p_t(x, y) f(y) \, \mathrm{vol}_{\mathcal{M}}(dy) \xrightarrow[t \downarrow 0]{} f(x),$$

with convergence in $C^\infty(\mathcal{M})$.

Markovian inequalities. We also use the Markovian properties of the heat semigroup. In the notation of (Grigoryan, 2013, Ch. 5, Section 5.3), the heat semigroup satisfies the following inequalities: if $f \geq 0$, then

$$P_t f \geq 0,$$

and if $f \leq 1$, then

$$P_t f \leq 1.$$

Consequently, if $0 \leq f \leq 1$, then

$$0 \leq P_t f \leq 1.$$

This is the form used in the proof of Proposition 4.3 to show that the heat-kernel averaging of an indicator function remains bounded between 0 and 1.

Stochastic completeness and conservative Markov kernels. Following (Grigoryan, 2013, Definition 8.17), a weighted manifold is called stochastically complete if its heat kernel satisfies

$$\int_{\mathcal{M}} p_t(x,y)\,\mathrm{vol}_{\mathcal{M}}(dy) = 1 \qquad \text{for all } x \in \mathcal{M},\ t > 0.$$

Combining stochastic completeness with the nonnegativity $p_t(x,y) \geq 0$ from (Grigoryan, 2013, Theorem 7.13), we obtain that

$$P_t(x, dy) := p_t(x,y)\,\mathrm{vol}_{\mathcal{M}}(dy)$$

is a conservative Markov kernel for every $t > 0$. In particular,

$$P_t(x, \mathcal{M}) = 1,$$

and for every bounded measurable function h,

$$\|P_t h\|_\infty \leq \|h\|_\infty.$$

Extension of the small-time limit to $C_b(\mathcal{M})$. The proof of Proposition 4.3 uses the small-time limit of the heat semigroup on bounded continuous test functions. Grigor'yan's heat-kernel theorem gives this limit first for compactly supported smooth functions; see (Grigoryan, 2013, Theorem 7.13). We record the standard extension.

Lemma A.3 (Small-time continuity on bounded continuous functions)**.** *Assume that $(\mathcal{M}, g)$ is stochastically complete, and let*

$$P_t(x, dy) := p_t(x,y)\,\mathrm{vol}_{\mathcal{M}}(dy)$$

be the heat-kernel Markov kernel. Then, for every $x \in \mathcal{M}$ and every $h \in C_b(\mathcal{M})$,

$$P_t h(x) = \int_{\mathcal{M}} h(y)\,P_t(x, dy) \xrightarrow[t\downarrow 0]{} h(x).$$

Proof. Fix $x \in \mathcal{M}$, and write

$$\nu_t^x(dy) := P_t(x, dy) = p_t(x,y)\,\mathrm{vol}_{\mathcal{M}}(dy).$$

By (Grigoryan, 2013, Theorem 7.13), for every $f \in C_c^\infty(\mathcal{M})$,

$$P_t f(x) = \int_{\mathcal{M}} f(y)\,\nu_t^x(dy) \xrightarrow[t\downarrow 0]{} f(x).$$

We first extend this convergence from $C_c^\infty(\mathcal{M})$ to $C_c(\mathcal{M})$. Let $f \in C_c(\mathcal{M})$, and let $\delta > 0$. Since $(\mathcal{M}, g)$ is a Riemannian manifold, $\mathcal{M}$ is smooth, and $C_c^\infty(\mathcal{M})$ is uniformly dense in $C_c(\mathcal{M})$; this follows from the locally compact Stone–Weierstrass theorem. Hence there exists $g \in C_c^\infty(\mathcal{M})$ such that

$$\|f - g\|_\infty < \delta.$$

Using the sub-Markov property of the heat semigroup,

$$|P_t(f-g)(x)| \leq \|f - g\|_\infty.$$

Therefore

$$|P_t f(x) - f(x)| \le |P_t(f-g)(x)| + |P_t g(x) - g(x)| + |g(x) - f(x)| \le 2\delta + |P_t g(x) - g(x)|.$$

Since $g \in C_c^\infty(\mathcal{M})$, there exists $t_0 > 0$ such that for all $0 < t < t_0$,

$$|P_t g(x) - g(x)| < \delta.$$

Therefore, for all $0 < t < t_0$,

$$|P_t f(x) - f(x)| \le 3\delta.$$

Since $\delta > 0$ was arbitrary, $P_t f(x) \to f(x)$.

We now prove the claim for $h \in C_b(\mathcal{M})$. Since $\mathcal{M}$ is a smooth manifold, smooth bump functions exist; hence there is $\chi \in C_c^\infty(\mathcal{M}) \subset C_c(\mathcal{M})$ such that

$$0 \le \chi \le 1, \qquad \chi(x) = 1,$$

see (Lee, 2012, Proposition 2.25). Then $h\chi \in C_c(\mathcal{M})$, so the C_c-convergence just proved gives

$$\int_{\mathcal{M}} h(y)\chi(y)\,\nu_t^x(dy) \xrightarrow[t\downarrow 0]{} h(x)\chi(x) = h(x).$$

Likewise, since $\chi \in C_c(\mathcal{M})$,

$$\int_{\mathcal{M}} \chi(y)\,\nu_t^x(dy) \xrightarrow[t\downarrow 0]{} \chi(x) = 1.$$

By stochastic completeness,

$$\nu_t^x(\mathcal{M}) = P_t(x, \mathcal{M}) = 1.$$

Therefore

$$\int_{\mathcal{M}} (1 - \chi(y))\,\nu_t^x(dy) = 1 - \int_{\mathcal{M}} \chi(y)\,\nu_t^x(dy) \xrightarrow[t\downarrow 0]{} 0.$$

Now decompose

$$P_t h(x) = \int_{\mathcal{M}} h(y)\chi(y)\,\nu_t^x(dy) + \int_{\mathcal{M}} h(y)(1-\chi(y))\,\nu_t^x(dy).$$

Hence

$$\begin{aligned} |P_t h(x) - h(x)| &\le \left| \int_{\mathcal{M}} h(y)\chi(y)\,\nu_t^x(dy) - h(x) \right| \\ &\quad + \int_{\mathcal{M}} |h(y)|(1-\chi(y))\,\nu_t^x(dy) \\ &\le \left| \int_{\mathcal{M}} h(y)\chi(y)\,\nu_t^x(dy) - h(x) \right| + \|h\|_\infty \int_{\mathcal{M}} (1-\chi(y))\,\nu_t^x(dy). \end{aligned}$$

Both terms on the right-hand side tend to 0. Therefore

$$P_t h(x) = \int_{\mathcal{M}} h(y)\,P_t(x, dy) \xrightarrow[t\downarrow 0]{} h(x),$$

as claimed. □

Geometric criteria for stochastic completeness. We now recall standard geometric criteria ensuring the stochastic-completeness assumption used above. Let $(\mathcal{M}, g)$ be a geodesically complete d-dimensional Riemannian manifold, and write

$$V(x_0, r) := \mathrm{vol}_{\mathcal{M}}(B(x_0, r)).$$

A volume-growth criterion of (Grigor'yan, 1999, Theorem 9.1) states that if, for some $x_0 \in \mathcal{M}$,

$$\int^\infty \frac{r\,dr}{\log V(x_0, r)} = \infty,$$

then $(\mathcal{M}, g)$ is stochastically complete. See also (Grigoryan, 2013, Ch. 11, Section 11.4, especially Theorem 11.8) for the corresponding criterion in Grigor'yan's monograph.

In particular, a Ricci lower bound implies stochastic completeness. Indeed, assume that

$$\mathrm{Ric}_{\mathcal{M}} \geq -(d-1)\kappa^2 g$$

for some $\kappa \geq 0$. Bishop–Gromov comparison applies to any complete d-dimensional Riemannian manifold satisfying this lower Ricci bound. It compares the volume growth of geodesic balls in $\mathcal{M}$ with the volume growth of balls in the simply connected d-dimensional model space of constant sectional curvature $-\kappa^2$. Consequently,

$$V_{\mathcal{M}}(x_0, r) \leq V_{\mathbb{H}^d_{-\kappa^2}}(r),$$

where the right-hand side denotes the ball volume in the corresponding hyperbolic model space, with the Euclidean model obtained when $\kappa = 0$. Since this model volume grows at most exponentially, there exist constants $A, C > 0$ such that, for all sufficiently large r,

$$V_{\mathcal{M}}(x_0, r) \leq Ae^{Cr}.$$

Consequently,

$$\log V_{\mathcal{M}}(x_0, r) \leq \log A + Cr,$$

and therefore

$$\int^{\infty} \frac{r\, dr}{\log V_{\mathcal{M}}(x_0, r)} \geq \int^{\infty} \frac{r\, dr}{\log A + Cr} = \infty.$$

The volume-growth criterion then gives stochastic completeness. This shows that every complete Riemannian manifold with Ricci curvature bounded from below is stochastically complete.

This criterion covers the geometric settings considered in the main text.

Compact manifolds. Every compact Riemannian manifold is geodesically complete by Hopf–Rinow. Its Ricci curvature is bounded from below: the Ricci tensor is a smooth symmetric 2-tensor field (Lee, 2019, Lemma 7.6), and the continuous function $(x, v) \mapsto \mathrm{Ric}_x(v, v)$ attains a finite minimum on the compact unit tangent bundle $S\mathcal{M}$. Hence compact Riemannian manifolds are stochastically complete by the Ricci-lower-bound criterion.

Euclidean spaces. For $\mathbb{R}^d$ with its standard metric, geodesic completeness is immediate and

$$\mathrm{Ric} = 0.$$

Thus $\mathbb{R}^d$ has Ricci curvature bounded from below and is stochastically complete.

Hyperbolic spaces. The d-dimensional hyperbolic space of constant sectional curvature $-\kappa^2$ is geodesically complete and satisfies

$$\mathrm{Ric} = -(d-1)\kappa^2 g.$$

Hence its Ricci curvature is bounded from below, and it is stochastically complete.

Products and $\mathrm{SE}(3)$. Products of covered examples are again covered. Indeed, geodesic completeness is preserved under products, and the Ricci tensor of a product metric is the product Ricci tensor. Therefore, if each factor is complete with Ricci curvature bounded from below, then so is the product. In particular,

$$\mathrm{SE}(3) \simeq \mathrm{SO}(3) \times \mathbb{R}^3,$$

equipped with the product metric used in our experiments, is stochastically complete: $\mathrm{SO}(3)$ is compact, while $\mathbb{R}^3$ has $\mathrm{Ric} = 0$.

SPD cone. The SPD cone endowed with the affine-invariant Riemannian metric used in our experiments is covered as well. Namely, on $\mathrm{SPD}(n)$ we use

$$g_X(U, V) = \mathrm{tr}(X^{-1}UX^{-1}V),$$

whose induced geodesic distance is

$$d_{\mathrm{AIRM}}(X, Y) = \left\| \log\left(X^{-1/2} Y X^{-1/2}\right) \right\|_F.$$

This is the standard affine-invariant metric, corresponding to the $(\alpha, \beta) = (1, 0)$ member of the affine-invariant family $g^{A(\alpha,\beta)}$ defined in (Thanwerdas & Pennec, 2023, Definition 3.3). For this family, (Thanwerdas & Pennec, 2023, Proposition 3.1) shows that $(\mathrm{SPD}(n), g^{A(\alpha,\beta)})$ is a Riemannian symmetric space, and therefore geodesically complete. Moreover, the curvature formulas in (Thanwerdas & Pennec, 2023, Table 5) show that, in particular for $(\alpha, \beta) = (1, 0)$, the sectional curvature is nonpositive and bounded below. Since Ricci curvature is the trace of sectional curvatures over an orthonormal basis (Lee, 2019, Ch. 7, "Ricci and Scalar Curvatures", after Lemma 7.6), a sectional-curvature lower bound $K \geq -K_0$ implies

$$\mathrm{Ric} \geq -(\dim \mathrm{SPD}(n) - 1)K_0\, g.$$

Thus $\mathrm{SPD}(n)$, and in particular $\mathrm{SPD}(3)$, equipped with the AIRM used in our experiments, is complete with Ricci curvature bounded from below, and is therefore stochastically complete.

B REVIEW ON GEOMETRIC DEEP LEARNING

Our approximation results are proved on the compact support K_ν, so we only need a compact-domain version of the usual feature-map transfer principle from geometric deep learning. The general philosophy is that one learns on a non-Euclidean domain by first representing points through a continuous feature map into a Euclidean space and then composing with a Euclidean approximator. A broad version of this principle is developed in (Kratsios & Bilokopytov, 2020); here we record the compact case needed in the paper.

Let $\mathcal{F} \subset C(\mathbb{R}^n, \mathbb{R})$ be dense under the topology of uniform convergence on compact subsets. Since K_ν is compact, approximation on K_ν is measured simply in the uniform norm

$$\|g\|_\infty := \sup_{x \in K_\nu} |g(x)|.$$

Given a continuous feature map $\varphi : K_\nu \to \mathbb{R}^n$, define the pullback class

$$\varphi^* \mathcal{F} := \{f \circ \varphi : \ f \in \mathcal{F}\}. \tag{18}$$

The key requirement is that φ separate points of K_ν.

Assumption B.1 (Injective feature map)**.** The feature map $\varphi : K_\nu \to \mathbb{R}^n$ is continuous and injective.

The following compact-domain transfer principle is a specialization of (Kratsios & Bilokopytov, 2020, Theorem 3.3).

Proposition B.2 (Transfer of Euclidean approximation through an injective feature map)**.** *Assume that $\mathcal{F} \subset C(\mathbb{R}^n, \mathbb{R})$ is dense under uniform convergence on compact subsets and that $\varphi : K_\nu \to \mathbb{R}^n$ is continuous and injective. Then $\varphi^* \mathcal{F}$ is dense in $C(K_\nu, \mathbb{R})$ under the uniform norm.*

For our purposes, a particularly useful class of feature maps is given by distance-to-landmark coordinates, following Gromov's distance-geometry viewpoint (Gromov, 1983). In the compact Riemannian setting, one can choose finitely many landmarks so that the associated distance map is continuous and injective. This provides an intrinsic feature map satisfying Assumption B.1.

Proposition B.3 (Finite distance-coordinate embedding, after Gromov)**.** *Assume that $\mathcal{M}$ is compact. Then there exist finitely many landmarks $x_1, \dots, x_N \in \mathcal{M}$ such that the map*

$$\varphi : \mathcal{M} \to \mathbb{R}^N, \qquad \varphi(x) := \big(d(x, x_1), \dots, d(x, x_N)\big),$$

is continuous and injective. In particular, $\varphi|_{K_\nu}$ satisfies Assumption B.1.

In the Cartan–Hadamard case there is an even simpler canonical choice of feature map. For any base point $z \in \mathcal{M}$, the Cartan–Hadamard theorem implies that the exponential map

$$\exp_z : T_z\mathcal{M} \to \mathcal{M}$$

is a global diffeomorphism. Hence its inverse

$$\mathrm{Log}_z : \mathcal{M} \to T_z\mathcal{M} \simeq \mathbb{R}^p$$

is globally defined and continuous, and therefore $\varphi := \mathrm{Log}_z|_{K_\nu}$ automatically satisfies Assumption B.1. Thus, on Cartan–Hadamard manifolds, Euclidean universal approximation transfers directly through this global chart.

C IMPLEMENTATION DETAILS

Unless otherwise specified, all experiments use the intrinsic quadratic transport cost

$$c(x, y) = \frac{1}{2} d(x, y)^2,$$

where d is the Riemannian geodesic distance on the corresponding manifold. Entropic RNOT parameterizes the target-side Schrödinger potential with per-batch gauge centering and a two-hidden-layer MLP of width 256, SiLU activations, Kaiming-normal hidden-layer initialization, and a small-initialized scalar output layer. Manifold inputs are embedded either through landmark/Gromov-distance features, used for $\mathbb{S}^2$, $\mathrm{SO}(3)$, and $\mathrm{SE}(3)$, or through Riemannian logarithmic coordinates, used for $\mathrm{SPD}(3)$ and $\mathbb{H}^2$. Layer normalization is applied after landmark-based embeddings and disabled after logarithmic embeddings.

The empirical semidual objective is optimized by minibatch stochastic gradient ascent using Adam with learning rate 10^{-3}, batch size 256, and cosine learning-rate decay. Unless otherwise stated, the entropic regularization is set by the median-cost heuristic

$$\varepsilon = 0.05 \times \mathrm{median}(C),$$

with the same value matched across Entropic RNOT and Sinkhorn-based methods. When a deterministic transport summary is required, we extract it from the entropic conditional distribution by heat-smoothed mode finding; the heat time is specified separately for each experiment. All experiments are implemented in JAX 0.6.2 with `float32` precision enabled and run on a single NVIDIA RTX A6000 GPU with 36 GB of memory. Sinkhorn references and baselines use OTT-JAX (Cuturi et al., 2022). Scaling results are averaged over three random seeds, with timing measured after explicit JIT warm-up.

C.1 BARYCENTRIC AND HEAT-SMOOTHED MAP EXTRACTION

Given a trained Entropic RNOT target-side Schrödinger potential g_θ, a source point x, and a discrete target support $\{y_j\}_{j=1}^M$, we first form the entropic conditional weights

$$w_j(x) = \frac{\beta_j \exp((g_\theta(y_j) - c(x, y_j))/\varepsilon)}{\sum_{k=1}^M \beta_k \exp((g_\theta(y_k) - c(x, y_k))/\varepsilon)},$$

where β_j denotes the target support mass. For uniform empirical targets, $\beta_j = 1/M$, and this factor cancels. These weights define the discrete conditional law

$$\pi_\theta(\cdot \mid x) = \sum_{j=1}^M w_j(x)\, \delta_{y_j}.$$

We consider two intrinsic deterministic summaries of this conditional law. The first is the Riemannian barycentric projection, defined as the Fréchet mean

$$T_{\mathrm{bar}}(x) \in \operatorname*{argmin}_{z \in \mathcal{M}} \frac{1}{2} \sum_{j=1}^M w_j(x)\, d(z, y_j)^2.$$

We approximate this minimizer by Karcher iterations. Starting from the target atom with largest conditional weight, $z_0 = y_{\arg\max_j w_j(x)}$, we repeat

$$v_\ell = \sum_{j=1}^M w_j(x) \operatorname{Log}_{z_\ell}(y_j), \qquad z_{\ell+1} = \operatorname{Exp}_{z_\ell}(\eta_{\mathrm{bar}} v_\ell),$$

followed by projection back to the manifold representation when required. In our implementation we use 32 iterations with $\eta_{\mathrm{bar}} = 0.5$. On Cartan–Hadamard manifolds, where squared distance is globally convex along geodesics, this is gradient descent on the Fréchet objective.

The second summary is the heat-smoothed mode used in our main evaluations. Ideally, for each conditional law one may smooth the weighted target atoms by the manifold heat kernel,

$$q_t(z) = \sum_{j=1}^M w_j(x)\, p_t(y_j, z),$$

where p_t is the heat kernel on $\mathcal{M}$. Since closed-form heat kernels are not available uniformly across the manifolds considered here, we use a Varadhan-type geodesic heat-kernel surrogate based on the leading short-time logarithmic asymptotic

$$\log p_t(y_j, z) \approx -\frac{d(y_j, z)^2}{4t} = -\frac{c(y_j, z)}{2t}.$$

This yields the intrinsic smoothed objective

$$\ell_t(z) = \log \sum_{j=1}^{M} w_j(x) \exp\left(-\frac{c(z, y_j)}{2t}\right),$$

which corresponds to smoothing by a geodesic Gaussian kernel. This surrogate keeps the smoothing intrinsic while avoiding manifold-specific heat-kernel normalization factors, curvature corrections, and volume-distortion terms.

We maximize ℓ_t by Riemannian gradient ascent. At the current iterate z_ℓ, define the heat-reweighted responsibilities

$$\alpha_j(z_\ell) = \frac{w_j(x) \exp(-c(z_\ell, y_j)/(2t))}{\sum_{k=1}^{M} w_k(x) \exp(-c(z_\ell, y_k)/(2t))}.$$

The ascent direction is the corresponding weighted log-map average,

$$v_\ell = \sum_{j=1}^{M} \alpha_j(z_\ell) \operatorname{Log}_{z_\ell}(y_j),$$

and we update

$$z_{\ell+1} = \operatorname{Exp}_{z_\ell}(\eta_{\text{heat}} v_\ell),$$

again projecting back to the manifold representation when necessary. The constant factor $1/(2t)$ in the exact gradient of the surrogate objective is absorbed into the step size. We use 32 ascent steps with $\eta_{\text{heat}} = 0.5$. To reduce sensitivity to initialization, the mode finder is run from multiple intrinsic initializers built from the conditional weights, including the heaviest target atom, and the candidate with largest final value of ℓ_t is selected.

For a discrete reference or baseline plan $P \in \mathbb{R}_+^{N \times M}$, the same extraction procedure is applied row-wise after normalizing each row:

$$w_j^{(i)} = \frac{P_{ij}}{\sum_{k=1}^{M} P_{ik}}.$$

Thus learned plans, Sinkhorn reference plans, and baseline plans are all converted into deterministic summaries using the same intrinsic procedure. For memory efficiency, the row-wise heat-smoothed extraction is evaluated in chunks.

C.2 Intrinsic-geometry benchmarks

This subsection provides the implementation details for the synthetic intrinsic-geometry benchmarks from Section 5.1. We describe the manifold models, the construction of source and target distributions, the numerical reference solutions, the baselines, the neural semidual model, and the evaluation protocol.

Manifold models and costs. We consider six manifold configurations spanning five manifold families.

Sphere $\mathbb{S}^2$. Points are represented as unit vectors in $\mathbb{R}^3$, with geodesic distance

$$d_{\mathbb{S}^2}(x, y) = \arccos(\langle x, y \rangle),$$

computed via the numerically stable $\operatorname{atan2}$ formulation.

Rotation group $\mathrm{SO}(3)$. Rotations are represented as unit quaternions $q \in \mathbb{R}^4$ with the antipodal identification $q \sim -q$ handled by sign normalization ($q_0 \geq 0$). The intrinsic distance is

$$d_{\mathrm{SO}(3)}(q_1, q_2) = 2 \arccos\big(\min(|\langle q_1, q_2 \rangle|, 1)\big),$$

computed via the same numerically stable $\mathrm{atan2}$ formulation as for $\mathbb{S}^2$, with antipodal identification handled by sign flipping when $\langle q_1, q_2 \rangle < 0$.

Rigid motions $\mathrm{SE}(3)$. Elements are represented as $(q, t) \in \mathbb{R}^7$ where q is a unit quaternion and $t \in \mathbb{R}^3$ is a translation. We use the weighted product metric

$$d^2_{\mathrm{SE}(3)}\big((q_1, t_1), (q_2, t_2)\big) = \alpha^2 \, d_{\mathrm{SO}(3)}(q_1, q_2)^2 + \|t_1 - t_2\|^2,$$

with coupling weight $\alpha = 2.0$ in the synthetic benchmark.

SPD manifold $\mathrm{SPD}(3)$*, affine-invariant.* We equip $\mathrm{SPD}(3)$ with the affine-invariant Riemannian metric (AIRM), for which

$$d_{\mathrm{AIRM}}(X, Y) = \big\|\log\big(X^{-1/2} Y X^{-1/2}\big)\big\|_F.$$

SPD manifold $\mathrm{SPD}(3)$*, log-Euclidean.* We additionally evaluate on $\mathrm{SPD}(3)$ equipped with the log-Euclidean metric

$$d_{\mathrm{LE}}(X, Y) = \|\log X - \log Y\|_F,$$

using the same underlying SPD support as the AIRM row. This produces a distinct reference plan and tests the method's ability to respect a different Riemannian structure on the same ambient space.

Hyperbolic plane $\mathbb{H}^2$. We use the Lorentz (hyperboloid) model

$$\mathbb{H}^2 = \{x \in \mathbb{R}^3 : -x_0^2 + x_1^2 + x_2^2 = -1, \; x_0 > 0\},$$

with geodesic distance

$$d_{\mathbb{H}^2}(x, y) = \mathrm{arcosh}\big(-\langle x, y \rangle_{\mathcal{M}}\big),$$

where $\langle \cdot, \cdot \rangle_{\mathcal{M}}$ is the Minkowski inner product.

Synthetic source and target distributions. For each manifold, source and target distributions are constructed as wrapped normal distributions: samples are drawn from a Gaussian in the tangent space at a prescribed center $p \in M$ and mapped to the manifold via the Riemannian exponential map,

$$v \sim \mathcal{N}(0, \sigma^2 I_{T_p M}), \qquad x = \exp_p(v).$$

The target distribution is obtained by shifting the center to a geodesically distant point, chosen to stress curvature effects and expose the failure of Euclidean approximations.

The specific benchmark instances are:

- $\mathbb{S}^2$. Source centered at the north pole $(0, 0, 1)$ with tangent scale $\sigma = 0.7$. Target centered at $(-0.5, 0, -0.866)$, approximately $150°$ from the north pole, with $\sigma = 0.7$. Both distributions wrap substantially around the sphere, making tangent-space linearization inaccurate.
- $\mathrm{SO}(3)$. Source centered at the identity quaternion. Target obtained by applying the exponential map to a tangent vector of norm 2.5 rad ($\approx 143°$ rotation), with tangent scale $\sigma = 0.8$. This places the target near the antipodal region where curvature distortion is maximal.
- $\mathrm{SPD}(3)$. Source centered at $\mathrm{diag}(4, 1, 0.25)$ (condition number 16) with $\sigma = 0.5$. Target centered at a $45°$ rotation of $\mathrm{diag}(0.25, 1, 4)$ in the 1–3 eigenplane, yielding center $\begin{pmatrix} 2.125 & 0 & -1.875 \\ 0 & 1 & 0 \\ -1.875 & 0 & 2.125 \end{pmatrix}$, with $\sigma = 0.5$. The non-commuting eigenbases ensure that AIRM and log-Euclidean geodesics genuinely diverge.
- $\mathrm{SPD}(3)(\mathrm{LE})$. Same source and target support as the AIRM row, but with the log-Euclidean cost replacing the affine-invariant cost. The discrete Sinkhorn reference is recomputed under the LE metric.
- $\mathrm{SE}(3)$. Source uniform on $\mathrm{SO}(3) \times [-4, 4]^3$ (Haar rotation, uniform translation). Target a $60°$ rotation about the z-axis paired with translation $(1.0, 0.5, -0.5)$, with $\sigma_{\mathrm{rot}} = 0.3$ and $\sigma_{\mathrm{trans}} = 0.5$ truncated to $[-4, 4]^3$. The weighted product metric ($\alpha = 2$) costs rotation $4\times$ translation, so the geodesic cost is dominated by the $\mathrm{SO}(3)$ component where tangent-space linearization is most distortive.
- $\mathbb{H}^2$. Source centered at the origin of the hyperboloid. Target at $\exp_o(2.0 \cdot e_1)$, a geodesic distance of ≈ 2.0 from the origin, with $\sigma = 0.5$.

Discrete reference entropic plan. For each benchmark, we compute a discrete reference solution on a shared evaluation support of $N = 200$ source and $M = 200$ target samples drawn from the respective distributions. The pairwise geodesic cost matrix $C_{ij} = \frac{1}{2}d(x_i, y_j)^2$ is formed using the intrinsic distance of each manifold, and the entropic OT problem is solved with the Sinkhorn algorithm in the log domain via OTT-JAX (Cuturi et al., 2022), using 200 iterations. The regularization parameter ε is set by the median-cost heuristic described above, with the median computed over 256×256 sampled cost pairs.

When a transport map is needed from the reference plan, we extract it using the common heat-smoothed mode-finding procedure with $t_{\text{heat}} = 100\varepsilon$.

Baselines. *Ambient Euclidean.* Points are treated as vectors in their ambient representation ($\mathbb{R}^3$ for $\mathbb{S}^2$ and $\mathbb{H}^2$; $\mathbb{R}^4$ for $\mathrm{SO}(3)$; $\mathbb{R}^7$ for $\mathrm{SE}(3)$; $\mathbb{R}^{3\times3}$ vectorized for $\mathrm{SPD}(3)$). Entropic OT is solved with squared Euclidean cost $c(x, y) = \frac{1}{2}\|x - y\|^2$ using the same Sinkhorn solver and matched ε.

Tangent-space. A reference point is computed as the intrinsic Fréchet mean of the source samples (50 gradient steps, step size 0.5). All source and target points are mapped to the tangent space at this reference via the Riemannian logarithm, Euclidean entropic OT is solved with squared Frobenius cost in that tangent space, and the transport summary is mapped back via the exponential map. This baseline is limited to a single chart.

Both baselines use 200 Sinkhorn iterations with the same ε as Entropic RNOT and the discrete reference.

Entropic RNOT model. Entropic RNOT uses the common gauge-centered target-side Schrödinger-potential parameterization described above. For the synthetic benchmarks, the manifold-specific input representation is chosen as follows:

- **Landmark/Gromov-distance embeddings.** For compact manifolds $(\mathbb{S}^2, \mathrm{SO}(3))$ and for the product manifold $\mathrm{SE}(3)$, we represent each input by its geodesic distances to 256 landmarks. The landmarks are selected by farthest-point sampling from a pool of 4,096 candidates drawn equally from the source and target distributions.
- **Logarithmic embeddings.** For non-compact manifolds $(\mathrm{SPD}(3)$ with both AIRM and log-Euclidean metrics, and $\mathbb{H}^2)$, we use Riemannian logarithmic coordinates at the manifold origin.

Optimization. For the synthetic benchmarks, training uses 3,000 stochastic-gradient iterations. The regularization parameter, optimizer, batch size, learning-rate schedule, and numerical precision follow the common setup above.

Transport map extraction. Given a trained Entropic RNOT potential g_θ, we evaluate the Gibbs conditional

$$\pi_\theta(dy \mid x) \propto \exp\left(\frac{g_\theta(y) - c(x, y)}{\varepsilon}\right) \nu(dy)$$

through normalized weights on the shared target support. The transport map is extracted via *heat-smoothed mode finding*: for each source point x_i, we find the mode of the heat-kernel–smoothed conditional density.

$$q_t(z) = \sum_j w_j \, p_t(y_j, z),$$

where w_j are the conditional weights and p_t is the manifold heat kernel at time $t_{\text{heat}} = 100\varepsilon$. The mode is found by multi-start Riemannian gradient ascent (32 steps, step size 0.5), with candidates initialized from the target atom with largest conditional weight.

Metric computation. All metrics are evaluated on the shared 200×200 support. Let $\widehat{\pi}$ denote the learned or baseline plan and let $\pi^\star$ denote the discrete manifold Sinkhorn reference plan on that support, computed using the intrinsic quadratic cost. To compare deterministic transport summaries, we apply the same heat-smoothed mode extractor to every plan. Namely, for each source point x_i, we

form the conditional target weights

$$\widehat{w}_j^{(i)} = \frac{\widehat{\pi}_{ij}}{\sum_k \widehat{\pi}_{ik}}, \qquad w_j^{\star,(i)} = \frac{\pi^\star_{ij}}{\sum_k \pi^\star_{ik}},$$

and define $\widehat{T}(x_i)$ and $T^\star(x_i)$ as the corresponding maximizers of the heat-smoothed target density. Thus $T^\star$ is the heat-smoothed mode of the discrete Sinkhorn reference plan, rather than an externally provided Monge map.

Plan KL. $\mathrm{KL}(\widehat{\pi} \,\|\, \pi^\star) = \sum_{i,j} \widehat{\pi}_{ij} \log(\widehat{\pi}_{ij}/\pi^\star_{ij})$, with a numerical floor of 10^{-30}.

Conditional W_1.

$$\mathrm{cW}_1(\widehat{\pi}, \pi^\star) = \frac{1}{N} \sum_{i=1}^{N} W_1\big(\widehat{\pi}(\cdot \mid x_i), \pi^\star(\cdot \mid x_i)\big),$$

where the inner W_1 is computed on the target manifold using the intrinsic distance. This conditional metric measures how accurately each method recovers the target conditional transport structure source point by source point, and does not allow errors to be hidden by rearranging mass across nearby source locations.

Map L^2. $\sqrt{(1/N)\sum_i d\big(\widehat{T}(x_i), T^\star(x_i)\big)^2}$, the RMS geodesic error of the heat-smoothed transport summaries.

Endpoint error. $(1/N)\sum_i d\big(\widehat{T}(x_i), T^\star(x_i)\big)$, the mean geodesic error of the transport summaries.

Hardware and software. The software stack, numerical precision, GPU, and Sinkhorn implementation are as described in the common implementation details above.

Default hyperparameters. Unless otherwise specified, the default choices across all synthetic benchmarks are summarized below.

Network depth	2
Hidden width	256
Batch size	256
Learning rate	10^{-3}
Regularization ε	$0.05 \times \mathrm{median}(C)$
Training iterations	3,000
Evaluation support size	200

C.3 Scalability and GPU efficiency

To examine scaling behavior, we use the same six manifold configurations and neural architecture as Appendix C.2, varying the support size $N_x = N_y = N \in \{128, 256, \ldots, 32{,}768\}$ over powers of two. For each N, fresh source and target samples are drawn, and timing is averaged over three seeds. We compare Entropic RNOT against discrete manifold Sinkhorn and Euclidean Sinkhorn. All methods use the matched median-cost regularization described in the common implementation details.

Timing excludes one-time JIT compilation via explicit warm-up passes. Peak memory is recorded via JAX's `peak_bytes_in_use` statistic; runs exceeding the device budget are marked as out-of-memory.

Measurement protocol. All timings are preceded by an explicit warm-up pass to exclude JAX JIT compilation from the timed block, and wrapped in `block_until_ready` to force synchronization. Because JAX's `peak_bytes_in_use` counter is monotonic per-process, we launch one subprocess per (manifold, method, N) cell via an orchestrator script; each subprocess runs all 3 seeds, exits, and writes its partial result to disk. Within a subprocess, throughput is measured only on the last seed so that the $\mathcal{O}(N \cdot B_y)$ transient cost matrices allocated during inference do not contaminate the training-peak snapshots of earlier seeds. Training-time peak memory reported in the figures is the mean over 3 uncontaminated seeds.

Throughput metrics. *Potential throughput* times the batched forward pass $x \mapsto g_\theta(x)$ on N source samples. *Transport throughput* additionally includes forming the neural entropic plan against a target batch of $B_y = 1{,}024$ samples and running heat-smoothed mode-finding at $t_{\text{heat}} = 100\varepsilon$. *Chunked transport throughput* processes source points in Python-level chunks of 256, yielding $\mathcal{O}(1)$-in-N inference memory ($\mathcal{O}(256 \cdot K \cdot D)$ peak by construction) at a modest throughput cost. All throughput figures are averaged over 10 timed calls following 3 warm-up calls.

Sinkhorn baselines. Both Sinkhorn baselines form the full $N \times N$ cost matrix and run 200 log-domain iterations via OTT-JAX (Cuturi et al., 2022); peak memory is recorded after `block_until_ready`, and runs that raise out-of-memory errors are marked as infeasible at that N. The Euclidean variant differs from the manifold variant only in using squared Euclidean cost in ambient coordinates, which isolates the contribution of intrinsic distance evaluation to the manifold Sinkhorn runtime.

C.4 Real-world pose refinement on $\mathrm{SE}(3)$ for protein–ligand docking

Motivation. A central computational problem in structure-based drug design is to refine candidate protein–ligand binding poses generated by docking (Trott & Olson, 2010). Given a receptor pocket and a ligand, docking software produces a ranked list of candidate poses, each specifying a position and orientation of the ligand within the pocket. These candidates are often noisy: high-ranked poses are not always correct, and low-ranked poses are not always wrong. Experimentally determining the true binding geometry requires X-ray crystallography or cryo-EM, which are expensive and slow.

In this work, we study a more specific problem than full de novo pose prediction: crystal-free refinement of docking pose ensembles using only information available at docking time. Concretely, we treat the candidate poses produced by a docking engine as empirical distributions on $\mathrm{SE}(3)$, and learn an Entropic RNOT refinement map that moves geometric outliers toward the docking engine's own top-ranked binding basin. Accordingly, our method should be interpreted as a docking-pose refinement or denoising procedure, not as a direct predictor of the crystal pose. The crystallographic pose $X_k^\star$ is used only for held-out evaluation, never for training or inference.

Dataset. We use CrossDocked2020 (Francoeur et al., 2020), a benchmark of protein–ligand complexes organized by pocket-similarity clusters. Let $\mathcal{K}$ denote the set of complexes and let $\mathcal{C}$ denote the set of pocket-similarity clusters, where each complex $k \in \mathcal{K}$ is assigned to a cluster $c(k) \in \mathcal{C}$. Two complexes in the same cluster share structurally related binding-site geometry. The dataset contains $|\mathcal{K}| = 3{,}765$ complexes across $|\mathcal{C}| = 1{,}302$ clusters. For each complex k, the available inputs are:

- receptor pocket heavy atoms $\mathcal{P}_k = \{p_{k,b}\}_{b=1}^{B_k} \subset \mathbb{R}^3$, where B_k is the number of non-hydrogen atoms in the binding site, the local protein environment that the ligand binds to;
- a canonical ligand conformer $L_k = \{\ell_{k,a}\}_{a=1}^{A_k} \subset \mathbb{R}^3$, where A_k is the number of ligand heavy atoms. This is a single known 3D geometry of the drug-like small molecule, used as the reference shape for rigid-body alignment;
- a crystallographic ligand pose $X_k^\star = \{x_{k,a}^\star\}_{a=1}^{A_k} \subset \mathbb{R}^3$, the experimentally determined binding geometry. We reserve the crystallographic ligand pose *for held-out evaluation only*, and it is s not used in training or pretraining.

Receptor-defined local frame. For each complex k, we define a deterministic orthonormal frame from receptor pocket heavy atoms $\mathcal{P}_k = \{p_{k,b}\}_{b=1}^{B_k} \subset \mathbb{R}^3$ alone. The origin is the pocket centroid $c_k = B_k^{-1} \sum_b p_{k,b}$. We select the three residue centroids $r_{k,1}, r_{k,2}, r_{k,3}$ nearest to c_k and apply Gram–Schmidt orthonormalization to obtain $Q_k = [e_{k,1}\ e_{k,2}\ e_{k,3}] \in \mathrm{SO}(3)$, with a sign convention that makes the frame deterministic. All ligand and pose coordinates are transformed into this frame via $\bar{x} = Q_k^\top (x - c_k)$, so that the subsequent $\mathrm{SE}(3)$ elements are defined relative to receptor geometry only. No ligand or crystal information is used.

Docking and rigid-body alignment. For each complex k, we generate $M = 40$ candidate docked poses using GNINA (McNutt et al., 2025) with the Vinardo (Quiroga & Villarreal, 2016) scoring

function. GNINA takes as input the receptor pocket structure $\mathcal{P}_k$ and the canonical ligand conformer L_k. The docking search box is centered at the pocket centroid c_k with side lengths equal to the coordinate-wise pocket extent plus a margin of $\delta_{\text{box}} = 4$ Å, defining a large search region from receptor geometry alone. We disable GNINA's CNN-based rescoring (`--cnn_scoring none`) to avoid data leakage from models trained on PDBbind crystal poses. This yields posed ligand coordinates $\{X_{k,m}\}_{m=1}^{M}$ with $X_{k,m} = \{x_{k,m,a}\}_{a=1}^{A_k} \subset \mathbb{R}^3$ and associated docking scores $\{s_{k,m}\}_{m=1}^{M}$. Note that both the canonical conformer L_k (the input geometry given to the docking tool) and the docked poses $X_{k,m}$ (the output of the docking tool) are available at inference time; neither requires the crystallographic pose $X_k^\star$. All coordinates are expressed in the receptor-defined local frame.

Each docked pose is converted to a rigid-body transform by Kabsch alignment to the canonical conformer:

$$(R_{k,m}, t_{k,m}) = \underset{R\in\mathrm{SO}(3),\, t\in\mathbb{R}^3}{\operatorname{argmin}} \frac{1}{A_k} \sum_{a=1}^{A_k} \|R\,\ell_{k,a} + t - x_{k,m,a}\|_2^2,$$

yielding a pose $g_{k,m} = (R_{k,m}, t_{k,m}) \in \mathrm{SE}(3)$ for each candidate. This representation is exact when the docked pose differs from the canonical conformer by rigid motion only. For flexible ligands, internal torsion changes during docking introduce error that rigid motion cannot capture. We quantify this by the rigid-fit residual

$$r_{k,m} = \left(\frac{1}{A_k} \sum_{a=1}^{A_k} \|R_{k,m}\,\ell_{k,a} + t_{k,m} - x_{k,m,a}\|_2^2 \right)^{1/2}.$$

We discard poses with $r_{k,m} > \tau_{\text{rigid}} = 2.5$ Å, retaining 155 complexes and 5,061 poses. After the source/target split and cluster-based train/test partition, this yields 117 training complexes (3,567 source, 559 target) across 75 clusters and 29 test complexes (814 source, 110 target) across 18 held-out clusters.

Since our pose representation models each docked ligand as a rigid transform of a fixed conformer, the near-rigid subset provides the most faithful setting for evaluating the method under its intended geometric assumptions. We therefore use this subset as our primary benchmark. At the same time, to test robustness beyond this idealised regime, we also evaluate on the full unfiltered docking ensembles. Performance remains comparable and is in several metrics even stronger without rigid filtering, suggesting that the method is reasonably robust to violations of the rigid-pose assumption. See Table 4. The consistency between the restricted and unfiltered evaluations also suggests that the conclusions drawn from the near-rigid benchmark are not merely an artifact of reduced sample size.

Source and target distributions. For each complex k, let $g_{k,1}^*$ denote the top-ranked pose by Vinardo score. We partition the retained poses into target and source sets by geometric proximity to the docking tool's highest-confidence prediction:

$$I_k^{\text{tgt}} = \{m : d_{\mathrm{SE}(3)}(g_{k,m}, g_{k,1}^*) \le \delta_{\text{prox}}\}, \qquad I_k^{\text{src}} = \{m : d_{\mathrm{SE}(3)}(g_{k,m}, g_{k,1}^*) > \delta_{\text{prox}}\},$$

with $\delta_{\text{prox}} = 5.0$ Å, the standard binding-mode tolerance used in docking evaluation (McNutt et al., 2025; Corso et al., 2022). Target poses are those that lie within the prescribed $\mathrm{SE}(3)$ neighborhood of the top-ranked prediction; source poses are geometric outliers. Complexes that yield empty source or target sets are discarded. This construction uses only the docking scores and the $\mathrm{SE}(3)$ metric; no crystallographic information is consulted. This construction makes the learning problem entirely crystal-free: the transport targets are defined solely from docking-generated poses and their geometry, without reference to the experimental structure. At the same time, the learned refinement is toward the docking engine's self-consistent top-ranked basin rather than directly toward the crystallographic pose.

The source and target empirical measures are

$$\mu_k = \frac{1}{|I_k^{\text{src}}|} \sum_{m\in I_k^{\text{src}}} \delta_{g_{k,m}}, \qquad \nu_k = \frac{1}{|I_k^{\text{tgt}}|} \sum_{m\in I_k^{\text{tgt}}} \delta_{g_{k,m}}.$$

Training and testing split. We split by pocket-similarity cluster to prevent structural leakage: if two complexes share a similar binding site, they are assigned to the same partition. With a 80/20 split, this yields 75 training clusters (117 complexes, 3,567 source and 559 target poses) and 18 held-out test clusters (29 complexes, 814 source and 110 target poses). No cluster appears in both partitions.

Empirical $\mathrm{SE}(3)$ **metric.** We equip $\mathrm{SE}(3)$ with the product metric $d_{\mathrm{SE}(3)}(g,h)^2 = \alpha^2 d_{\mathrm{SO}(3)}(R_g, R_h)^2 + \|t_g - t_h\|_2^2$. The weight α is set to the median radius of gyration of training ligands, $\alpha = \tilde{r}_{\mathrm{gyr}} = 2.44$ Å (computed from 117 training complexes), so that a rotation of θ rad contributes approximately $r_{\mathrm{gyr}} \cdot \theta$ Å to the distance, making rotational and translational components commensurate with heavy-atom RMSD. All training inputs are derived from the docking pipeline described above.

Training. Source and target poses are pooled across training complexes into empirical distributions μ and ν on $\mathrm{SE}(3)$. The entropic regularization is set to $\varepsilon = 0.05 \times \mathrm{median}(C) = 16.19$, where C is the pairwise $\mathrm{SE}(3)$ cost matrix on a calibration subsample. The Entropic RNOT semidual potential ψ_θ uses the common landmark-distance embedding and MLP architecture. Training follows the common optimizer setup, but uses $5{,}000$ steps for this experiment, taking approximately 14 seconds on one GPU.

Evaluation. Only at evaluation do we use the crystallographic pose $X_k^\star$. For each test complex k, we compute the crystal rigid-body transform $g_k^\star = (R_k^\star, t_k^\star) \in \mathrm{SE}(3)$ by the same Kabsch alignment to the canonical conformer in the receptor-defined local frame. For each source pose $g \in \mathrm{spt}(\mu_k)$, we compute the refined pose from the learned Entropic RNOT conditional using heat-smoothed mode finding with $t = \varepsilon$. At $t = \varepsilon$, the heat-smoothed mode concentrates on the locally dominant cluster of the entropic conditional rather than averaging across geometrically distant modes. This is particularly important on $\mathrm{SE}(3)$, where the conditional $\pi(\cdot \mid g)$ can be multimodal: a source pose far from the target region may have non-negligible weight on multiple distinct binding modes, and the Fréchet mean of these modes (recovered at large t) may lie in a geometrically invalid region of pose space between them.

Per-complex top-1 metrics (best transported pose per complex) are reported to enable comparison with other methods. All metrics are averaged over the 29 held-out test complexes (814 source poses, 18 clusters) and and bootstrapping with 100 resamples is used for 95% confidence intervals.

As a baseline, we also run GNINA's built-in local energy minimization (`--minimize`) on the same test source poses using the Vinardo scoring function, which refines each pose by optimization on the docking energy surface. This baseline uses the same information available at inference time (receptor, ligand, and docking scores) but operates independently on each pose without any learned transport.

As a further baseline, we run per-complex discrete Sinkhorn directly on $\mathrm{SE}(3)$: for each test complex, we form the intrinsic geodesic cost matrix, solve the entropic OT problem with matched ε (200 log-domain iterations via OTT-JAX), and extract the transport map by the same heat-smoothed mode-finding used for Entropic RNOT. This isolates the benefit of amortized cross-complex learning: with only a handful of target poses per complex, the per-complex discrete plan is severely under-determined, and the baseline's performance quantifies how much worse one does by solving each complex in isolation rather than pooling across the training distribution. This is visualized in Figure 3

D ALGORITHMS

Algorithm 1: Training Entropic RNOT via the neural semidual objective

1: **Input:** samples from μ and ν, regularization $\varepsilon > 0$, feature map φ, neural potential a_θ
2: **for** each training iteration **do**
3: Sample minibatches $\{x_i\}_{i=1}^B \sim \mu$ and $\{y_j\}_{j=1}^B \sim \nu$
4: Evaluate and center the target-side potential

$$g_\theta(y_j) = a_\theta(y_j) - \frac{1}{B}\sum_{\ell=1}^{B} a_\theta(y_\ell)$$

5: Compute the empirical soft c-transform

$$f_\theta^\varepsilon(x_i) = -\varepsilon \log\left(\frac{1}{B}\sum_{j=1}^{B}\exp\left(\frac{g_\theta(y_j) - c(x_i, y_j)}{\varepsilon}\right)\right)$$

6: Form $\widehat{\mathcal{J}}_\varepsilon(\theta)$ and take a gradient ascent step
7: **end for**
8: **Output:** Entropic RNOT potential g_θ

Algorithm 2: Intrinsic barycentric map induced by Entropic RNOT

1: **Input:** source point x, target support $\{y_j\}_{j=1}^K \sim \nu$, regularization $\varepsilon > 0$, neural potential ψ_θ
2: Evaluate the target-side potential $\psi_\theta(y_j)$
3: Compute the entropic conditional weights

$$w_j(x) = \frac{\exp\left(\frac{\psi_\theta(y_j) - c(x,y_j)}{\varepsilon}\right)}{\sum_{k=1}^K \exp\left(\frac{\psi_\theta(y_k) - c(x,y_k)}{\varepsilon}\right)}$$

4: Initialize z_0 from the heaviest atom $y_{j^\star}$, where

$$j^\star \in \arg\max_{j=1,\dots,K} w_j(x)$$

5: **for** $\ell = 0, \dots, T-1$ **do**
6: Compute the intrinsic barycenter direction

$$v_\ell = \sum_{j=1}^{K} w_j(x)\mathrm{Log}_{z_\ell}(y_j)$$

7: Update

$$z_{\ell+1} = \mathrm{Exp}_{z_\ell}(\eta\, v_\ell)$$

8: **end for**
9: **Output:** intrinsic entropic barycentric projection $T_\varepsilon(x) = z_T$

Algorithm 3: Heat-smoothed transport extraction from an Entropic RNOT plan

Input: transport plan $P \in \mathbb{R}_+^{N\times M}$, target support $\{y_j\}_{j=1}^M \subset \mathcal{M}$, diffusion time $t > 0$, manifold operations $\mathrm{Log}, \mathrm{Exp}$, initializer set $\{z_0^{(m)}\}_{m=1}^L$

for each source index $i = 1, \dots, N$ **do**

Normalize the i-th plan row into conditional weights

$$w_j^{(i)} = \frac{P_{ij}}{\sum_{k=1}^M P_{ik}}, \qquad j = 1, \dots, M$$

for each initialization $m = 1, \dots, L$ **do**

Set $z_0^{(m)} = z_0^{(m)} \in \mathcal{M}$

for $\ell = 0, \dots, T-1$ **do**

Compute the heat-kernel reweighted conditional coefficients

$$\alpha_{j,\ell}^{(m)} = \frac{w_j^{(i)} \exp\big(-c(z_\ell^{(m)}, y_j)/(2t)\big)}{\sum_{k=1}^M w_k^{(i)} \exp\big(-c(z_\ell^{(m)}, y_k)/(2t)\big)}$$

Form the intrinsic ascent direction

$$v_\ell^{(m)} = \sum_{j=1}^M \alpha_{j,\ell}^{(m)} \mathrm{Log}_{z_\ell^{(m)}}(y_j)$$

Take a Riemannian gradient ascent step

$$z_{\ell+1}^{(m)} = \mathrm{Exp}_{z_\ell^{(m)}}\big(\eta\, v_\ell^{(m)}\big)$$

end for

Evaluate the heat-smoothed log-density

$$\log q_t\big(z_T^{(m)}\big) = \log\left(\sum_{j=1}^M w_j^{(i)}\, p_t(y_j, z_T^{(m)})\right) \approx \log\left(\sum_{j=1}^M w_j^{(i)} \exp\big(-c(y_j, z_T^{(m)})/(2t)\big)\right)$$

end for

Select the best mode across initializations

$$T(x_i) = z_T^{(m^\star)}, \qquad m^\star \in \arg\max_{m=1,\dots,L} \log q_t\big(z_T^{(m)}\big)$$

end for

Output: transported points $\{T(x_i)\}_{i=1}^N \subset \mathcal{M}$

Algorithm 4: Heat-smoothed Gibbs mode extraction on a manifold

Input: conditional weights $\{w_j\}_{j=1}^M$, target support $\{y_j\}_{j=1}^M \subset \mathcal{M}$, diffusion time $t > 0$, step size η, initializers $\{z_0^{(m)}\}_{m=1}^L$

for each initialization $m = 1, \dots, L$ **do**

  Set $z_0^{(m)} \in \mathcal{M}$

  for $\ell = 0, \dots, T-1$ **do**

    Compute heat-smoothed weights

$$\alpha_{j,\ell}^{(m)} = \frac{w_j \exp\big(-c(z_\ell^{(m)}, y_j)/(2t)\big)}{\sum_{k=1}^M w_k \exp\big(-c(z_\ell^{(m)}, y_k)/(2t)\big)}$$

    Form the intrinsic ascent direction

$$v_\ell^{(m)} = \sum_{j=1}^M \alpha_{j,\ell}^{(m)} \mathrm{Log}_{z_\ell^{(m)}}(y_j)$$

    Update by exponential map

$$z_{\ell+1}^{(m)} = \mathrm{Exp}_{z_\ell^{(m)}}\big(\eta\, v_\ell^{(m)}\big)$$

  end for

  Compute the final objective

$$\log q_t\big(z_T^{(m)}\big) = \log\left(\sum_{j=1}^M w_j\, p_t(y_j, z_T^{(m)})\right)$$

end for

Return the maximizer

$$z^\star \in \arg\max_{m=1,\dots,L} \log q_t\big(z_T^{(m)}\big)$$

Output: heat-smoothed mode $z^\star$

E Additional Experimental Results

Additional synthetic results

Table 3 expands Table 1 with two additional metrics — conditional W_1 (cW_1) at the plan level and intrinsic RMS geodesic error ($\mathrm{Map}\ L^2$) at the map level. We also include one additional manifold, $\mathrm{SPD}(3)$ (LE), which evaluates the same SPD support under the log-Euclidean cost instead of the affine-invariant Riemannian metric.

Table 3: Synthetic transport benchmarks across manifold geometries. Lower is better for all metrics. Plan errors are computed relative to a discrete manifold Sinkhorn reference on a sampled support.

Manifold	Method	Plan KL ↓	cW_1 ↓	Map L^2 ↓	Endpoint error ↓
$\mathbb{S}^2$	Ambient Euclidean	0.7247 ± 0.0397	0.2275 ± 0.0011	0.3332 ± 0.0000	0.2141 ± 0.0000
	Tangent-space	0.4864 ± 0.0157	0.2138 ± 0.0009	0.3322 ± 0.0019	0.2094 ± 0.0009
	Entropic RNOT	**0.0411** ± **0.0027**	**0.0692** ± **0.0015**	**0.0880** ± **0.0035**	**0.0683** ± **0.0017**
$\mathrm{SO}(3)$	Ambient Euclidean	1.3975 ± 0.0338	0.6747 ± 0.0041	0.9426 ± 0.0145	0.7658 ± 0.0079
	Tangent-space	0.1462 ± 0.0034	0.1973 ± 0.0026	0.4619 ± 0.0164	0.2536 ± 0.0045
	Entropic RNOT	**0.0325** ± **0.0025**	**0.1175** ± **0.0046**	**0.3290** ± **0.0298**	**0.1635** ± **0.0049**
$\mathrm{SPD}(3)$	Ambient Euclidean	1.8590 ± 0.0320	0.8751 ± 0.0050	0.6820 ± 0.0049	0.6115 ± 0.0047
	Tangent-space	1.2355 ± 0.0176	0.7049 ± 0.0026	0.4687 ± 0.0029	0.4295 ± 0.0024
	Entropic RNOT	**0.0089** ± **0.0009**	**0.0670** ± **0.0039**	**0.0448** ± **0.0032**	**0.0428** ± **0.0032**
$\mathrm{SPD}(3)$ (LE)	Ambient Euclidean	1.6603 ± 0.0377	0.7680 ± 0.0038	0.5997 ± 0.0057	0.5299 ± 0.0046
	Tangent-space	1.5911 ± 0.0384	0.7475 ± 0.0037	0.5332 ± 0.0059	0.4840 ± 0.0047
	Entropic RNOT	**0.0280** ± **0.0030**	**0.1083** ± **0.0056**	**0.0875** ± **0.0047**	**0.0786** ± **0.0044**
$\mathrm{SE}(3)$	Ambient Euclidean	1.6818 ± 0.0363	0.8195 ± 0.0048	0.6826 ± 0.0036	0.6396 ± 0.0031
	Tangent-space	1.2940 ± 0.0231	0.6640 ± 0.0031	0.6104 ± 0.0039	0.5023 ± 0.0024
	Entropic RNOT	**0.0531** ± **0.0017**	**0.1606** ± **0.0020**	**0.1398** ± **0.0020**	**0.1189** ± **0.0015**
$\mathbb{H}^2$	Ambient Euclidean	0.9677 ± 0.0206	0.3127 ± 0.0004	0.3612 ± 0.0024	0.2451 ± 0.0013
	Tangent-space	0.1011 ± 0.0048	0.1170 ± 0.0015	0.1433 ± 0.0021	0.1067 ± 0.0016
	Entropic RNOT	**0.0101** ± **0.0012**	**0.0485** ± **0.0034**	**0.0438** ± **0.0032**	**0.0430** ± **0.0032**

Additional pose refinement results

Table 4 extends Table 2 to the full unfiltered CrossDocked2020 ensemble, dropping the 2.5 Å rigid-fit residual cutoff used in the main table to restrict evaluation to near-rigid poses. The same per-complex top-1 metrics are reported on the broader set, which includes poses exhibiting non-trivial conformational deviation from rigid alignment.

Table 4: Post-docking pose refinement on $\mathrm{SE}(3)$ (CrossDocked2020, 54 held-out clusters). Per-complex top-1 metrics with 95% bootstrap CIs. Crystal pose used for evaluation only. No rigid filter used.

Method	RMSD (Å) ↓	Median (Å) ↓	@2Å ↑	@5Å ↑
No refinement	$9.14_{[8.07,10.25]}$	$9.10_{[7.64,9.90]}$	$9.3\%_{[2.7,18.5]}$	$20.4\%_{[11.1,28.7]}$
GNINA –minimize	$9.17_{[7.77,11.02]}$	$8.92_{[6.39,9.85]}$	$0.0\%_{[0.0,0.0]}$	$18.8\%_{[6.2,34.4]}$
Sinkhorn SE(3)	$10.08_{[7.13,13.23]}$	$4.03_{[2.10,7.43]}$	$29.6\%_{[21.2,40.7]}$	$53.7\%_{[40.7,63.0]}$
Entropic RNOT	$\mathbf{1.68}_{[1.43,2.00]}$	$\mathbf{1.57}_{[1.33,1.82]}$	$\mathbf{72.2\%}_{[60.1,83.3]}$	$\mathbf{98.1\%}_{[93.5,100.0]}$

Entropic versus non-entropic RNOT

Table 5 compares Entropic RNOT with the non-entropic RNOT model of (Micheli et al., 2026), which requires an inner solve. Under matched architecture and batch size, the two methods reach comparable measure approximation on $\mathbb{S}^2$: pushforward Wasserstein distances to ν are similar, while Entropic RNOT is substantially faster and more memory efficient. Training time is shown in Figure 4.

Table 5: Entropic RNOT versus non-entropic RNOT on $\mathbb{S}^2$, mean $\pm$ std over 5 training seeds. Quality is the geodesic W_p between $T_{\#}\mu$ and a held-out target sample; cost is training wall-clock and peak GPU memory.

Metric	Entropic RNOT (entropic)	RNOT (non-entropic)
Quality		
$W_1(T_{\#}\mu, \nu) \downarrow$	0.1614 ± 0.0091	0.1850 ± 0.0068
$W_2(T_{\#}\mu, \nu) \downarrow$	0.1988 ± 0.0123	0.2263 ± 0.0083
Cost (N_{steps} = 3000 / 200, batch = 256)		
Train wall-clock (s) $\downarrow$	24.2 ± 0.6	82.3 ± 0.1
Peak memory (MB) $\downarrow$	88	201

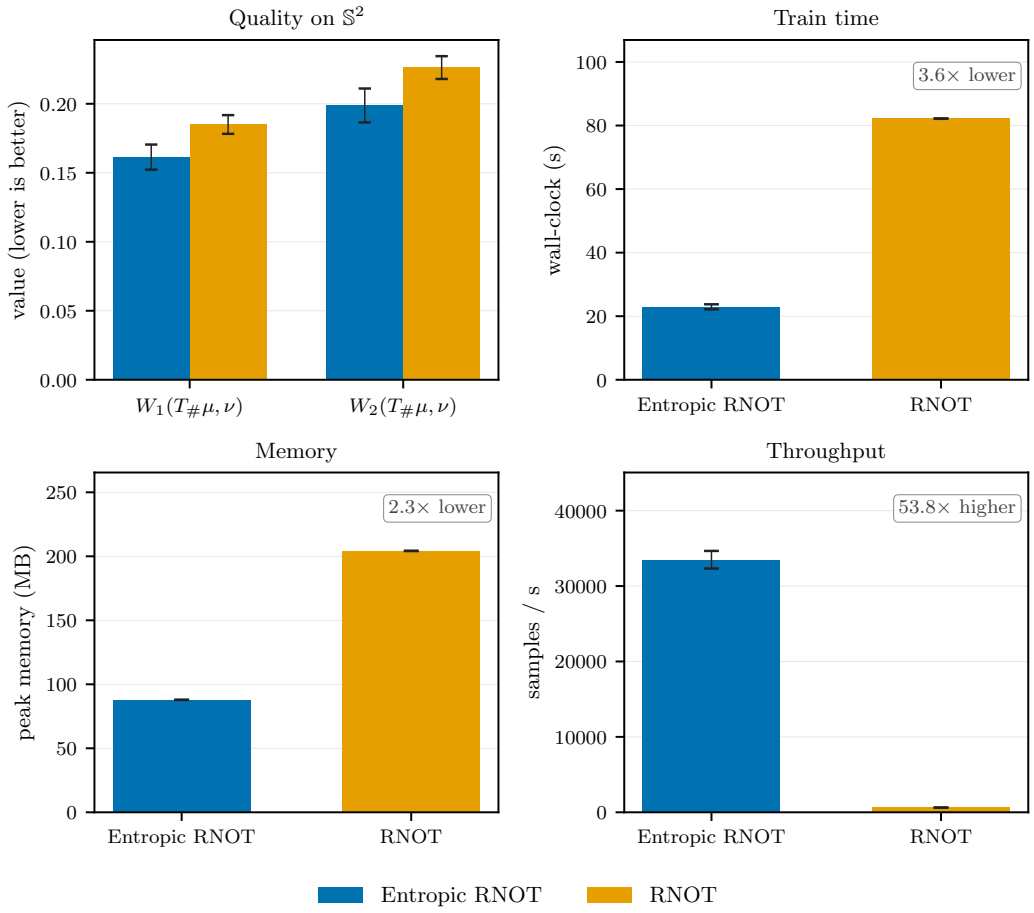


Figure 4: Entropic RNOT versus non-entropic RNOT on $\mathbb{S}^2$, mean $\pm$ std over training seeds. *Leftmost*: pushforward Wasserstein distances W_1, W_2 between $T_{\#}\mu$ and a held-out target sample of ν. *Right three panels*: training wall-clock, peak GPU memory, and training throughput under matched architecture and batch size; annotations report the relative ratio of non-entropic RNOT versus Entropic RNOT.

Table 6 extends this comparison to the remaining geometries in the main synthetic benchmark. Because RNOT returns a deterministic unregularized map, plan-level KL against the entropic Sinkhorn reference is not meaningful for it; we therefore compare the pushforward Wasserstein distances $W_1(T_{\#}\mu, \nu)$ and $W_2(T_{\#}\mu, \nu)$, computed against an independent held-out target sample, under matched source and target distributions, manifold representation, architecture, batch size, and per-seed initialization. On the compactly supported $\mathrm{SPD}(3)$ task, RNOT diverged across all five seeds,

whereas Entropic RNOT trained normally, so we report divergence rather than a numerical score. Entropic RNOT attains lower W_1 and W_2 and substantially higher inference throughput on every manifold where RNOT returns a valid map.

Table 6: Entropic RNOT versus non-entropic RNOT across manifold geometries, mean ± std over 5 training seeds. W_p is the geodesic Wasserstein distance between $T_{\#}\mu$ and a held-out target sample; throughput is inference points per second.

	$W_1 \downarrow$		$W_2 \downarrow$		Inference pt/s ↑	
Manifold	Ours	RNOT	Ours	RNOT	Ours	RNOT
$\mathrm{SO}(3)$	**0.4066** ± **0.0088**	0.4706 ± 0.0300	**0.4680** ± **0.0142**	0.5266 ± 0.0328	**104,050**	1,328
$\mathrm{SE}(3)$	**0.6305** ± **0.0031**	0.7795 ± 0.0235	**0.6729** ± **0.0035**	0.8179 ± 0.0233	**78,487**	1,177
$\mathbb{H}^2$	**0.2111** ± **0.0045**	0.5708 ± 0.0815	**0.2485** ± **0.0032**	0.6333 ± 0.0914	**137,863**	1,670
$\mathrm{SPD}(3)$	**0.6968** ± **0.0009**	diverged	**0.7287** ± **0.0010**	diverged	**951**	diverged

Heat-smoothed transport map and barycentric-projection comparison

Table 7 compares two transport extractors on the Cartan–Hadamard manifolds ($\mathrm{SPD}(3)$, $\mathrm{SPD}(3)$ (LE), and $\mathbb{H}^2$): the heat-smoothed mode and the Riemannian barycentric projection, which Proposition 4.2 shows makes the recovery guarantee well-defined on these manifolds. We use a large heat time t; in this regime, the heat-smoothed mode-finding iteration approaches the Riemannian barycenter. The two extractors agree to within 10^{-3} on every row. We therefore use barycentric projection for $\mathrm{SPD}(3)$, $\mathrm{SPD}(3)$ (LE), and $\mathbb{H}^2$, and the heat-smoothed mode for the compact and mixed-compact manifolds, where the Fréchet mean may fail to be unique.

Table 7: Effect of transport extractors on Cartan–Hadamard manifolds

	Heat-smoothed mode		Riemannian barycenter	
Method	Map $L^2 \downarrow$	Endpt err ↓	Map $L^2 \downarrow$	Endpt err ↓
$\mathrm{SPD}(3)$				
Ambient Euclidean	0.6806	0.6101	0.6802	0.6098
Tangent-space	0.4688	0.4293	0.4676	0.4286
Entropic RNOT	0.0439	0.0419	0.0438	0.0419
$\mathrm{SPD}(3)$ *(LE)*				
Ambient Euclidean	0.5988	0.5291	0.5983	0.5288
Tangent-space	0.5323	0.4831	0.5318	0.4828
Ours	0.0859	0.0771	0.0857	0.0770
$\mathbb{H}^2$				
Ambient Euclidean	0.3612	0.2448	0.3606	0.2447
Tangent-space	0.1428	0.1063	0.1427	0.1062
Ours	0.0424	0.0416	0.0424	0.0415

F BROADER IMPACTS

Our work is focused on generative modelling. The five manifold benchmarks ($\mathbb{S}^2$, $\mathrm{SO}(3)$, $\mathrm{SPD}(3)$, $\mathrm{SE}(3)$, $\mathbb{H}^2$) cover settings such as directional statistics, orientation estimation, covariance comparison, rigid-pose modelling, and hyperbolic representation learning, none of which carry an immediate dual-use concern. The most direct downstream application is the $\mathrm{SE}(3)$ protein–ligand pose-refinement experiment, which is framed as a refinement or denoising procedure for an existing docking ensemble rather than an end-to-end docking or generative pose-prediction system; the gap from pose refinement to de-novo molecule design is therefore substantial, but likely within current technical limits. Like all structure-based design tools, the underlying transport machinery could in principle be incorporated into a pipeline that optimises harmful small molecules; mitigation in this domain is governed by access controls on chemistry resources, restricted-substance regulations, and screening of synthesis providers, none of which are altered by the contributions of this paper.

G PROOFS

G.1 PROOF OF THEOREM 4.1

We work under the standing assumptions of Section 4. In particular, K_μ and K_ν are compact, the cost is $c(x,y) = \frac{1}{2}d(x,y)^2$, and the normalized entropic Schrödinger potential $g^\star_\varepsilon$ is continuous on K_ν. We use the pullback class $\varphi^*\mathcal{F}$ and the centered pullback class $\mathsf{C}_\nu(\varphi^*\mathcal{F})$ as defined in (6) and (7).

We first record that the centering operator C_ν, defined in (7), preserves uniform density on the centered subspace.

Lemma G.1 (Density of the centered class)**.** *Let $\mathcal{A} \subset C(K_\nu)$ be dense in $C(K_\nu)$ with respect to $\|\cdot\|_{L^\infty(K_\nu)}$. Then:*

(i) every $g \in \mathsf{C}_\nu(\mathcal{A})$ satisfies

$$\int_{K_\nu} g \, d\nu = 0;$$

(ii) for every $h \in C(K_\nu)$ with

$$\int_{K_\nu} h \, d\nu = 0,$$

there exists a sequence $(h_m)_{m\in\mathbb{N}} \subset \mathsf{C}_\nu(\mathcal{A})$ such that

$$\|h_m - h\|_{L^\infty(K_\nu)} \xrightarrow[m\to\infty]{} 0.$$

Proof. For (i), let $g \in \mathsf{C}_\nu(\mathcal{A})$. Then $g = \mathsf{C}_\nu a$ for some $a \in \mathcal{A}$, and hence

$$\int_{K_\nu} g \, d\nu = \int_{K_\nu} \left(a - \int_{K_\nu} a \, d\nu\right) d\nu = \int_{K_\nu} a \, d\nu - \left(\int_{K_\nu} a \, d\nu\right) \nu(K_\nu) = 0,$$

because $\nu(K_\nu) = 1$.

For (ii), let $h \in C(K_\nu)$ satisfy

$$\int_{K_\nu} h \, d\nu = 0.$$

Since $\mathcal{A}$ is dense in $C(K_\nu)$, there exists a sequence $(a_m)_{m\in\mathbb{N}} \subset \mathcal{A}$ such that

$$\|a_m - h\|_{L^\infty(K_\nu)} \to 0.$$

Define

$$h_m := \mathsf{C}_\nu a_m = a_m - \int_{K_\nu} a_m \, d\nu.$$

Then $h_m \in \mathsf{C}_\nu(\mathcal{A})$. Moreover, since h is centered,

$$\int_{K_\nu} a_m \, d\nu = \int_{K_\nu} (a_m - h) \, d\nu,$$

and therefore

$$h_m - h = a_m - h - \int_{K_\nu} (a_m - h)\, d\nu.$$

Thus

$$\begin{aligned}\|h_m - h\|_{L^\infty(K_\nu)} &\le \|a_m - h\|_{L^\infty(K_\nu)} + \left| \int_{K_\nu} (a_m - h)\, d\nu \right| \\ &\le \|a_m - h\|_{L^\infty(K_\nu)} + \int_{K_\nu} |a_m - h|\, d\nu \\ &\le 2\|a_m - h\|_{L^\infty(K_\nu)}.\end{aligned}$$

The right-hand side tends to 0, so

$$\|h_m - h\|_{L^\infty(K_\nu)} \to 0.$$

This proves (ii), and hence the lemma. □

We next record the elementary stability of the soft c-transform under uniform perturbations.

Lemma G.2 (Supremum-norm stability of the soft c-transform)**.** *For all $g, h \in C(K_\nu)$,*

$$\|\mathcal{T}_\nu^\varepsilon g - \mathcal{T}_\nu^\varepsilon h\|_{L^\infty(K_\mu)} \le \|g - h\|_{L^\infty(K_\nu)}.$$

Proof. Set

$$\delta := \|g - h\|_{L^\infty(K_\nu)}.$$

Then, for every $y \in K_\nu$,

$$h(y) - \delta \le g(y) \le h(y) + \delta.$$

Hence, for every $x \in K_\mu$,

$$e^{-\delta/\varepsilon} \int_{K_\nu} e^{(h(y)-c(x,y))/\varepsilon}\, d\nu(y) \le \int_{K_\nu} e^{(g(y)-c(x,y))/\varepsilon}\, d\nu(y) \le e^{\delta/\varepsilon} \int_{K_\nu} e^{(h(y)-c(x,y))/\varepsilon}\, d\nu(y).$$

Taking logarithms and multiplying by $-\varepsilon$ gives

$$|\mathcal{T}_\nu^\varepsilon g(x) - \mathcal{T}_\nu^\varepsilon h(x)| \le \delta \qquad \text{for every } x \in K_\mu.$$

Taking the supremum over $x \in K_\mu$ proves the claim. □

The next proposition shows that the normalized target-side entropic potential can be uniformly approximated by functions from the centered pullback class.

Proposition G.3 (Approximation of the normalized entropic potential)**.** *Assume that $\mathcal{F} \subset C(\mathbb{R}^n, \mathbb{R})$ is dense under the ucc topology and that the feature map $\varphi : K_\nu \to \mathbb{R}^n$ satisfies Assumption B.1. Let $g_\varepsilon^\star \in C(K_\nu)$ be the normalized target-side Schrödinger potential, chosen so that*

$$\int_{K_\nu} g_\varepsilon^\star\, d\nu = 0,$$

and set

$$f_\varepsilon^\star := \mathcal{T}_\nu^\varepsilon g_\varepsilon^\star.$$

Then there exists a sequence $(h_m)_{m\in\mathbb{N}} \subset \mathsf{C}_\nu(\varphi^*\mathcal{F})$ such that, as $m \to \infty$,

$$\|h_m - g_\varepsilon^\star\|_{L^\infty(K_\nu)} \to 0, \qquad \|\mathcal{T}_\nu^\varepsilon h_m - f_\varepsilon^\star\|_{L^\infty(K_\mu)} \to 0,$$

and

$$\mathcal{J}_\varepsilon(h_m) \to \mathrm{OT}_\varepsilon.$$

Proof. By Assumption B.1 and the transfer principle from Appendix B, the pullback class $\varphi^*\mathcal{F}$ is dense in $C(K_\nu)$ with respect to $\|\cdot\|_{L^\infty(K_\nu)}$. By Proposition A.1, in the compact-support setting we may choose the normalized target-side Schrödinger potential $g_\varepsilon^\star$ as an element of $C(K_\nu)$. Since

$$\int_{K_\nu} g_\varepsilon^\star\, d\nu = 0,$$

Lemma G.1 yields a sequence $(h_m)_{m\in\mathbb{N}} \subset \mathsf{C}_\nu(\varphi^*\mathcal{F})$ such that

$$\|h_m - g_\varepsilon^\star\|_{L^\infty(K_\nu)} \to 0.$$

Applying Lemma G.2, we obtain

$$\|\mathcal{T}_\nu^\varepsilon h_m - \mathcal{T}_\nu^\varepsilon g_\varepsilon^\star\|_{L^\infty(K_\mu)} \le \|h_m - g_\varepsilon^\star\|_{L^\infty(K_\nu)} \to 0.$$

Since

$$\mathcal{T}_\nu^\varepsilon g_\varepsilon^\star = f_\varepsilon^\star,$$

it follows that

$$\|\mathcal{T}_\nu^\varepsilon h_m - f_\varepsilon^\star\|_{L^\infty(K_\mu)} \to 0.$$

It remains to show convergence of the semidual values. Since $h_m \in \mathsf{C}_\nu(\varphi^*\mathcal{F})$, each h_m is centered:

$$\int_{K_\nu} h_m \, d\nu = 0.$$

Together with the normalization of $g_\varepsilon^\star$, this gives

$$\int_{K_\nu} (h_m - g_\varepsilon^\star) \, d\nu = 0.$$

Therefore,

$$\begin{aligned} |\mathcal{J}_\varepsilon(h_m) - \mathcal{J}_\varepsilon(g_\varepsilon^\star)| &= \left| \int_{K_\mu} \left(\mathcal{T}_\nu^\varepsilon h_m - \mathcal{T}_\nu^\varepsilon g_\varepsilon^\star\right) d\mu \right| \\ &\le \|\mathcal{T}_\nu^\varepsilon h_m - \mathcal{T}_\nu^\varepsilon g_\varepsilon^\star\|_{L^\infty(K_\mu)} \\ &\le \|h_m - g_\varepsilon^\star\|_{L^\infty(K_\nu)} \to 0. \end{aligned}$$

Since $g_\varepsilon^\star$ is a semidual maximizer,

$$\mathcal{J}_\varepsilon(g_\varepsilon^\star) = \mathrm{OT}_\varepsilon,$$

we conclude that

$$\mathcal{J}_\varepsilon(h_m) \to \mathrm{OT}_\varepsilon.$$

□

We next show how approximation of the semidual value translates into convergence of the induced Gibbs plans.

Proposition G.4 (Gap identity and convergence of Gibbs plans)**.** *For every bounded measurable function $g : K_\nu \to \mathbb{R}$, define*

$$d\pi_g^\varepsilon(x,y) := \exp\left(\frac{\mathcal{T}_\nu^\varepsilon g(x) + g(y) - c(x,y)}{\varepsilon}\right) d\mu(x)\, d\nu(y).$$

Then π_g^ε is a probability measure with first marginal μ, and

$$\varepsilon \,\mathrm{KL}(\pi_\varepsilon^\star \,\|\, \pi_g^\varepsilon) = \mathcal{J}_\varepsilon(g_\varepsilon^\star) - \mathcal{J}_\varepsilon(g). \tag{19}$$

Consequently,

$$\|\pi_g^\varepsilon - \pi_\varepsilon^\star\|_{\mathrm{TV}} \le \sqrt{\frac{\mathcal{J}_\varepsilon(g_\varepsilon^\star) - \mathcal{J}_\varepsilon(g)}{2\varepsilon}}. \tag{20}$$

Proof. By definition of $\mathcal{T}_\nu^\varepsilon g$, for every $x \in K_\mu$,

$$\int_{K_\nu} \exp\left(\frac{\mathcal{T}_\nu^\varepsilon g(x) + g(y) - c(x,y)}{\varepsilon}\right) d\nu(y) = 1.$$

Integrating this identity with respect to μ shows that π_g^ε is a probability measure. The same identity also implies that its first marginal is μ.

Next, using the Gibbs representation of $\pi^\star_\varepsilon$, we have

$$\log \frac{d\pi^\star_\varepsilon}{d\pi^\varepsilon_g}(x,y) = \frac{f^\star_\varepsilon(x) + g^\star_\varepsilon(y) - \mathcal{T}^\varepsilon_\nu g(x) - g(y)}{\varepsilon}.$$

Integrating with respect to $\pi^\star_\varepsilon$, and using that the marginals of $\pi^\star_\varepsilon$ are μ and ν, gives

$$\begin{aligned}\varepsilon \,\mathrm{KL}(\pi^\star_\varepsilon \,\|\, \pi^\varepsilon_g) &= \int_{K_\mu} \left(f^\star_\varepsilon - \mathcal{T}^\varepsilon_\nu g\right) d\mu + \int_{K_\nu} \left(g^\star_\varepsilon - g\right) d\nu \\ &= \mathcal{J}_\varepsilon(g^\star_\varepsilon) - \mathcal{J}_\varepsilon(g),\end{aligned}$$

because $f^\star_\varepsilon = \mathcal{T}^\varepsilon_\nu g^\star_\varepsilon$. Since $g^\star_\varepsilon$ is a semidual maximizer,

$$\mathcal{J}_\varepsilon(g^\star_\varepsilon) - \mathcal{J}_\varepsilon(g) \geq 0.$$

The total-variation bound then follows from Pinsker's inequality, using the probabilistic convention for total variation. □

We can now conclude the proof of Theorem 4.1.

Proof of Theorem 4.1. By Proposition G.3, there exists a sequence $(g_m)_{m\in\mathbb{N}} \subset \mathsf{C}_\nu(\varphi^*\mathcal{F})$ such that

$$\mathcal{J}_\varepsilon(g_m) \to \mathrm{OT}_\varepsilon \qquad \text{as } m \to \infty.$$

For each m, let

$$\pi^\varepsilon_m := \pi^\varepsilon_{g_m}$$

denote the Gibbs plan induced by g_m. Applying Proposition G.4, we obtain

$$\varepsilon \,\mathrm{KL}(\pi^\star_\varepsilon \,\|\, \pi^\varepsilon_m) = \mathcal{J}_\varepsilon(g^\star_\varepsilon) - \mathcal{J}_\varepsilon(g_m).$$

Since $g^\star_\varepsilon$ is a semidual maximizer,

$$\mathcal{J}_\varepsilon(g^\star_\varepsilon) = \mathrm{OT}_\varepsilon,$$

it follows that

$$\varepsilon \,\mathrm{KL}(\pi^\star_\varepsilon \,\|\, \pi^\varepsilon_m) = \mathrm{OT}_\varepsilon - \mathcal{J}_\varepsilon(g_m) \to 0.$$

Hence

$$\mathrm{KL}(\pi^\star_\varepsilon \,\|\, \pi^\varepsilon_m) \to 0.$$

Using (20), we also obtain

$$\|\pi^\varepsilon_m - \pi^\star_\varepsilon\|_{\mathrm{TV}} \leq \sqrt{\frac{\mathrm{OT}_\varepsilon - \mathcal{J}_\varepsilon(g_m)}{2\varepsilon}} \to 0.$$

Finally, total variation convergence implies weak convergence. Therefore,

$$\pi^\varepsilon_m \rightharpoonup \pi^\star_\varepsilon.$$

□

G.2 Proof of Proposition 4.2

We work under the assumptions of Proposition 4.2. Since $(\mathcal{M}, g)$ is Cartan–Hadamard, $(\mathcal{M}, d)$ is a global NPC space. We use standard facts about barycenters in such spaces: probability measures in $\mathcal{P}_2(\mathcal{M})$, and in particular compactly supported probability measures, admit unique barycenters for the squared-distance functional, the barycenter map is 1-Lipschitz with respect to W_1, and the variance inequality holds (Sturm, 2003, Propositions 4.3, 4.4 and Theorem 6.3).

Set

$$D_\nu := \mathrm{diam}(K_\nu) := \max_{y,y'\in K_\nu} d(y,y') < \infty.$$

The maximum is finite because K_ν is compact and d is continuous.

For any probability measure $\pi \in \mathcal{P}(K_\mu \times K_\nu)$ with first marginal μ, the disintegration theorem on standard Borel spaces gives a regular conditional kernel $x \mapsto \pi_x \in \mathcal{P}(K_\nu)$, unique μ-almost surely, such that

$$\pi(dx, dy) = \mu(dx)\, \pi_x(dy).$$

Since π_x is supported on the compact set K_ν, we have $\pi_x \in \mathcal{P}_2(\mathcal{M})$ for μ-almost every x. Hence its barycenter is well defined and unique:

$$T_\pi(x) := \mathrm{bar}(\pi_x) := \operatorname*{arg\,min}_{z\in\mathcal{M}} \frac{1}{2}\int_{\mathcal{M}} d(z,y)^2\,\pi_x(dy).$$

We will write

$$T_\varepsilon := T_{\pi^\star_\varepsilon}, \qquad \widehat{T}^\varepsilon_m := T_{\pi^\varepsilon_m}$$

for the barycentric projections associated with the optimal entropic coupling and the approximating Gibbs plans, respectively.

Before proving convergence, we first verify that these barycentric projections are measurable. This is needed because the estimates below involve integrals of the form

$$\int_{K_\mu} d\big(T_\pi(x), T_{\widetilde{\pi}}(x)\big)^q\,\mu(dx), \qquad q\in\{1,2\},$$

so the maps T_π must be well-defined as measurable maps, up to μ-null sets.

Lemma G.5 (Measurability of barycentric projections)**.** *For every probability measure $\pi \in \mathcal{P}(K_\mu \times K_\nu)$ with first marginal μ, the map*

$$T_\pi(x) = \mathrm{bar}(\pi_x)$$

is Borel measurable, up to modification on a μ-null set.

Proof. Since K_μ and K_ν are compact metric spaces, they are standard Borel spaces. Therefore, by the existence of regular conditional distributions and the disintegration theorem (Kallenberg, 2002, Theorems 6.3–6.4), there exists a μ-almost surely unique Markov kernel $x \mapsto \pi_x \in \mathcal{P}(K_\nu)$ such that

$$\pi(dx,dy) = \mu(dx)\,\pi_x(dy).$$

We first view this kernel as a measurable map into $\mathcal{P}(K_\nu)$ equipped with the weak Borel σ-algebra. By definition of a Markov kernel, for every Borel set $B \subset K_\nu$, the map

$$x \mapsto \pi_x(B)$$

is measurable. Since K_ν is compact metric, it is separable metrizable, and the Borel σ-algebra on $\mathcal{P}(K_\nu)$ induced by the weak topology is generated by the evaluation maps

$$e_B : \rho \mapsto \rho(B), \qquad B \in \mathcal{B}(K_\nu);$$

see (Bertsekas & Shreve, 1978, Proposition 7.25). For each such B, the composition of $x \mapsto \pi_x$ with e_B is precisely

$$(e_B \circ (x \mapsto \pi_x))(x) = \pi_x(B),$$

which is measurable by the Markov-kernel property. Since the maps e_B generate the weak Borel σ-algebra on $\mathcal{P}(K_\nu)$, it follows that $x \mapsto \pi_x$ is Borel measurable as a map into $\mathcal{P}(K_\nu)$ equipped with its weak Borel σ-algebra.

Since K_ν is compact, W_1-convergence on $\mathcal{P}(K_\nu)$ is equivalent to weak convergence: by the standard characterization of Wasserstein convergence, W_1-convergence is equivalent to weak convergence plus convergence of first moments, and the first-moment condition is automatic on compact sets (Villani, 2009, Theorem 6.9). Hence the weak and W_1 topologies induce the same Borel σ-algebra on $\mathcal{P}(K_\nu)$, so $x \mapsto \pi_x$ is also measurable as a map into $(\mathcal{P}(K_\nu), W_1)$.

The barycenter map is 1-Lipschitz with respect to W_1 in global NPC spaces (Sturm, 2003, Theorem 6.3). Therefore the composition

$$x \mapsto \pi_x \mapsto \mathrm{bar}(\pi_x)$$

is Borel measurable. Thus T_π is Borel measurable, up to modification on a μ-null set. □

We use the probabilistic convention for total variation: for probability measures $\rho, \widetilde{\rho}$ on a common measurable space,

$$\|\rho - \widetilde{\rho}\|_{\mathrm{TV}} := \sup_A |\rho(A) - \widetilde{\rho}(A)|.$$

Equivalently, if λ is any finite measure such that $\rho, \widetilde{\rho} \ll \lambda$, then

$$\|\rho - \widetilde{\rho}\|_{\mathrm{TV}} = \frac{1}{2} \int \left| \frac{d\rho}{d\lambda} - \frac{d\widetilde{\rho}}{d\lambda} \right| d\lambda.$$

We will also use the variational representation

$$\|\rho - \widetilde{\rho}\|_{\mathrm{TV}} = \sup_{\substack{g \text{ measurable} \\ |g| \le 1/2}} \left\{ \int g \, d\rho - \int g \, d\widetilde{\rho} \right\} = \frac{1}{2} \sup_{\substack{f \text{ measurable} \\ \|f\|_\infty \le 1}} \left| \int f \, d(\rho - \widetilde{\rho}) \right|.$$

This is the total-variation special case of the variational representation of f-divergences, obtained by taking $f(t) = \frac{1}{2}|t-1|$; see (Polyanskiy & Wu, 2025, Theorem 7.26 and Example 7.3).

We next record a fiberwise identity for total variation. This is a direct consequence of disintegration and the Radon–Nikodym theorem.

Lemma G.6 (Fiberwise total variation identity)**.** *Let $\pi, \widetilde{\pi} \in \mathcal{P}(K_\mu \times K_\nu)$ have first marginal μ, and write*

$$\pi(dx, dy) = \mu(dx)\, \pi_x(dy), \qquad \widetilde{\pi}(dx, dy) = \mu(dx)\, \widetilde{\pi}_x(dy).$$

Then

$$\|\pi - \widetilde{\pi}\|_{\mathrm{TV}} = \int_{K_\mu} \|\pi_x - \widetilde{\pi}_x\|_{\mathrm{TV}} \, \mu(dx).$$

Proof. Let

$$\lambda(dx, dy) := \mu(dx)\big(\pi_x + \widetilde{\pi}_x\big)(dy).$$

Then λ is a finite measure and $\pi, \widetilde{\pi} \ll \lambda$. Indeed, if $\lambda(A) = 0$ for a Borel set $A \subset K_\mu \times K_\nu$, then, writing

$$A_x := \{y \in K_\nu : (x, y) \in A\},$$

we have

$$0 = \lambda(A) = \int_{K_\mu} \big(\pi_x(A_x) + \widetilde{\pi}_x(A_x)\big)\, \mu(dx).$$

Since the integrand is nonnegative, this implies $\pi_x(A_x) = \widetilde{\pi}_x(A_x) = 0$ for μ-almost every x. Hence

$$\pi(A) = \int_{K_\mu} \pi_x(A_x)\, \mu(dx) = 0, \qquad \widetilde{\pi}(A) = \int_{K_\mu} \widetilde{\pi}_x(A_x)\, \mu(dx) = 0.$$

Thus $\pi, \widetilde{\pi} \ll \lambda$.

By the Radon–Nikodym theorem for kernels on standard Borel spaces, we may choose jointly measurable versions of the fiberwise densities

$$r(x, y) := \frac{d\pi_x}{d(\pi_x + \widetilde{\pi}_x)}(y), \qquad \widetilde{r}(x, y) := \frac{d\widetilde{\pi}_x}{d(\pi_x + \widetilde{\pi}_x)}(y).$$

We claim that

$$\frac{d\pi}{d\lambda}(x, y) = r(x, y), \qquad \frac{d\widetilde{\pi}}{d\lambda}(x, y) = \widetilde{r}(x, y).$$

To verify the first identity, let Φ be a bounded measurable function on $K_\mu \times K_\nu$. Then

$$\begin{aligned}
\int_{K_\mu \times K_\nu} \Phi(x, y) r(x, y)\, \lambda(dx, dy) &= \int_{K_\mu} \left[\int_{K_\nu} \Phi(x, y) r(x, y)\, (\pi_x + \widetilde{\pi}_x)(dy) \right] \mu(dx) \\
&= \int_{K_\mu} \left[\int_{K_\nu} \Phi(x, y)\, \pi_x(dy) \right] \mu(dx) \\
&= \int_{K_\mu \times K_\nu} \Phi(x, y)\, \pi(dx, dy).
\end{aligned}$$

Therefore $r = d\pi/d\lambda$. The proof that $\widetilde{r} = d\widetilde{\pi}/d\lambda$ is identical.

Using the density representation of total variation with respect to the common dominating measure λ, we obtain

$$2\|\pi - \widetilde{\pi}\|_{\mathrm{TV}} = \int_{K_\mu \times K_\nu} |r(x,y) - \widetilde{r}(x,y)|\, \lambda(dx,dy)$$
$$= \int_{K_\mu} \left(\int_{K_\nu} |r(x,y) - \widetilde{r}(x,y)|\, (\pi_x + \widetilde{\pi}_x)(dy) \right) \mu(dx).$$

For each fixed x, the measure $\pi_x + \widetilde{\pi}_x$ dominates both π_x and $\widetilde{\pi}_x$, with densities $r(x,\cdot)$ and $\widetilde{r}(x,\cdot)$, respectively. Therefore, again by the density representation of total variation,

$$\int_{K_\nu} |r(x,y) - \widetilde{r}(x,y)|\, (\pi_x + \widetilde{\pi}_x)(dy) = 2\|\pi_x - \widetilde{\pi}_x\|_{\mathrm{TV}} \qquad \text{for } \mu\text{-a.e. } x.$$

Substituting this into the previous display gives

$$2\|\pi - \widetilde{\pi}\|_{\mathrm{TV}} = 2 \int_{K_\mu} \|\pi_x - \widetilde{\pi}_x\|_{\mathrm{TV}}\, \mu(dx).$$

Dividing by 2 proves the identity. □

The next estimate converts total variation convergence of plans with first marginal μ into L^1 and L^2 convergence of their barycentric projections.

Proposition G.7 (From plan convergence to barycentric map convergence)**.** *Let $\pi, \widetilde{\pi} \in \mathcal{P}(K_\mu \times K_\nu)$ have first marginal μ. Then*

$$\int_{K_\mu} d\big(T_\pi(x), T_{\widetilde{\pi}}(x)\big)\, \mu(dx) \le D_\nu\, \|\pi - \widetilde{\pi}\|_{\mathrm{TV}}, \tag{21}$$

and

$$\int_{K_\mu} d\big(T_\pi(x), T_{\widetilde{\pi}}(x)\big)^2\, \mu(dx) \le 2D_\nu^2\, \|\pi - \widetilde{\pi}\|_{\mathrm{TV}}. \tag{22}$$

Proof. Write

$$\pi(dx,dy) = \mu(dx)\, \pi_x(dy), \qquad \widetilde{\pi}(dx,dy) = \mu(dx)\, \widetilde{\pi}_x(dy).$$

By Lemma G.5, the maps T_π and $T_{\widetilde{\pi}}$ are Borel measurable. Since d is continuous, the functions

$$x \mapsto d\big(T_\pi(x), T_{\widetilde{\pi}}(x)\big), \qquad x \mapsto d\big(T_\pi(x), T_{\widetilde{\pi}}(x)\big)^2$$

are measurable.

By the W_1-contraction property of barycenters in global NPC spaces (Sturm, 2003, Theorem 6.3),

$$d\big(T_\pi(x), T_{\widetilde{\pi}}(x)\big) \le W_1(\pi_x, \widetilde{\pi}_x) \qquad \mu\text{-a.e. } x.$$

We now bound the fiberwise W_1 distance by fiberwise total variation. Since both conditional laws are supported on K_ν, they belong to $\mathcal{P}_1(K_\nu)$. By the Kantorovich–Rubinstein duality formula (Villani, 2009, Theorem 5.10),

$$W_1(\pi_x, \widetilde{\pi}_x) = \sup_{\mathrm{Lip}(\phi) \le 1} \left| \int_{K_\nu} \phi\, d(\pi_x - \widetilde{\pi}_x) \right|.$$

Fix any 1-Lipschitz $\phi : K_\nu \to \mathbb{R}$. Since π_x and $\widetilde{\pi}_x$ are probability measures,

$$(\pi_x - \widetilde{\pi}_x)(K_\nu) = 0.$$

Thus subtracting constants from ϕ does not change the integral:

$$\int_{K_\nu} \phi\, d(\pi_x - \widetilde{\pi}_x) = \int_{K_\nu} (\phi - a)\, d(\pi_x - \widetilde{\pi}_x) \qquad \text{for every } a \in \mathbb{R}.$$

Choose $a := \inf_{K_\nu} \phi$. Then

$$0 \le \phi - a \le \mathrm{osc}_{K_\nu}(\phi).$$

If $\operatorname{osc}_{K_\nu}(\phi) = 0$, the integral is zero. Otherwise, set

$$h := \frac{\phi - a}{\operatorname{osc}_{K_\nu}(\phi)}.$$

Then $0 \le h \le 1$. Since

$$(\pi_x - \widetilde{\pi}_x)(K_\nu) = 0,$$

we may subtract $1/2$ from h without changing the integral:

$$\int_{K_\nu} h \, d(\pi_x - \widetilde{\pi}_x) = \int_{K_\nu} \left(h - \frac{1}{2}\right) d(\pi_x - \widetilde{\pi}_x).$$

Moreover, $\left|h - \frac{1}{2}\right| \le 1/2$. Hence, by the variational representation of total variation stated above,

$$\left| \int_{K_\nu} h \, d(\pi_x - \widetilde{\pi}_x) \right| \le \|\pi_x - \widetilde{\pi}_x\|_{\mathrm{TV}}.$$

Therefore,

$$\left| \int_{K_\nu} \phi \, d(\pi_x - \widetilde{\pi}_x) \right| \le \operatorname{osc}_{K_\nu}(\phi) \, \|\pi_x - \widetilde{\pi}_x\|_{\mathrm{TV}}.$$

Because ϕ is 1-Lipschitz and K_ν has diameter D_ν,

$$\operatorname{osc}_{K_\nu}(\phi) := \sup_{y, y' \in K_\nu} |\phi(y) - \phi(y')| \le D_\nu.$$

Thus

$$\left| \int_{K_\nu} \phi \, d(\pi_x - \widetilde{\pi}_x) \right| \le D_\nu \, \|\pi_x - \widetilde{\pi}_x\|_{\mathrm{TV}}.$$

Taking the supremum over all 1-Lipschitz ϕ yields

$$W_1(\pi_x, \widetilde{\pi}_x) \le D_\nu \, \|\pi_x - \widetilde{\pi}_x\|_{\mathrm{TV}}.$$

Combining this with the barycenter contraction gives

$$d\big(T_\pi(x), T_{\widetilde{\pi}}(x)\big) \le D_\nu \, \|\pi_x - \widetilde{\pi}_x\|_{\mathrm{TV}} \qquad \mu\text{-a.e. } x.$$

Integrating over K_μ and using Lemma G.6, we get

$$\int_{K_\mu} d\big(T_\pi(x), T_{\widetilde{\pi}}(x)\big) \, \mu(dx) \le D_\nu \int_{K_\mu} \|\pi_x - \widetilde{\pi}_x\|_{\mathrm{TV}} \, \mu(dx) = D_\nu \, \|\pi - \widetilde{\pi}\|_{\mathrm{TV}}.$$

This proves (21).

It remains to prove the L^2 estimate. Fix $y_0 \in K_\nu$. By the variance inequality for barycenters in global NPC spaces (Sturm, 2003, Proposition 4.4), for μ-almost every x,

$$d\big(T_\pi(x), y_0\big)^2 \le \int_{K_\nu} d(y_0, y)^2 \, \pi_x(dy).$$

Since π_x is supported on K_ν and $y_0 \in K_\nu$, we have $d(y_0, y) \le D_\nu$ for all $y \in K_\nu$. Therefore

$$d\big(T_\pi(x), y_0\big)^2 \le \int_{K_\nu} D_\nu^2 \, \pi_x(dy) = D_\nu^2.$$

Thus

$$d\big(T_\pi(x), y_0\big) \le D_\nu.$$

The same argument applied to $\widetilde{\pi}_x$ gives

$$d\big(T_{\widetilde{\pi}}(x), y_0\big) \le D_\nu.$$

By the triangle inequality,

$$d\big(T_\pi(x), T_{\widetilde{\pi}}(x)\big) \le d\big(T_\pi(x), y_0\big) + d\big(y_0, T_{\widetilde{\pi}}(x)\big) \le 2D_\nu \qquad \mu\text{-a.e. } x.$$

Hence, setting

$$a(x) := d\big(T_\pi(x), T_{\widetilde{\pi}}(x)\big),$$

we have $0 \le a(x) \le 2D_\nu$, and therefore

$$a(x)^2 \le 2D_\nu a(x).$$

Equivalently,

$$d\big(T_\pi(x), T_{\widetilde{\pi}}(x)\big)^2 \le 2D_\nu \, d\big(T_\pi(x), T_{\widetilde{\pi}}(x)\big).$$

Integrating and applying the L^1 bound (21),

$$\int_{K_\mu} d\big(T_\pi(x), T_{\widetilde{\pi}}(x)\big)^2 \, \mu(dx) \le 2D_\nu \int_{K_\mu} d\big(T_\pi(x), T_{\widetilde{\pi}}(x)\big) \, \mu(dx) \le 2D_\nu^2 \, \|\pi - \widetilde{\pi}\|_{\mathrm{TV}}.$$

This proves (22). □

We now prove the proposition.

Proof of Proposition 4.2. By Theorem 4.1, there exists a sequence $(g_m)_{m\in\mathbb{N}} \subset \mathsf{C}_\nu(\varphi^*\mathcal{F})$ such that, if π_m^ε denotes the induced Gibbs plan, then

$$\|\pi_m^\varepsilon - \pi_\varepsilon^\star\|_{\mathrm{TV}} \xrightarrow[m\to\infty]{} 0.$$

By Lemma G.5, both

$$\widehat{T}_m^\varepsilon = T_{\pi_m^\varepsilon}, \qquad T_\varepsilon = T_{\pi_\varepsilon^\star}$$

are Borel measurable, so the $L^1(\mu)$ and $L^2(\mu)$ distances below are well defined.

Applying Proposition G.7 with

$$\pi = \pi_m^\varepsilon, \qquad \widetilde{\pi} = \pi_\varepsilon^\star,$$

we obtain

$$\int_{K_\mu} d\big(\widehat{T}_m^\varepsilon(x), T_\varepsilon(x)\big) \, \mu(dx) \le D_\nu \, \|\pi_m^\varepsilon - \pi_\varepsilon^\star\|_{\mathrm{TV}} \xrightarrow[m\to\infty]{} 0,$$

and

$$\int_{K_\mu} d\big(\widehat{T}_m^\varepsilon(x), T_\varepsilon(x)\big)^2 \, \mu(dx) \le 2D_\nu^2 \, \|\pi_m^\varepsilon - \pi_\varepsilon^\star\|_{\mathrm{TV}} \xrightarrow[m\to\infty]{} 0.$$

Therefore

$$\widehat{T}_m^\varepsilon \to T_\varepsilon \qquad \text{in } L^2(\mu).$$

The $L^1(\mu)$ convergence follows either from the first bound above or from Cauchy–Schwarz, since μ is a probability measure. □

G.3 Proof of Proposition 4.3

Proof. Let $(g_m)_{m\in\mathbb{N}} \subset \mathsf{C}_\nu(\varphi^*\mathcal{F})$ be the approximating sequence given by Theorem 4.1. Thus the associated Gibbs plans satisfy

$$\|\pi_m^\varepsilon - \pi_\varepsilon^\star\|_{\mathrm{TV}} \xrightarrow[m\to\infty]{} 0.$$

Fix $t > 0$. By the heat-kernel facts recalled in Appendix A.3, the heat kernel $p_t(y, z)$ associated with the Riemannian heat semigroup is nonnegative and, under stochastic completeness, satisfies

$$\int_{\mathcal{M}} p_t(y, z) \, \mathrm{vol}_{\mathcal{M}}(dz) = 1.$$

Hence

$$P_t(y, dz) := p_t(y, z) \, \mathrm{vol}_{\mathcal{M}}(dz)$$

is a conservative Markov kernel on $\mathcal{M}$. This is the only point in the proof where stochastic completeness is used.

For a probability measure $\eta \in \mathcal{P}(\mathcal{M} \times \mathcal{M})$, let ηP_t denote the measure obtained by applying P_t in the second variable:

$$(\eta P_t)(C) := \int_{\mathcal{M}\times\mathcal{M}} \int_{\mathcal{M}} \mathbb{1}_C(x, z) \, P_t(y, dz) \, \eta(dx, dy), \qquad C \in \mathcal{B}(\mathcal{M} \times \mathcal{M}).$$

Equivalently, if $(x, y) \sim \eta$ and, conditionally on y, $z \sim P_t(y, \cdot)$, then $(x, z) \sim \eta P_t$. By the definition of the heat-smoothed transport surrogate,

$$\Pi^{\varepsilon}_{m,t} = \pi^{\varepsilon}_m P_t, \qquad \Pi^{\star}_{\varepsilon,t} = \pi^{\star}_{\varepsilon} P_t.$$

We first prove absolute continuity of the smoothed conditional laws. Since $\mathcal{M}$ is a Riemannian manifold, it is a standard Borel space, and the plans admit regular conditional distributions. Let $\pi^{\varepsilon}_{m,x}$ be a disintegration of π^{ε}_m with respect to its first marginal μ. Then, for μ-a.e. x,

$$Q^{\varepsilon}_{m,t}(x, dz) = \int_{\mathcal{M}} P_t(y, dz)\, \pi^{\varepsilon}_{m,x}(dy).$$

Using the heat-kernel representation recalled in Appendix A.3, for every measurable $B \subseteq \mathcal{M}$ we have

$$Q^{\varepsilon}_{m,t}(x, B) = \int_{\mathcal{M}} \left(\int_B p_t(y, z)\, \mathrm{vol}_{\mathcal{M}}(dz) \right) \pi^{\varepsilon}_{m,x}(dy).$$

Since $p_t(y, z) \geq 0$, Tonelli's theorem gives

$$Q^{\varepsilon}_{m,t}(x, B) = \int_B \left(\int_{\mathcal{M}} p_t(y, z)\, \pi^{\varepsilon}_{m,x}(dy) \right) \mathrm{vol}_{\mathcal{M}}(dz).$$

Hence

$$Q^{\varepsilon}_{m,t}(x, dz) = \left(\int_{\mathcal{M}} p_t(y, z)\, \pi^{\varepsilon}_{m,x}(dy) \right) \mathrm{vol}_{\mathcal{M}}(dz).$$

Therefore

$$Q^{\varepsilon}_{m,t}(x, \cdot) \ll \mathrm{vol}_{\mathcal{M}} \qquad \text{for } \mu\text{-a.e. } x,$$

with density

$$q^{\varepsilon}_{m,t}(x, z) = \int_{\mathcal{M}} p_t(y, z)\, \pi^{\varepsilon}_{m,x}(dy).$$

The same argument applied to a disintegration $\pi^{\star}_{\varepsilon,x}$ of $\pi^{\star}_{\varepsilon}$ shows that

$$Q^{\star}_{\varepsilon,t}(x, \cdot) \ll \mathrm{vol}_{\mathcal{M}} \qquad \text{for } \mu\text{-a.e. } x.$$

We now prove the total-variation contraction estimate. Let $C \in \mathcal{B}(\mathcal{M} \times \mathcal{M})$, and define

$$H_C(x, y) := \int_{\mathcal{M}} \mathbb{1}_C(x, z)\, P_t(y, dz).$$

Since P_t is a Markov kernel, $0 \leq H_C \leq 1$. Moreover,

$$\Pi^{\varepsilon}_{m,t}(C) - \Pi^{\star}_{\varepsilon,t}(C) = \int_{\mathcal{M}\times\mathcal{M}} H_C(x, y)\, d(\pi^{\varepsilon}_m - \pi^{\star}_{\varepsilon})(x, y).$$

We now prove the total-variation contraction estimate. Let $C \in \mathcal{B}(\mathcal{M} \times \mathcal{M})$, and define

$$H_C(x, y) := \int_{\mathcal{M}} \mathbb{1}_C(x, z)\, P_t(y, dz).$$

Since P_t is a Markov kernel, $0 \leq H_C \leq 1$. Moreover,

$$\Pi^{\varepsilon}_{m,t}(C) - \Pi^{\star}_{\varepsilon,t}(C) = \int_{\mathcal{M}\times\mathcal{M}} H_C(x, y)\, d(\pi^{\varepsilon}_m - \pi^{\star}_{\varepsilon})(x, y).$$

Since π^{ε}_m and $\pi^{\star}_{\varepsilon}$ are probability measures,

$$(\pi^{\varepsilon}_m - \pi^{\star}_{\varepsilon})(\mathcal{M} \times \mathcal{M}) = 0.$$

Thus

$$\int H_C\, d(\pi^{\varepsilon}_m - \pi^{\star}_{\varepsilon}) = \int \left(H_C - \frac{1}{2} \right) d(\pi^{\varepsilon}_m - \pi^{\star}_{\varepsilon}),$$

and $\left| H_C - \frac{1}{2} \right| \leq 1/2$. By the variational representation of total variation,

$$\left| \Pi^{\varepsilon}_{m,t}(C) - \Pi^{\star}_{\varepsilon,t}(C) \right| \leq \|\pi^{\varepsilon}_m - \pi^{\star}_{\varepsilon}\|_{\mathrm{TV}}.$$

Taking the supremum over $C \in \mathcal{B}(\mathcal{M} \times \mathcal{M})$ gives

$$\|\Pi^{\varepsilon}_{m,t} - \Pi^{\star}_{\varepsilon,t}\|_{\mathrm{TV}} \le \|\pi^{\varepsilon}_{m} - \pi^{\star}_{\varepsilon}\|_{\mathrm{TV}}.$$

Taking the supremum over $C \in \mathcal{B}(\mathcal{M} \times \mathcal{M})$ gives

$$\|\Pi^{\varepsilon}_{m,t} - \Pi^{\star}_{\varepsilon,t}\|_{\mathrm{TV}} \le \|\pi^{\varepsilon}_{m} - \pi^{\star}_{\varepsilon}\|_{\mathrm{TV}}.$$

Together with Theorem 4.1, this yields

$$\|\Pi^{\varepsilon}_{m,t} - \Pi^{\star}_{\varepsilon,t}\|_{\mathrm{TV}} \xrightarrow[m\to\infty]{} 0.$$

In particular,

$$\Pi^{\varepsilon}_{m,t} \rightharpoonup \Pi^{\star}_{\varepsilon,t} \qquad (m \to \infty)$$

for every fixed $t > 0$.

It remains to show that the population-level heat smoothing has vanishing bias:

$$\Pi^{\star}_{\varepsilon,t} \rightharpoonup \pi^{\star}_{\varepsilon} \qquad (t \downarrow 0).$$

Let $\Phi \in C_b(\mathcal{M} \times \mathcal{M})$. Then

$$\int_{\mathcal{M}\times\mathcal{M}} \Phi(x,z)\, \Pi^{\star}_{\varepsilon,t}(dx,dz) = \int_{\mathcal{M}\times\mathcal{M}} (P_t\Phi_x)(y)\, \pi^{\star}_{\varepsilon}(dx,dy), \qquad \Phi_x(z) := \Phi(x,z).$$

By the extension of the small-time heat-kernel limit to bounded continuous functions recalled in Appendix A.3, for every fixed $(x, y) \in \mathcal{M} \times \mathcal{M}$,

$$(P_t\Phi_x)(y) \xrightarrow[t\downarrow 0]{} \Phi_x(y) = \Phi(x,y),$$

because $\Phi_x \in C_b(\mathcal{M})$. Moreover, since P_t is a Markov kernel,

$$|(P_t\Phi_x)(y) - \Phi(x,y)| \le 2\|\Phi\|_{\infty}.$$

Dominated convergence therefore gives

$$\int \Phi\, d\Pi^{\star}_{\varepsilon,t} \xrightarrow[t\downarrow 0]{} \int \Phi\, d\pi^{\star}_{\varepsilon}.$$

Hence

$$\Pi^{\star}_{\varepsilon,t} \rightharpoonup \pi^{\star}_{\varepsilon} \qquad (t \downarrow 0).$$

Finally, let $t_m \downarrow 0$. For any $\Phi \in C_b(\mathcal{M} \times \mathcal{M})$,

$$\left|\int \Phi\, d\Pi^{\varepsilon}_{m,t_m} - \int \Phi\, d\pi^{\star}_{\varepsilon}\right| \le \left|\int \Phi\, d(\Pi^{\varepsilon}_{m,t_m} - \Pi^{\star}_{\varepsilon,t_m})\right| + \left|\int \Phi\, d\Pi^{\star}_{\varepsilon,t_m} - \int \Phi\, d\pi^{\star}_{\varepsilon}\right|.$$

For the first term, using the probabilistic convention for total variation,

$$\left|\int \Phi\, d(\Pi^{\varepsilon}_{m,t_m} - \Pi^{\star}_{\varepsilon,t_m})\right| \le 2\|\Phi\|_{\infty}\|\Pi^{\varepsilon}_{m,t_m} - \Pi^{\star}_{\varepsilon,t_m}\|_{\mathrm{TV}}.$$

By the contraction estimate proved above,

$$\|\Pi^{\varepsilon}_{m,t_m} - \Pi^{\star}_{\varepsilon,t_m}\|_{\mathrm{TV}} \le \|\pi^{\varepsilon}_{m} - \pi^{\star}_{\varepsilon}\|_{\mathrm{TV}} \xrightarrow[m\to\infty]{} 0.$$

The second term tends to zero because $t_m \downarrow 0$ and

$$\Pi^{\star}_{\varepsilon,t} \rightharpoonup \pi^{\star}_{\varepsilon} \qquad (t \downarrow 0).$$

Therefore

$$\Pi^{\varepsilon}_{m,t_m} \rightharpoonup \pi^{\star}_{\varepsilon},$$

which completes the proof. □