LEARNING OF POPULATION DYNAMICS: INVERSE OPTIMIZATION MEETS JKO SCHEME

Anonymous authors
Paper under double-blind review

ABSTRACT

Learning population dynamics involves recovering the underlying process that governs particle evolution, given evolutionary snapshots of samples at discrete time points. Recent methods frame this as an energy minimization problem in probability space and leverage the celebrated JKO scheme for efficient time discretization. In this work, we introduce `iJKOnet`, an approach that combines the JKO framework with inverse optimization techniques to learn population dynamics. Our method relies on a conventional *end-to-end* adversarial training procedure and does not require restrictive architectural choices, e.g., input-convex neural networks. We establish theoretical guarantees for our methodology and demonstrate improved performance over prior JKO-based methods.

1 INTRODUCTION

Modeling population dynamics is a fundamental challenge in many scientific domains, including biology (Schiebinger et al., 2019; Moon et al., 2019), ecology (Ayala et al., 1973), meteorology (Fisher et al., 2009; Sigrist et al., 2015; Verma et al., 2024; Price et al., 2025), transportation flows in urban networks (Medina-Salgado et al., 2022), and epidemiology (Wang et al., 2021; Kosma et al., 2023), among others. The task is to infer the underlying stochastic dynamics of a system – typically modeled by stochastic differential equations (SDEs) – from observed marginal distributions at discrete time points. While this problem has been studied extensively in settings where individual trajectories are available (Krishnan et al., 2017; Li et al., 2020; Brogat-Motte et al., 2024), such data are often unavailable in practice. In many real-world scenarios, we only observe *population-level data*, where it is infeasible to continuously track individual entities. Instead, we are forced to rely on temporally separated and mutually independent snapshots of the population.

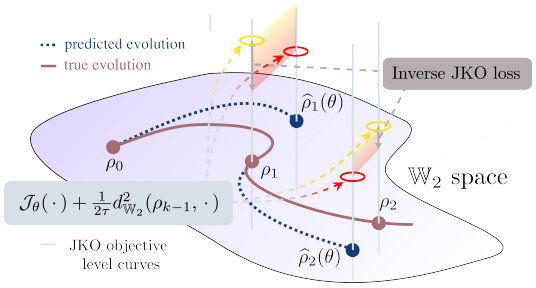


Figure 1: `iJKOnet` working scheme: our method minimizes the **gap** between the optimal values of parametric (θ) JKO functional and suboptimal values obtained at ground truth population measures.

In *single-cell genomics* (Macosko et al., 2015), for example, measuring the state of individual cells typically involves destructive sampling – a process where cells are destroyed during measurement, preventing any further observation of their future behavior. As a result, the data consist only of isolated profiles of cellular populations taken at discrete time points. Reconstructing the continuous developmental trajectories of cells from such fragmented data poses a challenge. A similar situation arises in *financial markets* (Gontis et al., 2010; Yang et al., 2023), where analysts often have access only to marginal distributions of asset prices at specific times, making it necessary to infer the underlying dynamics that govern these distributions. In *crowd dynamics* (Maury et al., 2010; 2011; Wan et al., 2023), modeling the temporal evolution of population densities is crucial for understanding and managing pedestrian flows. In this case, data on individual trajectories are also typically unavailable, and only aggregate distributions at various time points are observed.

A promising approach to address the absence of particle trajectories is the Jordan–Kinderlehrer–Otto (JKO) scheme (Jordan et al., 1998), which models the evolution of a particle system as a sequence of distributions that gradually approach the minimum of a total energy functional while remaining

close to the previous distributions. However, implementing the JKO scheme involves solving an optimization problem over the space of probability measures, which is computationally demanding. The first attempt to leverage JKO for learning population dynamics was introduced in (Bunne et al., 2022b, `JKOnet`). While innovative, this approach relies on a complex learning objective and is limited to potential energy functionals – meaning it cannot capture stochasticity in the dynamics. A recent work, (Terpin et al., 2024, `JKOnet*`), proposes replacing the JKO optimization step with its first-order optimality conditions. This relaxation allows modeling more general energy functionals and reduces computational complexity, but it does not support end-to-end training. Instead, it requires precomputing optimal transport couplings (Cuturi, 2013) between subsequent time snapshots, which limits its scalability and generalization.

Contributions. In this work, we present a novel approach for recovering the system dynamics based on observed population-level data. Our key contributions are as follows:

1. **Methodology:** In §3.2, we cast the problem of reconstructing energy functionals within the JKO framework as an inverse optimization task. This perspective leads to a novel min-max optimization objective for population dynamics recovery.
2. **Algorithm:** We equip our population dynamics methodology with a conventional end-to-end adversarial learning procedure (§3.3). Importantly, our practical scheme does not pose restrictions on the architectures of the utilized neural networks, which contributes to the scalability.
3. **Theoretical Guarantees:** Under suitable assumptions, we show that our method can accurately recover the underlying energy functional governing the observed population dynamics (§ 3.4).

In §5 and Appendix B, we evaluate our approach on a range of synthetic and real-world datasets, including single-cell genomics. The results show that our method demonstrates improved performance over previous JKO-based approaches for learning population dynamics.

Notation. Let $\mathcal{X}$ be a compact subset of $\mathbb{R}^D$ equipped with the Euclidean norm $\|\cdot\|_2$. Let $\mathcal{P}(\mathcal{X})$ denote the set of probability measures on $\mathcal{X}$, and let $\mathcal{P}_{ac}(\mathcal{X})$ denote its subset of probability measures absolutely continuous with respect to the Lebesgue measure. For $\rho \in \mathcal{P}_{ac}(\mathcal{X})$, we use ρ to denote both the measure and its density function with respect to the Lebesgue measure. For measures $\mu, \nu \in \mathcal{P}(\mathcal{X})$, we denote the set of couplings (*transportation plans*) between them by $\Pi(\mu, \nu)$. For a measure $\rho \in \mathcal{P}(\mathcal{X})$ and measurable map $T : \mathcal{X} \to \mathcal{X}$, we denote by $T\sharp\rho$ the associated *push-forward* measure, and $\mathrm{id}_{\mathcal{X}} : \mathcal{X} \to \mathcal{X}$ is the identity mapping. $\nabla \cdot F = \sum_{d=1}^{D} \frac{\partial F_d}{\partial x_d}$ denotes the *divergence* operator for a continuously differentiable vector field $F : \mathcal{X} \to \mathcal{X}$.

2 Background

To describe the evolution of population measures, we use the theoretical framework of *Wasserstein gradient flows* (WGFs). In what follows, we first introduce the preliminary concept of Optimal Transport between probability measures (Villani et al., 2008; Santambrogio, 2015). Then we provide sufficient theoretical background on WGFs; see (Santambrogio, 2017; Figalli & Glaudo, 2023) for an overview and (Ambrosio et al., 2008) for a comprehensive study of WGFs theory. Finally, we get the reader acquainted with the JKO scheme, which is the cornerstone of our approach.

Optimal Transport. The (squared) *Wasserstein-2 distance* $d_{\mathbb{W}_2}$ between two probability measures $\mu, \nu \in \mathcal{P}(\mathcal{X})$ is defined as the solution to the *Kantorovich* problem (Kantorovich, 1942):

$$d^2_{\mathbb{W}_2}(\mu, \nu) \stackrel{\text{def}}{=} \min_{\pi \in \Pi(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{X}} \|x - y\|_2^2 \, \mathrm{d}\pi(x, y), \tag{1}$$

where the distribution π^* delivering the minimum to (1) is called the *optimal coupling* (or *plan*). If $\mu \in \mathcal{P}_{ac}(\mathcal{X})$, then the Brenier's theorem (Brenier, 1991) establishes equivalence between the Kantorovich formulation (1) and *Monge's* problem (Monge, 1781):

$$d^2_{\mathbb{W}_2}(\mu, \nu) = \min_{T:T\sharp\mu=\nu} \int_{\mathcal{X}} \|x - T(x)\|_2^2 \, \mathrm{d}\mu(x), \tag{2}$$

where the optimal transport map T^* is known as the *Monge map*. Moreover, there exists a *unique* (up to an additive constant) convex potential $\psi^* : \mathcal{X} \to \mathbb{R}$ such that $T^* = \nabla\psi^*$ and $(\nabla\psi^*)\sharp\mu = \nu$ (McCann, 1995). In this setting, the optimal coupling in (1) is given by $\pi^* = [\mathrm{id}_{\mathcal{X}}, \nabla\psi^*]\sharp\mu$.

Wasserstein Gradient Flows. For clarity, we first consider the concept of gradient flows in Euclidean space $\mathbb{R}^D$ and then move to the space of probability measures $\mathcal{P}(\mathcal{X})$. For Euclidean space $\mathbb{R}^D$, a *gradient flow* is an absolutely continuous curve $x(t)$ starting at $x_0 \in \mathbb{R}^D$ that minimizes

an energy functional $J : \mathbb{R}^D \to \mathbb{R}$ "as fast as possible". To find such a curve, one needs to solve an ordinary differential equation (ODE) (Teschl, 2012) of the form $x'(t) = -\nabla J(x(t))$ with initial condition $x(0) = x_0$. The same idea is applicable to the space of probability measures $\mathcal{P}(\mathcal{X})$. If we equip $\mathcal{P}(\mathcal{X})$ with the Wasserstein-2 distance $d_{\mathbb{W}_2}$, we obtain the *Wasserstein space* $\mathbb{W}_2(\mathcal{X}) = (\mathcal{P}(\mathcal{X}), d_{\mathbb{W}_2})$ – a complete and separable metric space (Bogachev & Kolesnikov, 2012). In this case, for an energy functional $\mathcal{J} : \mathcal{P}(\mathcal{X}) \to \mathbb{R}$, the gradient flow in $\mathbb{W}_2(\mathcal{X})$, called the *Wasserstein gradient flow (WGF)*, is an absolutely continuous curve $\rho_t : \mathbb{R}_+ \to \mathcal{P}(\mathcal{X})$ starting at ρ_0 that follows the steepest descent direction of $\mathcal{J}$, i.e., solves

$$\partial_t \rho_t = -\nabla_{\mathbb{W}_2} \mathcal{J}(\rho_t), \quad \rho_{(t=0)} = \rho_0, \tag{3}$$

where $\nabla_{\mathbb{W}_2} \mathcal{J}(\rho_t)$ denotes the *Wasserstein gradient* in $\mathbb{W}_2$ given by $\nabla_{\mathbb{W}_2} \mathcal{J}(\rho) = -\nabla \cdot (\rho \nabla \frac{\delta \mathcal{J}}{\delta \rho})$ with $\frac{\delta \mathcal{J}}{\delta \rho}$ denoting the first variation (Chewi et al., 2025) of energy $\mathcal{J}$. Thus, equation (3) can be rewritten in the form of the *continuity equation*, expressing mass conservation under the velocity field v_t:

$$\partial_t \rho_t + \nabla \cdot (\rho_t v_t) = 0, \; v_t = -\nabla \frac{\delta \mathcal{J}}{\delta \rho}(\rho_t). \tag{4}$$

There exists an intriguing connection between WGFs and certain partial differential equations (PDEs) (Evans, 2022). In particular, different energy functionals in the Wasserstein space correspond to distinct PDEs (Santambrogio, 2015), some of which are associated with diffusion processes appearing in practice (Gómez-Castro, 2024; Bailo et al., 2024b).

Examples of PDEs as WGFs. Consider the *free energy* functional (Carrillo et al., 2003, Eq. (1.3)):

$$\mathcal{J}_{\text{FE}}(\rho) = \underbrace{\int_{\mathcal{X}} V(x) \, \mathrm{d}\rho(x)}_{\mathcal{V}(\rho)} + \underbrace{\int_{\mathcal{X}\times\mathcal{X}} W(x-y) \, \mathrm{d}\rho(x)\mathrm{d}\rho(y)}_{\mathcal{W}(\rho)} + \underbrace{\int_{\mathcal{X}} U(\rho(x)) \, \mathrm{d}x}_{\mathcal{U}(\rho)}, \tag{5}$$

where $\mathcal{V}$, $\mathcal{W}$, and $\mathcal{U}$ correspond to the system's *potential*, *interaction*, and *internal* energies, respectively. The PDE corresponding to the WGF driven by this functional is known as the *aggregation-diffusion* equation (ADE). It describes the evolution of density ρ_t under a corresponding velocity field v_t (4). This evolution reflects a balance of three effects: *drift* – driven by the potential $V(x)$, modeling an external field (e.g., gravity, electric potential); *interaction* – governed by the symmetric interaction kernel $W(x-y)$, accounting for non-local effects (e.g., particle interactions, long-range forces); and *diffusion* – modeled by the internal energy $\mathcal{U}(\rho)$, representing the energy associated with the local state of the system (e.g., thermal energy, chemical energy). A *weak* solution to the equation exists if the *no-flux conditions*: $\nabla V \cdot \vec{n} = 0$ on $\partial\mathcal{X}$ holds, where $\vec{n}$ is the outward normal vector on the boundary, and if $W \in \mathcal{C}(\mathcal{X} \times \mathcal{X})$, i.e. W is a continuous function on $\mathcal{X} \times \mathcal{X}$. These conditions ensure that no mass crosses the boundary of $\mathcal{X}$; mass can only be redistributed within $\mathcal{X}$. For a detailed discussion of the existence and classification of solutions, see (Gómez-Castro, 2024).

Such energy functionals are ubiquitous in real-world applications, including physics (Carrillo & Gvalani, 2021), biology (Keller & Segel, 1971; Potts & Painter, 2024; Potts, 2024), economics (Fiaschi & Ricci, 2025), machine learning (Suzuki et al., 2023; Chizat et al., 2024; Nitanda, 2024), and nonlinear optimization (Bailo et al., 2024a), to name a few. For further discussion and references, see the recent surveys (Carrillo et al., 2019a; Gómez-Castro, 2024; Bailo et al., 2024b).

To demonstrate the flexibility of the energy formulation in (5), we examine a few representative cases. When the system's energy is purely internal and given by the negative differential entropy, $\mathcal{U}(\rho) = -\mathcal{H}(\rho) \stackrel{\text{def}}{=} \int \rho(x) \log \rho(x) \mathrm{d}x$, continuity equation (4) reduces to the classical heat equation: $\partial_t \rho = \nabla^2 \rho$ (Vázquez, 2017). Alternatively, when the energy is $\mathcal{J}_{\text{FP}}(\rho) = \mathcal{V}(\rho) - \beta\mathcal{H}(\rho)$, the resulting PDE is the *linear Fokker-Planck equation* with diffusion coefficient β:

$$\partial_t \rho_t = \nabla \cdot (\nabla V(x) \, \rho_t) + \beta \nabla^2 \rho_t, \tag{6}$$

A fruitful branch of theoretical and practical research stems from the connection between the Fokker-Planck PDE and stochastic differential equations (SDEs) (Risken & Frank, 1996; Bogachev et al., 2022), the latter describes the stochastic evolution of *particles*. In particular, equation (6) is equivalent (Weinan et al., 2021) to the following *Itô SDE*:

$$\mathrm{d}X_t = -\nabla V(X_t) \, \mathrm{d}t + \sqrt{2\beta} \, \mathrm{d}W_t, \tag{7}$$

where $X = \{X_t\}_{t \geq 0}$ is a $\mathbb{R}^D$-valued stochastic process and $W = \{W_t\}_{t \geq 0}$ is a standard Wiener process (Särkkä & Solin, 2019). In other words, if $X_t \sim \rho_t$ evolves according to (7), then the density ρ_t evolves according to the Fokker-Planck equation (6) in the space of probability measures.

WGFs provide a compelling framework for modeling PDEs and their associated SDEs across various domains, giving rise to a range of methods for approximating solutions to WGFs. These include deep learning-based approaches (Mokrov et al., 2021; Alvarez-Melis et al., 2022; Altekrüger et al., 2023), particle-based methods such as Stein Variational Gradient Descent (SVGD) (Liu & Wang, 2016) and its extensions (Das & Nagaraj, 2023; Tankala et al., 2025), as well as classical discretization techniques in Wasserstein space (Carrillo et al., 2022, see §1.2). Many of these advances build on the celebrated Jordan-Kinderlehrer-Otto (JKO) scheme (Jordan et al., 1998), which we review next.

The JKO Scheme. A classical method for solving ODEs in Euclidean space is the *implicit Euler scheme*. For an (Euclidean) gradient flow $x'(t) = -\nabla J(x(t))$ and $\tau > 0$, this scheme approximately solves this ODE by iteratively applying the proximal operator (Parikh et al., 2014, §1.1): $x_{k+1} = \mathrm{prox}_{\tau J}(x_k) = \arg\min_{x \in \mathbb{R}^D} \left\{ J(x) + \frac{1}{2\tau}\|x - x_k\|_2^2 \right\}$. Compared to the explicit Euler scheme, this approach offers improved numerical stability (Butcher, 2016). Jordan, Kinderlehrer, and Otto (Jordan et al., 1998) extended this idea to the space of probability measures, introducing a variational time discretization of the Fokker–Planck equation (6), now known as the *JKO scheme*:

$$\rho_{k+1}^\tau = \arg\min_{\rho \in \mathcal{P}(\mathcal{X})} \left\{ \mathcal{J}(\rho) + \frac{1}{2\tau} d_{\mathbb{W}_2}^2(\rho, \rho_k^\tau) \right\} = \mathrm{JKO}_{\tau\mathcal{J}}(\rho_k^\tau), \quad \rho_0^\tau = \rho_0, \tag{8}$$

where $\tau > 0$ is the time step. As $\tau \to 0$, the sequence $\rho_k^\tau, k \in \mathbb{N}$ converges to the continuous solution ρ_t of the Fokker–Planck equation (6). Later, the JKO scheme's convergence was generalized (Ambrosio et al., 2008, Thm. 4.0.4) for the free energy functional $\mathcal{J}_{\mathrm{FE}}$ (5). Note that since $d_{\mathbb{W}_2}^2(\rho, \rho_k^\tau) \geq 0$, the energy is non-increasing along the sequence: $\mathcal{J}(\rho_{k+1}^\tau) \leq \mathcal{J}(\rho_k^\tau)$. This monotonicity property is often utilized in JKO-based algorithms (Salim et al., 2020) in Wasserstein space.

3 iJKOnet Method

We begin by formally stating the problem addressed in our work (§3.1), then we develop our methodology (§3.2). Finally, we discuss the practical (§3.3) and theoretical aspects (§3.4) of our approach.

3.1 Problem statement

In our work, we address the problem of recovering the underlying energy functional $\mathcal{J}^*$ that governs the evolution of a density $\rho_t \in \mathcal{P}_{ac}(\mathcal{X})$ in (4) based on marginal population measures (Bunne et al., 2022b; Lavenant et al., 2024). Specifically, we are given independent samples from marginals $\{\rho_k\}_{k=0}^K$ of the evolving distribution ρ_t at corresponding time points $t_0 < t_1 < \cdots < t_K$. Importantly, each distribution ρ_k may be represented by a different number of samples. As discussed in §2, the JKO scheme approaches the continuous-time dynamics as the step sizes $\Delta t_k = t_{k+1} - t_k$ are small enough. This motivates our modeling assumption that the ground truth sequence of measures $\{\rho_k\}_{k=0}^K$ follows the scheme $\rho_{k+1} = \mathrm{JKO}_{\Delta t_k \mathcal{J}^*}(\rho_k)$.

Although the time intervals Δt_k between observations may vary, corresponding to non-uniform step sizes in the JKO scheme, we assume equal spacing $\Delta t_k = \tau$ in the remaining text for simplicity.

3.2 Methodology and loss derivation

In this section, we assume $\rho_k \in \mathcal{P}_{ac}(\mathcal{X})$ and minimization is always taken over $\rho \in \mathcal{P}_{ac}(\mathcal{X})$ in order to utilize Brenier's theorem. The key idea of our method builds on *inverse* optimization techniques (Chan et al., 2025) and is demonstrated in Figure 1. Thanks to assumption $\rho_{k+1} = \mathrm{JKO}_{\tau\mathcal{J}^*}(\rho_k)$, we can derive an inequality that becomes an equality if a candidate functional $\mathcal{J}$ matches the ground truth functional $\mathcal{J}^*$, since ρ_{k+1} is obtained via the JKO step associated with $\mathcal{J}^*$:

$$\min_{\rho^k} \left\{ \mathcal{J}(\rho^k) + \frac{1}{2\tau} d_{\mathbb{W}_2}^2(\rho_k, \rho^k) \right\} \leq \mathcal{J}(\rho_{k+1}) + \frac{1}{2\tau} d_{\mathbb{W}_2}^2(\rho_k, \rho_{k+1}). \tag{9}$$

Moving the right-hand side to the left yields an expression that is always upper-bounded by zero, regardless of the choice of $\mathcal{J}$. Maximizing the resulting gap with respect to $\mathcal{J}$ encourages the candidate functional to approximate the true functional $\mathcal{J}^*$. This yields the following objective:

$$\max_{\mathcal{J}} \sum_{k=0}^{K-1} \left[\min_{\rho^k} \left\{ \mathcal{J}(\rho^k) + \frac{1}{2\tau} d_{\mathbb{W}_2}^2(\rho_k, \rho^k) \right\} - \mathcal{J}(\rho_{k+1}) - \underbrace{\frac{1}{2\tau} d_{\mathbb{W}_2}^2(\rho_k, \rho_{k+1})}_{\text{independent of } \mathcal{J},\, \rho^k} \right] =$$

$$\max_{\mathcal{J}} \sum_{k=0}^{K-1} \min_{\rho^k} \left[\mathcal{J}(\rho^k) - \mathcal{J}(\rho_{k+1}) + \frac{1}{2\tau} d^2_{\mathbb{W}_2}(\rho_k, \rho) \right] + \text{Const.} \tag{10}$$

Thanks to our assumption $\rho_k \in \mathcal{P}_{ac}(\mathcal{X})$, Brenier's theorem (Brenier, 1991) ensures that each distribution ρ^k appearing in (10) can be written as a pushforward $\rho^k = T^k \sharp \rho_k$ for some transport map $T^k : \mathcal{X} \to \mathcal{X}$. In addition, the Wasserstein distance (2) admits the upper-bound $d^2_{\mathbb{W}_2}(\rho_k, \rho) \leq \int_{\mathcal{X}} \|x - T^k(x)\|^2 d\rho_k(x)$. Since the minimizations over ρ^k (or equivalently over T^k) are independent across k, the sum and the minimization can be interchanged. Applying these observations, we end up with our *final loss objective* equivalent to (10):

$$\max_{\mathcal{J}} \min_{T^k} \mathcal{L}(\mathcal{J}, T^k) \stackrel{\text{def}}{=} \max_{\mathcal{J}} \min_{T^k} \sum_{k=0}^{K-1} \left[\mathcal{J}(T^k \sharp \rho_k) - \mathcal{J}(\rho_{k+1}) + \frac{1}{2\tau} \int_{\mathcal{X}} \|x - T^k(x)\|_2^2 \rho_k(x)\, \mathrm{d}x \right]. \tag{11}$$

The key idea is that the inner minimization over maps T^k approximates a JKO step for a given energy functional $\mathcal{J}$; that is, each optimal map $T^k_{\mathcal{J}} = \arg\min_{T^k} \left[\mathcal{J}(T^k \sharp \rho_k) + \frac{1}{2\tau} \int_{\mathcal{X}} \|x - T^k(x)\|_2^2 \rho_k(x)\, \mathrm{d}x \right]$ pushes ρ_k to the JKO-updated distribution $\hat{\rho}_{k+1} = \text{JKO}_{\tau\mathcal{J}}(\rho_k)$. The outer maximization then drives $\mathcal{J}$ toward the true functional $\mathcal{J}^*$, causing the pushforward distributions $\hat{\rho}_{k+1} = T^k_{\mathcal{J}} \sharp \rho_k$ to approach ρ_{k+1}, see illustration in Figure 1.

3.3 Practical aspects: method parametrization and learning procedure

We denote by $\theta \in \Theta$ and $\varphi \in \Phi$ the parameters of sufficiently expressive function classes used to approximate the candidate functional $\mathcal{J}$ and the transport maps T^k, respectively. Specifically, $\mathcal{J}_\theta$ and T^k_φ are their parameterized counterparts.

Mapping Parameterization. Building upon (Benamou et al., 2016), prior works (Mokrov et al., 2021; Alvarez-Melis et al., 2022; Bunne et al., 2022b) typically parameterize the transport maps as $T^k_\varphi = \nabla \psi^k_\varphi$, where ψ^k_φ are modeled using input-convex neural networks (ICNNs) (Amos et al., 2017). However, ICNNs suffer from poor scalability in high-dimensional settings (Korotin et al., 2021b). In contrast, since our objective (11) imposes *no convexity constraints*, we parameterize T^k_φ directly using standard architectures like MLPs or ResNets (He et al., 2016) This relaxation simplifies optimization and yields improved empirical stability. Alternatively, one can employ more task-specific parametrizations, such as triangular maps (Baptista et al., 2024).

Energy functional parametrization. Following (Terpin et al., 2024), we parameterize the energy functional $\mathcal{J}$ as a specific instance of the free energy formulation in (5), where the internal energy term is chosen to be scaled negative differential entropy:

$$\mathcal{J}_\theta(\rho) = \underbrace{\int_{\mathcal{X}} V_{\theta_1}(x)\, \mathrm{d}\rho(x)}_{\mathcal{V}_{\theta_1}(\rho)} + \underbrace{\int_{\mathcal{X}\times\mathcal{X}} W_{\theta_2}(x-y)\, \mathrm{d}\rho(x)\mathrm{d}\rho(y)}_{\mathcal{W}_{\theta_2}(\rho)} + \underbrace{\theta_3 \int_{\mathcal{X}} \log \rho(x)\, \mathrm{d}\rho(x)}_{\mathcal{U}_{\theta_3}(\rho) = -\theta_3 \mathcal{H}(\rho)}. \tag{12}$$

Here $\theta = \{\theta_1, \theta_2, \theta_3\}$ represents the set of learnable parameters: θ_1 and θ_2 are parameters of neural networks that define the potential $V_{\theta_1} : \mathcal{X} \to \mathbb{R}$ and the interaction kernel $W_{\theta_2} : \mathcal{X} \to \mathbb{R}$, respectively; $\theta_3 \in \mathbb{R}$ is a learnable scalar diffusion coefficient, see (7).

Entropy evaluation. All the terms in the objective function (11) can be estimated using Monte Carlo integration (Metropolis & Ulam, 1949), except for the internal energy term $\mathcal{U}_{\theta_3}$. To handle this remaining component, we follow (Mokrov et al., 2021, Theorem 1), (Alvarez-Melis et al., 2022, Eq. (11)) and apply the change-of-variables formula, yielding:

$$\mathcal{U}_{\theta_3}(T^k_\varphi \sharp \rho_k) = \mathcal{U}_{\theta_3}(\rho_k) - \theta_3 \int_{\mathcal{X}} \log |\det \nabla_x T^k_\varphi(x)|\, \mathrm{d}\rho_k(x), \tag{13}$$

where $\nabla_x T^k_\varphi$ denotes the Jacobian of the *invertible* transport map T^k_φ. We note that, during training, the computation of $\log\det$ – even for non-invertible mappings T^k_φ – can be effectively handled by modern programming tools; see Appendix C.3 for details. To efficiently compute the gradient of the $\log\det$ term in (13) with respect to φ, one can use the Hutchinson trace estimator (Hutchinson, 1989), as proposed by (Finlay et al., 2020; Huang et al., 2021). However, in our experimental setting (§5), we find that computing the full Jacobian is tractable and sufficient. For estimating the negative

differential entropy $\mathcal{H}(\rho_k)$ in (13), we employ the Kozachenko–Leonenko nearest-neighbor estimator (Kozachenko, 1987; Berrett et al., 2019). Notably, entropy values for the population measures $\{\rho_k\}_{k=0}^K$ can be precomputed prior to training.

Learning Objective. To facilitate gradient-based optimization, we employ Monte Carlo estimation to approximate the loss in (11). At each time step t_k (for $k = 0, \dots, K$), we draw a batch of B_k samples $X_k = \{x_k^1, \dots, x_k^{B_k}\} \sim \rho_k$ and optimize the following *empirical loss objective*:

$$\widehat{\mathcal{L}}(\theta, \varphi) = \sum_{k=0}^{K-1} \left[\widehat{\mathcal{J}}_\theta(T_\varphi^k(X_k)) - \widehat{\mathcal{J}}_\theta(X_{k+1}) + \frac{1}{2\tau} \sum_{j=1}^{B_k} \|x_k^j - T_\varphi^k(x_k^j)\|_2^2 \right], \tag{14}$$

where $T_\varphi^k(X_k)$ denotes the batched pushforward used to compute the empirical estimates $\widehat{\mathcal{J}}_\theta$ of the functional $\mathcal{J}_\theta$. We optimize the objective in (14) with respect to θ and φ using a standard gradient descent–ascent scheme; implementation details are provided in §5 and Appendix C, with the complete training procedure summarized in Algorithm 1 therein.

3.4 Theoretical aspects

The central idea behind our loss formulation (§3.2) is to minimize the *inverse gap* in (9), which measures the discrepancy between the optimal value (left-hand side) and the expected value (right-hand side) of the JKO step for a candidate functional $\mathcal{J}$. While this objective is intuitively justified, it remains to be shown whether the minimizer $\mathcal{J}_{\text{min}}$ of (11) truly approximates the ground-truth energy functional $\mathcal{J}^*$. Our Theorem 3.1 addresses this question, demonstrating that, under suitable assumptions, optimizing (11) indeed recovers $\mathcal{J}^*$ up to an additive constant that does not affect the dynamics governed by (4). Full details and proofs are provided in Appendix D.

We state our theorem for $K = 1$. When considering $K > 1$, the statement of the theorem holds true independently for each timestep $k = 0, 1, \dots, K-1$, and the results extend straightforwardly to this more general setting. We further assume that both the ground-truth functional $\mathcal{J}^*$ and the candidate functional $\mathcal{J}$ are purely of potential energy form, i.e., $\mathcal{J}^*(\rho) = \mathcal{J}^*_{\text{PE}}(\rho) = \int_{\mathcal{X}} V^*(x)\, \mathrm{d}\rho(x)$; $\mathcal{J}(\rho) = \mathcal{J}_{\text{PE}}(\rho) = \int_{\mathcal{X}} V(x)\, \mathrm{d}\rho(x)$. For notational simplicity, we denote the loss as $\mathcal{L}(V, T)$ in place of $\mathcal{L}(\mathcal{J}_{\text{PE}}, T)$. For technical purposes, we introduce the modified version of a potential as $V_q := \tau V + \frac{1}{2}\|\cdot\|_2^2 : \mathcal{X} \to \mathbb{R}$; subscript q stands for "quadratic".

Theorem 3.1 (Quality bounds for recovered potential energy)**.** *Let* $\varepsilon(V) \overset{def}{=} \mathcal{L}(V^*, T_{V^*}) - \mathcal{L}(V, T_V)$ *be the gap between the optimal and optimized value of inverse JKO loss* (11) *with internal* $\min_T$ *problem solved exactly, i.e.,* $T_V \overset{def}{=} \min_T \mathcal{L}(V, T)$*. Let* $\mathcal{X}$ *be a convex set; (modified) potentials* V_q *be strictly convex and* $\frac{1}{\beta}$*-smooth (see the definition in Appendix D). Then there exists* $C = C(\tau, \beta)$ *such that following inequality holds:*

$$\int_{\mathcal{X}} \left\| \nabla V^*(y) - \nabla V(y) \right\|^2 \mathrm{d}\rho_1(y) \leq C\varepsilon(V). \tag{15}$$

Notably, equation (15) compares *gradients* of the recovered and ground truth potentials, hiding the appearance of a redundant additive constant. Importantly, the assumptions of the theorem are not particularly restrictive in practice. Specifically, the smoothness of potentials can be ensured by employing smooth activation functions such as `CELU` (Barron, 2017), `SiLU` (Hendrycks & Gimpel, 2016), `SoftPlus`, and others. Strict convexity, in turn, can be enforced through architectural design choices (Amos et al., 2017). Furthermore, the strict convexity of the (modified) potentials V_q can often be assumed when the step size τ is sufficiently small. In our experiments (§5), we parameterize V using standard MLPs and observe that this approach performs adequately.

To the best of our knowledge, this work is the first to provide a quality analysis (Theorem 3.1) of JKO-based solvers for population dynamics. At present, our analysis focuses exclusively on the *potential* energy component. Extending this framework to incorporate *interaction* and *internal* energies presents an interesting direction for future research. In contrast, (Wu & Wang, 2025) analyzes the JKO scheme in terms of convergence to the true density of the Fokker-Planck equation (6), accounting for parameter uncertainty in iterative updates. Our work differs by focusing on the ability to restore energy components from data, rather than reconstructing the full density.

4 Related Works

This section reviews research directions most relevant to our work. For an extended discussion of related works, see Appendix A.1. First, we discuss existing methods for modeling the dynamics of ρ_k given a known energy functional $\mathcal{J}$ (§4.1). Next, we focus on approaches closely related to our setting – specifically, learning population dynamics from observed data (§4.2) via the JKO scheme.

4.1 Solving aggregation-diffusion equation with given energy

This line of research focuses on methods for *solving* aggregation-diffusion equations derived from WGFs with given energy functionals of the form (5). A variety of numerical approaches exist, including mesh-based schemes (Cancès et al., 2023; Jüngel et al., 2024), particle methods (Craig & Bertozzi, 2016; Campos Pinto et al., 2018; Carrillo et al., 2019b), and JKO-based schemes (Carrillo et al., 2022). For a comprehensive overview, see (Bailo et al., 2024b, §5). Recently, the research field was empowered by deep learning techniques. In particular, (Alvarez-Melis et al., 2022; Mokrov et al., 2021; Fan et al., 2022; Park et al., 2023; Lee et al., 2024) employ gradient-based optimization and neural-network parametrization to solve WGFs with given energy $\mathcal{J}$ as to special cases of (5).

4.2 Learning energy for aggregation-diffusion equation via JKO scheme

This line of research focuses on methods for *learning* population dynamics (§3.1) using the JKO scheme, as in `JKOnet` (Bunne et al., 2022b) and `JKOnet`* (Terpin et al., 2024). We discuss these methods in detail below, as they are the most closely related to our approach. Additional methods for learning population dynamics are reviewed in Appendix A.1.1.

JKOnet (Bunne et al., 2022b) formulates the task of population dynamics recovery as a bi-level optimization problem aimed at minimizing the discrepancy between observed distributions ρ_k and model predictions $\hat{\rho}_k$:

$$\begin{aligned}\mathcal{L}_{\texttt{JKOnet}}(\theta,\varphi) = \sum_{k=0}^{K-1} d^2_{\mathbb{W}_2}(\hat{\rho}_k,\rho_k), \quad \text{s.t. } \hat{\rho}_0=\rho_0, \quad \hat{\rho}_{k+1} = \nabla\psi_k^*\sharp\hat{\rho}_k,\\ \psi_k^* \overset{\text{def}}{=} \operatorname*{arg\,min}_{\varphi:\psi_\varphi\in\texttt{CVX}} \mathcal{J}_\theta(\nabla\psi_\varphi\sharp\hat{\rho}_k) + \frac{1}{2\tau}\int_{\mathcal{X}} \|x-\nabla\psi_\varphi(x)\|^2 \,\mathrm{d}\hat{\rho}_k(x),\end{aligned} \tag{16}$$

where `CVX` denotes the set of continuously differentiable convex functions from $\mathcal{X}$ to $\mathbb{R}$. The transport maps ψ_φ are parametrized either using ICNNs (Bunne et al., 2022a) or through the 'Monge gap' approach (Uscidda & Cuturi, 2023). However, the energy functional is limited to the potential energy term $\mathcal{V}_\theta(\rho)$, excluding interaction and internal components. This framework has two key limitations: **(i)** solving the bi-level optimization problem in (16) is computationally challenging and requires specific techniques like unrolling optimizer's steps, and **(ii)** extending the method to more realistic settings with richer energy structures is questionable due to computational complexity.

JKOnet* (Terpin et al., 2024) addresses the limitations of the original `JKOnet` by replacing the full JKO optimization problem (8) with its first-order optimality conditions. This reformulation allows the use of more expressive energy functionals (12). The method minimizes the following objective, utilizing a *precomputed* optimal transport plans π_k between ρ_k and ρ_{k+1}:

$$\begin{aligned}\mathcal{L}_{\texttt{JKOnet}^*}(\theta) = \sum_{k=0}^{K-1}\int_{\mathcal{X}\times\mathcal{X}} \Big\| \nabla V_{\theta_1}(x_{k+1}) + \int_{\mathcal{X}} \nabla U_{\theta_2}(x_{k+1}-y_{k+1})\,\mathrm{d}\rho_{k+1}(y_{k+1})\\ +\theta_3\frac{\nabla\rho_{k+1}(x_{k+1})}{\rho_{k+1}(x_{k+1})} + \frac{1}{\tau}(x_{k+1}-x_k)\Big\|^2 \,\mathrm{d}\pi_k(x_k,x_{k+1}).\end{aligned} \tag{17}$$

The authors emphasize several advantages over `JKOnet`, including reduced computational costs and support for more general energy functionals. However, their method requires an additional optimization round to precompute optimal transport plans π_k, introducing extra sources of inaccuracy and rendering the approach non-end-to-end. Moreover, the authors of (Terpin et al., 2024) employ discrete OT solvers to compute π_k, which may fail to accurately represent OT maps between the underlying distributions, particularly in high-dimensional settings (Deb et al., 2021).

iJKOnet (Ours) integrates the strengths of both `JKOnet` and `JKOnet`*. From `JKOnet`, it inherits parameterized models for the energy functional $\mathcal{J}_\theta$ and the transport map T_φ, avoiding reliance on

precomputed discrete optimal transport plans π_k, which can hinder flexible mappings between ρ_k and ρ_{k+1}. From JKOnet*, it adopts a more expressive formulation of the energy functional (12) – capable of capturing complex structures. Unlike JKOnet*, our formulation does not rely on analytical expressions of the JKO step (8) optimality conditions; instead, it solves the step directly via the inner minimization in (11). This approach makes the method more readily extendable to general energy forms, such as porous medium energies (Alvarez-Melis et al., 2022), although exploring such extensions is beyond the scope of the current work.

5 Experiments

Our implementation is primarily based on the publicly available code[1] from JKOnet* (Terpin et al., 2024), written in JAX (Bradbury et al., 2018), and will be made publicly available. Since our approach builds on the JKO framework, in §5.1, we first reproduce the experimental setup from JKOnet* for learning a known potential energy. Although we found the original codebase well-structured and accessible, we identified several inconsistencies in the data generation process, which we discuss in detail there. Finally, in §5.2, we compare our method against popular non-JKO baselines on single-cell RNA-seq datasets. Extended comparisons can be found in Appendix B.

JKO-based Models. As demonstrated in Appendix B.1.2 and B.1.3, learning the interaction and internal energy components directly from samples is challenging – likely because their estimation requires computing integrals over functions that must themselves be estimated. Consequently, for large-scale experiments, we restrict our model to the potential energy parametrization, imposing the inductive bias $\theta_2 = \theta_3 = 0$. Following (Terpin et al., 2024), we refer to this variant as $\texttt{iJKOnet}_V$. For moderate-dimensional tasks, we explore all combinations of parametrizations to illustrate the challenges of full energy recovery. For a fair comparison, we also implemented a time-varying potential energy parametrization, as described in Appendix B of the original paper, and denote this variant by $\texttt{iJKOnet}_{t,V}$. However, as discussed in Appendix A.2, this parametrization does not correspond to a single, consistent energy functional governing the dynamics. Instead, it actually solves a sequence of K independent optimal transport problems between consecutive snapshots (Bunne et al., 2023). Moreover, because JKOnet (Bunne et al., 2022b) is computationally expensive, we limit our comparisons to experiments with moderate dimensionality, with the corresponding results reported in Appendix B.2. Technical details of model training are provided in Appendix C.

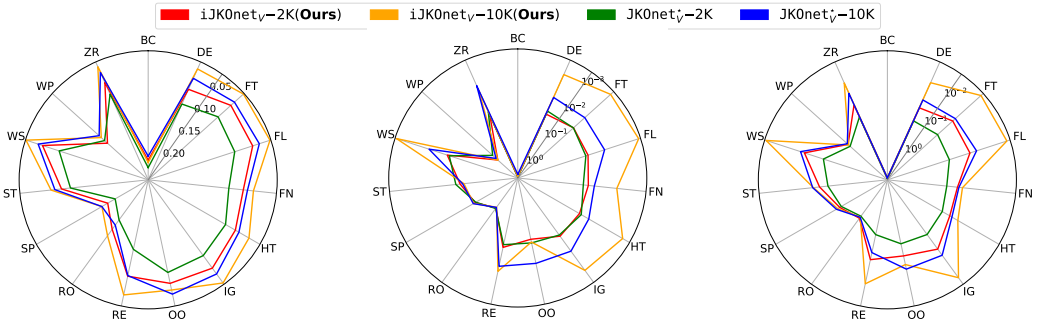
Metrics. We evaluate the next-step distribution $\hat{\rho}_{k+1}$ from ρ_k and compare it to the ground-truth distribution ρ_{k+1} using the following metrics: **(a)** Earth Mover's Distance (EMD) (or $\mathcal{W}_1$) (Rubner et al., 1998) and $\mathcal{W}_2$, **(b)** *Bures-Wasserstein Unexplained Variance Percentage* ($\text{Bd}^2_{\mathbb{W}_2}$-UVP) (Korotin et al., 2021c), **(c)** *$\mathcal{L}_2$-based Unexplained Variance Percentage* ($\mathcal{L}_2$-UVP) (Korotin et al., 2021a), which measures the discrepancy between the ground-truth functional F^* and its reconstruction $\widehat{F}$, and **(d)** *Maximum Mean Discrepancy* (MMD) (Gretton et al., 2012). Formal definitions and implementation details for all metrics are provided in Appendix C.1.

5.1 Learning potential energy

As discussed in §1, the population dynamics setting assumes that particle trajectories are not tracked across time – i.e., each time snapshot should consist of independently sampled particles without temporal correlation. This implies that data at each time step should be regenerated from scratch. However, we identified an inconsistency in the original codebase[1]: particle trajectories are preserved across time steps, resulting in temporally correlated samples. We refer to this as the **paired** setup, where each particle x_k is directly linked to x_{k+1} along a trajectory. In contrast, the intended, temporally uncorrelated setting is referred to as the **unpaired** setup. Table 3 in Appendix B.1.1 shows that switching between setup has a substantial impact on performance. A detailed discussion of this phenomenon is provided in Appendix A.3.

Following (Terpin et al., 2024, §4.1), we begin by examining how our method learns potentials in the 2D **paired** setup, which allows for a direct visual comparison with JKOnet*. These experiments are conducted on the synthetic dataset from (Terpin et al., 2024, Appendix B), with results presented in Figure 4 and further details provided in Appendix B.1.1. We then repeat the same procedure in the 2D **unpaired** setup, analyzing how performance scales with the number of samples (2K and 10K). The corresponding results are shown in Figure 2. They indicate that our method outperforms

[1] https://github.com/antonioterpin/jkonet-star



(a) EMD (b) B$d^2_{\mathbb{W}_2}$-UVP (c) $\mathcal{L}_2$-UVP

Figure 2: Numerical results from §5.1 for the **unpaired** setup. The reported absolute values show that, while increasing the number of samples generally improves performance across metrics, certain potentials remain challenging, highlighting the difficulty of this setup.

JKOnet* on nearly all potentials, and that for certain cases, increasing the number of samples does not improve performance, highlighting the difficulty of the unpaired setup.

5.2 Learning population dynamics for single-cell data

Following (Chen et al., 2023) and (Shen et al., 2025), in this section, we apply our method to modeling the dynamics of single-cell RNA data, with extended comparisons provided in Appendix B.2.

Dataset. We use the Embryoid Body (**EB**) dataset (Moon et al., 2019) and follow the preprocessing pipeline described in (Tong et al., 2020). The **EB** dataset comprises a collection of 5 timepoints sampled over a period of 30 days.

Non-JKO Models. We evaluate our method against classic algorithms for trajectory inference such as TrajectoryNet (Tong et al., 2020), MIOFLOW (Huguet et al., 2022), DMSB (Chen et al., 2023), NLSB (Koshizuka & Sato, 2023) and the recently proposed MMSB (Shen et al., 2025). For these methods, the reported metrics are taken from the corresponding referenced papers.

Leave-two-out in 5D. Following (Shen et al., 2025), we perform experiments using a leave-two-out setup. Since the **EB** dataset contains five timesteps, we remove the second ($\mathbf{t_2}$) and fourth ($\mathbf{t_4}$) timesteps from training data, and then evaluate how well our method can reconstruct the data from the remaining $\mathbf{t_1}$ and $\mathbf{t_3}$ timesteps. The results are demonstrated in Table 1. One can see that our method with the "t, V" parameterization yields the best metrics, whereas the use of alternative parameterizations (with interaction and internal energies) results in poorer/non-even metrics. Vanilla-SB corresponds to the IPF (GP) method (Vargas et al., 2021); further details are provided in Appendix B.6 of (Shen et al., 2025).

Table 1: **5D** experiment. $d_{\mathbb{W}_2}$ distance (↓) comparison across t_2 and t_4. Results for baselines (Shen et al., 2025).

Method	$\mathbf{t_2}$	$\mathbf{t_4}$
Vanilla-SB	1.49 ± 0.063	1.55 ± 0.034
DMSB	1.13 ± 0.082	1.45 ± 0.16
TrajectoryNet	2.03 ± 0.04	1.93 ± 0.08
MMSB	1.27 ± 0.028	1.57 ± 0.048
Static		
JKOnet$^*_V$	1.145 ± 0.033	2.529 ± 0.014
JKOnet$^*_{V+U}$	1.099 ± 0.119	2.537 ± 0.054
JKOnet$^*_{V+W}$	1.419 ± 0.173	2.510 ± 0.094
JKOnet$^*_{W+U}$	1.887 ± 0.017	1.739 ± 0.037
JKOnet*	1.361 ± 0.257	2.557 ± 0.042
Static (Ours)		
iJKOnet$_V$	1.082 ± 0.011	1.147 ± 0.001
iJKOnet$_{V+U}$	$\mathit{1.065 \pm 0.018}$	$\mathit{1.150 \pm 0.004}$
iJKOnet$_{V+W}$	2.865 ± 0.166	1.376 ± 0.015
iJKOnet$_{W+U}$	1.649 ± 0.005	0.868 ± 0.005
iJKOnet	3.577 ± 0.166	1.395 ± 0.032
Time-varying		
JKOnet$^*_{t,V}$	4.414 ± 1.499	2.771 ± 0.197
iJKOnet$_{t,V}$ **(Ours)**	$\mathbf{0.983 \pm 0.037}$	$\mathbf{0.849 \pm 0.021}$

Our method outperforms previous approaches both in terms of reconstruction metrics and execution time, as can be seen by comparing Table 12 in Appendix E of (Shen et al., 2025) with Table 7.

Leave-one-out in 100D. Following (Chen et al., 2023), we conduct experiments using a leave-one-out setup. One of the timesteps $\mathbf{t_2}$, $\mathbf{t_3}$, or $\mathbf{t_4}$ is omitted, and we evaluate the method's ability to reconstruct the distribution of the left-out timestep. The results are shown in Table 2. Our method achieves performance comparable to DMSB in the all-time (w/o LO) setting while using a simpler, simulation-free optimization routine that requires no trajectory caching. Consequently, it outperforms DMSB in execution time, as shown by comparing Table 12 in Appendix E of (Shen et al., 2025) (for 5D) with Table 7 (for 100D).

Table 2: **100D** experiment. MMD distance (↓). Comparison of methods across different leave-one-out splits. The results are averaged across 3 runs. Results for baseline methods: (Chen et al., 2023).

Method	LO-$\mathbf{t_2}$	LO-$\mathbf{t_3}$	LO-$\mathbf{t_4}$	w/o LO
NLSB (Koshizuka & Sato, 2023)	0.38	0.37	0.37	0.66
MIOFLOW (Huguet et al., 2022)	0.23	0.90	0.23	0.23
DMSB (Chen et al., 2023)	$\mathbf{0.042 \pm 0.020}$	$\mathbf{0.033 \pm 0.003}$	$\mathbf{0.040 \pm 0.020}$	$\mathbf{0.032 \pm 0.003}$
JKOnet$^\star_V$ (Terpin et al., 2024)	0.220 ± 0.025	0.293 ± 0.018	0.235 ± 0.006	0.229 ± 0.052
iJKOnet$_V$ **(Ours)**	$\mathit{0.137 \pm 0.001}$	$\mathit{0.123 \pm 0.001}$	0.097 ± 0.002	$\mathit{0.085 \pm 0.024}$
JKOnet$^\star_{t,V}$ (Terpin et al., 2024)	0.575 ± 0.119	0.619 ± 0.157	0.456 ± 0.056	0.477 ± 0.098
iJKOnet$_{t,V}$ **(Ours)**	0.848 ± 0.043	0.370 ± 0.233	$\mathit{0.055 \pm 0.007}$	0.124 ± 0.243

6 Discussion

Contributions. We introduce the novel `iJKOnet` method, which provides a general framework for recovering any type of energy functional governing the evolution of a system; in this work, we focus on the free energy parametrization. Our work bridges the gap between inverse optimization theory and computational gradient flows. The proposed method outperforms previous JKO-based approaches in all comparisons and achieves comparable results to non-JKO methods on real-world tasks. We also establish theoretical guarantees for recovering dynamics driven solely by potential energy.

Limitations. Our method can not recover other types of internal energy functionals except for the Entropy, nor can it handle time-dependent interaction energies, although theoretical results for such cases exist (Ferreira & Valencia-Guevara, 2018). Furthermore, it does not account for birth–death dynamics as in recent trajectory inference methods (Zhang et al., 2025b). The reliance on entropy estimation may also lead to performance degradation in higher-dimensional settings, and we find that learning interaction energies remains particularly challenging. As a result, jointly optimizing all energy parameters $(\theta_1, \theta_2, \theta_3)$ can cause instability and convergence to inaccurate potential estimates. Finally, from a theoretical standpoint, our analysis currently provides guarantees only for potential energy (Theorem (3.1)); extending it to other energy types is an avenue for future work.

LLM Usage. Large Language Models (LLMs) were used exclusively to assist with sentence rephrasing and improving text clarity. All scientific content, results, and interpretations in this paper were produced solely by the authors.

References

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. In Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining, pp. 2623–2631, 2019.

Vitalii Aksenov and Martin Eigel. An eulerian approach to regularized jko scheme with low-rank tensor decompositions for bayesian inversion. Journal of Scientific Computing, 105(1):3, 2025.

Michael Samuel Albergo and Eric Vanden-Eijnden. Building normalizing flows with stochastic interpolants. In The Eleventh International Conference on Learning Representations, 2023. URL https://openreview.net/forum?id=li7qeBbCR1t.

Fabian Altekrüger, Johannes Hertrich, and Gabriele Steidl. Neural Wasserstein gradient flows for discrepancies with riesz kernels. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), Proceedings of the 40th International Conference on Machine Learning, volume 202 of Proceedings of Machine Learning Research, pp. 664–690. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/altekruger23a.html.

David Alvarez-Melis, Yair Schiff, and Youssef Mroueh. Optimizing functionals on the space of probabilities with input convex neural networks. Transactions on Machine Learning Research, 2022. ISSN 2835-8856.

Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. Gradient flows: in metric spaces and in the space of probability measures. Springer Science & Business Media, 2008.

Brandon Amos, Lei Xu, and J Zico Kolter. Input convex neural networks. In International conference on machine learning, pp. 146–155. PMLR, 2017.

Francisco Andrade, Gabriel Peyré, and Clarice Poon. Learning from samples: Inverse problems over measures via sharpened fenchel-young losses. arXiv preprint arXiv:2505.07124, 2025.

Lazar Atanackovic, Xi Zhang, Brandon Amos, Mathieu Blanchette, Leo J Lee, Yoshua Bengio, Alexander Tong, and Kirill Neklyudov. Meta flow matching: Integrating vector fields on the wasserstein manifold. In The Thirteenth International Conference on Learning Representations, 2025. URL https://openreview.net/forum?id=9SYczU3Qgm.

Francisco J Ayala, Michael E Gilpin, and Joan G Ehrenfeld. Competition between species: theoretical models and experimental tests. Theoretical population biology, 4(3):331–356, 1973.

Jean Baccou and Jacques Liandrat. Subdivision scheme for discrete probability measure-valued data. Applied Mathematics Letters, 158:109233, 2024.

Rafael Bailo, Alethea Barbaro, Susana N Gomes, Konstantin Riedl, Tim Roith, Claudia Totzeck, and Urbain Vaes. Cbx: Python and julia packages for consensus-based interacting particle methods. Journal of Open Source Software, 9(98):6611, 2024a.

Rafael Bailo, José A Carrillo, and David Gómez-Castro. Aggregation-diffusion equations for collective behaviour in the sciences. arXiv preprint arXiv:2405.16679, 2024b.

Amartya Banerjee, Harlin Lee, Nir Sharon, and Caroline Moosmüller. Efficient trajectory inference in wasserstein space using consecutive averaging. In International Conference on Artificial Intelligence and Statistics, pp. 2260–2268. PMLR, 2025.

Arindam Banerjee, Srujana Merugu, Inderjit S Dhillon, and Joydeep Ghosh. Clustering with bregman divergences. Journal of machine learning research, 6(Oct):1705–1749, 2005.

Ricardo Baptista, Youssef Marzouk, and Olivier Zahm. On the representation and learning of monotone triangular transport maps. Foundations of Computational Mathematics, 24(6):2063–2108, 2024.

Jonathan T Barron. Continuously differentiable exponential linear units. arXiv preprint arXiv:1704.07483, 2017.

Nico Battich, Joep Beumer, Buys De Barbanson, Lenno Krenning, Chloé S Baron, Marvin E Tanenbaum, Hans Clevers, and Alexander Van Oudenaarden. Sequencing metabolically labeled transcripts in single cells reveals mrna turnover strategies. Science, 367(6482):1151–1156, 2020.

Amir Beck. First-Order Methods in Optimization, volume 25 of MOS-SIAM Series on Optimization. SIAM, Philadelphia, 2017. ISBN 9781611974980. doi: 10.1137/1.9781611974997.

Jean-David Benamou and Yann Brenier. A computational fluid mechanics solution to the monge-kantorovich mass transfer problem. Numerische Mathematik, 84(3):375–393, 2000.

Jean-David Benamou, Guillaume Carlier, Quentin Mérigot, and Edouard Oudet. Discretization of functionals involving the monge–ampère operator. Numerische mathematik, 134(3):611–636, 2016.

Jean-David Benamou, Thomas O Gallouët, and François-Xavier Vialard. Second-order models for optimal transport and cubic splines on the wasserstein space. Foundations of Computational Mathematics, 19(5):1113–1143, 2019.

Jules Berman and Benjamin Peherstorfer. CoLoRA: Continuous low-rank adaptation for reduced implicit neural modeling of parameterized partial differential equations. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), Proceedings of the 41st International Conference on Machine Learning, volume 235 of Proceedings of Machine Learning Research, pp. 3565–3583. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/berman24b.html.

Jules Berman, Tobias Blickhan, and Benjamin Peherstorfer. Parametric model reduction of mean-field and stochastic systems via higher-order action matching. Advances in Neural Information Processing Systems, 37:56588–56618, 2024.

Thomas B Berrett, Richard J Samworth, and Ming Yuan. Efficient multivariate entropy estimation via k-nearest neighbour distances. The Annals of Statistics, 47(1):288–318, 2019.

Vladimir I Bogachev and Aleksandr V Kolesnikov. The monge-kantorovich problem: achievements, connections, and perspectives. Russian Mathematical Surveys, 67(5):785, 2012.

Vladimir I Bogachev, Nicolai V Krylov, Michael Röckner, and Stanislav V Shaposhnikov. Fokker–Planck–Kolmogorov Equations, volume 207. American Mathematical Society, 2022.

Clément Bonet, Christophe Vauthier, and Anna Korba. Flowing datasets with wasserstein over wasserstein gradient flows. arXiv preprint arXiv:2506.07534, 2025.

James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Chris Leary, Dougal Maclaurin, George Necula, Adam Paszke, Jake VanderPlas, Skye Wanderman-Milne, and Qiao Zhang. JAX: composable transformations of Python+NumPy programs, 2018. URL http://github.com/jax-ml/jax.

Yann Brenier. Polar factorization and monotone rearrangement of vector-valued functions. Communications on pure and applied mathematics, 44(4):375–417, 1991.

Luc Brogat-Motte, Riccardo Bonalli, and Alessandro Rudi. Learning controlled stochastic differential equations. arXiv preprint arXiv:2411.01982, 2024.

Charlotte Bunne, Andreas Krause, and Marco Cuturi. Supervised training of conditional monge maps. Advances in Neural Information Processing Systems, 35:6859–6872, 2022a.

Charlotte Bunne, Laetitia Papaxanthos, Andreas Krause, and Marco Cuturi. Proximal optimal transport modeling of population dynamics. In International Conference on Artificial Intelligence and Statistics, pp. 6511–6528. PMLR, 2022b.

Charlotte Bunne, Stefan G Stark, Gabriele Gut, Jacobo Sarabia Del Castillo, Mitch Levesque, Kjong-Van Lehmann, Lucas Pelkmans, Andreas Krause, and Gunnar Rätsch. Learning single-cell perturbation responses using neural optimal transport. Nature methods, 20(11):1759–1768, 2023.

Daniel Burkhardt, Jonathan Bloom, Robrecht Cannoodt, Malte D Luecken, Smita Krishnaswamy, Christopher Lance, Angela O Pisco, and Fabian J Theis. Multimodal single-cell integration across time, individuals, and batches. NeurIPS Competitions, pp. 12, 2022.

Ivan Butakov, Aleksandr Tolmachev, Sofia Malanchuk, Anna Neopryatnaya, and Alexey Frolov. Mutual information estimation via normalizing flows. Advances in Neural Information Processing Systems, 37:3027–3057, 2024.

John Charles Butcher. Numerical methods for ordinary differential equations. John Wiley & Sons, 2016.

Martin Campos Pinto, José A Carrillo, Frédérique Charles, and Young-Pil Choi. Convergence of a linearly transformed particle method for aggregation equations. Numerische Mathematik, 139: 743–793, 2018.

Clément Cancès, Maxime Herda, and Annamaria Massimini. Finite volumes for a generalized poisson-nernst-planck system with cross-diffusion and size exclusion. In International Conference on Finite Volumes for Complex Applications, pp. 57–73. Springer, 2023.

José A Carrillo and Rishabh S Gvalani. Phase transitions for nonlinear nonlocal aggregation-diffusion equations. Communications in mathematical physics, 382(1):485–545, 2021.

José A Carrillo, Robert J McCann, and Cédric Villani. Kinetic equilibration rates for granular media and related equations: entropy dissipation and mass transportation estimates. Revista Matematica Iberoamericana, 19(3):971–1018, 2003.

José A Carrillo, Katy Craig, and Yao Yao. Aggregation-diffusion equations: dynamics, asymptotics, and singular limits. Active Particles, Volume 2: Advances in Theory, Models, and Applications, pp. 65–108, 2019a.

José A Carrillo, Katy Craig, Li Wang, and Chaozhen Wei. Primal dual methods for wasserstein gradient flows. Foundations of Computational Mathematics, pp. 1–55, 2022.

José Antonio Carrillo, Katy Craig, and Francesco S Patacchini. A blob method for diffusion. Calculus of Variations and Partial Differential Equations, 58:1–53, 2019b.

Timothy CY Chan, Rafid Mahmood, and Ian Yihang Zhu. Inverse optimization: Theory and applications. Operations Research, 73(2):1046–1074, 2025.

Ricky T. Q. Chen and Yaron Lipman. Flow matching on general geometries. In The Twelfth International Conference on Learning Representations, 2024. URL https://openreview.net/forum?id=g7ohDlTITL.

Ricky TQ Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. Advances in neural information processing systems, 31, 2018a.

Tianrong Chen, Guan-Horng Liu, and Evangelos Theodorou. Likelihood training of schrödinger bridge using forward-backward SDEs theory. In International Conference on Learning Representations, 2022a. URL https://openreview.net/forum?id=nioAdKCEdXB.

Tianrong Chen, Guan-Horng Liu, Molei Tao, and Evangelos Theodorou. Deep momentum multi-marginal schrödinger bridge. Advances in Neural Information Processing Systems, 36:57058–57086, 2023.

Yongxin Chen, Giovanni Conforti, and Tryphon T Georgiou. Measure-valued spline curves: An optimal transport viewpoint. SIAM Journal on Mathematical Analysis, 50(6):5947–5968, 2018b.

Yongxin Chen, Tryphon T Georgiou, and Michele Pavon. The most likely evolution of diffusing and vanishing particles: Schrodinger bridges with unbalanced marginals. SIAM Journal on Control and Optimization, 60(4):2016–2039, 2022b.

Xiuyuan Cheng, Jianfeng Lu, Yixin Tan, and Yao Xie. Convergence of flow-based generative models via proximal gradient descent in wasserstein space. IEEE Transactions on Information Theory, 2024.

Ziheng Cheng, Shiyue Zhang, Longlin Yu, and Cheng Zhang. Particle-based variational inference with generalized wasserstein gradient flow. Advances in Neural Information Processing Systems, 36:66583–66609, 2023.

Andrei Chertkov and Ivan Oseledets. Solution of the fokker–planck equation by cross approximation method in the tensor train format. Frontiers in Artificial Intelligence, 4:668215, 2021.

Sinho Chewi, Julien Clancy, Thibaut Le Gouic, Philippe Rigollet, George Stepaniants, and Austin Stromme. Fast and smooth interpolation on wasserstein space. In International Conference on Artificial Intelligence and Statistics, pp. 3061–3069. PMLR, 2021.

Sinho Chewi, Jonathan Niles-Weed, and Philippe Rigollet. Statistical Optimal Transport: École d'Été de Probabilités de Saint-Flour XLIX – 2019, volume 2364 of Lecture Notes in Mathematics. Springer Cham, 2025. ISBN 978-3-031-85159-9. doi: 10.1007/978-3-031-85160-5.

Lenaic Chizat, Gabriel Peyré, Bernhard Schmitzer, and François-Xavier Vialard. An interpolating distance between optimal transport and fisher–rao metrics. Foundations of Computational Mathematics, 18(1):1–44, 2018a.

Lenaic Chizat, Gabriel Peyré, Bernhard Schmitzer, and François-Xavier Vialard. Unbalanced optimal transport: Dynamic and kantorovich formulations. Journal of Functional Analysis, 274(11): 3090–3123, 2018b.

Lénaïc Chizat, Maria Colombo, Xavier Fernández-Real, and Alessio Figalli. Infinite-width limit of deep linear neural networks. Communications on Pure and Applied Mathematics, 77(10): 3958–4007, 2024.

Jaemoo Choi, Jaewoong Choi, and Myungjoo Kang. Scalable wasserstein gradient flow for generative modeling through unbalanced optimal transport. In International Conference on Machine Learning, pp. 8629–8650. PMLR, 2024.

Julien Clancy and Felipe Suarez. Wasserstein-fisher-rao splines. arXiv e-prints, pp. arXiv–2203, 2022.

Katy Craig and Andrea Bertozzi. A blob method for the aggregation equation. Mathematics of computation, 85(300):1681–1717, 2016.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. Advances in neural information processing systems, 26, 2013.

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 preprint arXiv:2201.12324, 2022.

Aniket Das and Dheeraj Nagaraj. Provably fast finite particle variants of svgd via virtual particle stochastic approximation. Advances in Neural Information Processing Systems, 36:49748–49760, 2023.

Constantinos Daskalakis, Andrew Ilyas, Vasilis Syrgkanis, and Haoyang Zeng. Training gans with optimism. In International Conference on Learning Representations, 2018.

Valentin De Bortoli, James Thornton, Jeremy Heng, and Arnaud Doucet. Diffusion schrödinger bridge with applications to score-based generative modeling. Advances in neural information processing systems, 34:17695–17709, 2021.

Nabarun Deb, Promit Ghosal, and Bodhisattva Sen. Rates of estimation of optimal transport maps using plug-in estimators via barycentric projections. Advances in Neural Information Processing Systems, 34:29736–29753, 2021.

DeepMind, Igor Babuschkin, Kate Baumli, Alison Bell, Surya Bhupatiraju, Jake Bruce, Peter Buchlovsky, David Budden, Trevor Cai, Aidan Clark, Ivo Danihelka, Antoine Dedieu, Claudio Fantacci, Jonathan Godwin, Chris Jones, Ross Hemsley, Tom Hennigan, Matteo Hessel, Shaobo Hou, Steven Kapturowski, Thomas Keck, Iurii Kemaev, Michael King, Markus Kunesch, Lena Martens, Hamza Merzic, Vladimir Mikulik, Tamara Norman, George Papamakarios, John Quan, Roman Ring, Francisco Ruiz, Alvaro Sanchez, Laurent Sartran, Rosalia Schneider, Eren Sezener, Stephen Spencer, Srivatsan Srinivasan, Miloš Stanojević, Wojciech Stokowiec, Luyu Wang, Guangyao Zhou, and Fabio Viola. The DeepMind JAX Ecosystem, 2020. URL `http://github.com/google-deepmind`.

Michael Ziyang Diao, Krishna Balasubramanian, Sinho Chewi, and Adil Salim. Forward-backward gaussian variational inference via jko in the bures-wasserstein space. In International Conference on Machine Learning, pp. 7960–7991. PMLR, 2023.

Nira Dyn and Nir Sharon. Subdivision schemes in metric spaces. arXiv preprint arXiv:2509.08070, 2025.

Rossin Erbe, Genevieve Stein-O'Brien, and Elana J Fertig. Transcriptomic forecasting with neural ordinary differential equations. Patterns, 4(8), 2023.

Lawrence C Evans. Partial differential equations, volume 19. American Mathematical Society, 2022.

Jiaojiao Fan, Qinsheng Zhang, Amirhossein Taghvaei, and Yongxin Chen. Variational wasserstein gradient flow. In International Conference on Machine Learning, pp. 6185–6215. PMLR, 2022.

Jiaojiao Fan, Shu Liu, Shaojun Ma, Hao-Min Zhou, and Yongxin Chen. Neural monge map estimation and its applications. Transactions on Machine Learning Research, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=2mZSlQscj3. Featured Certification.

Lucas CF Ferreira and Julio C Valencia-Guevara. Gradient flows of time-dependent functionals in metric spaces and applications to pdes. Monatshefte für Mathematik, 185(2):231–268, 2018.

Davide Fiaschi and Cristiano Ricci. The spatial evolution of economic activities and the emergence of cities. Journal of Economic Theory, pp. 105971, 2025.

Alessio Figalli and Federico Glaudo. An Invitation to Optimal Transport, Wasserstein Distances, and Gradient Flows. EMS Press, Berlin, 2nd edition, 2023. ISBN 978-3-98547-050-1. doi: 10.4171/ETB/25. URL https://ems.press/books/etb/258.

Chris Finlay, Jörn-Henrik Jacobsen, Levon Nurbekyan, and Adam Oberman. How to train your neural ode: the world of jacobian and kinetic regularization. In International conference on machine learning, pp. 3154–3164. PMLR, 2020.

Mike Fisher, Jorge Nocedal, Yannick Trémolet, and Stephen J Wright. Data assimilation in weather forecasting: a case study in pde-constrained optimization. Optimization and Engineering, 10(3): 409–426, 2009.

Benyamin Ghojogh, Ali Ghodsi, Fakhri Karray, and Mark Crowley. Reproducing kernel hilbert space, mercer's theorem, eigenfunctions, nystr\" om method, and use of kernels in machine learning: Tutorial and survey. arXiv preprint arXiv:2106.08443, 2021.

David Gómez-Castro. Beginner's guide to aggregation-diffusion equations. SeMA Journal, 81(4): 531–587, 2024.

V Gontis, J Ruseckas, and A Kononovičius. A long-range memory stochastic model of the return in financial markets. Physica A: Statistical Mechanics and its Applications, 389(1):100–106, 2010.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. Communications of the ACM, 63(11):139–144, 2020.

Erhard Gramer. Probability measures with given marginals and conditionals: I-projections and conditional iterative proportional fitting. Statistics & Risk Modeling, 18(3):311–330, 2000.

Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. The journal of machine learning research, 13(1):723–773, 2012.

Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. Advances in neural information processing systems, 30, 2017.

Fuqun Han, Stanley Osher, and Wuchen Li. Tensor train based sampling algorithms for approximating regularized wasserstein proximal operators. SIAM/ASA Journal on Uncertainty Quantification, 13(2):775–804, 2025.

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

Doron Haviv, Aram-Alexandre Pooladian, Dana Pe'er, and Brandon Amos. Wasserstein flow matching: Generative modeling over families of distributions. In Forty-second International Conference on Machine Learning, 2025. URL https://openreview.net/forum?id=MRmI68k3gd.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778, 2016.

Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). arXiv preprint arXiv:1606.08415, 2016.

Wanli Hong, Yuliang Shi, and Jonathan Niles-Weed. Trajectory inference with smooth schrödinger bridges. In Forty-second International Conference on Machine Learning, 2025. URL https://openreview.net/forum?id=GHyvvWu1XC.

Chin-Wei Huang, Ricky T. Q. Chen, Christos Tsirigotis, and Aaron Courville. Convex potential flows: Universal probability distributions with optimal transport and convex optimization. In International Conference on Learning Representations, 2021.

Guillaume Huguet, Daniel Sumner Magruder, Alexander Tong, Oluwadamilola Fasina, Manik Kuchroo, Guy Wolf, and Smita Krishnaswamy. Manifold interpolating optimal-transport flows for trajectory inference. Advances in neural information processing systems, 35:29705–29718, 2022.

Michael F Hutchinson. A stochastic estimator of the trace of the influence matrix for laplacian smoothing splines. Communications in Statistics-Simulation and Computation, 18(3):1059–1076, 1989.

Jan-Christian Hütter and Philippe Rigollet. Minimax estimation of smooth optimal transport maps. The Annals of Statistics, 49(2):1166 – 1194, 2021. doi: 10.1214/20-AOS1997. URL https://doi.org/10.1214/20-AOS1997.

Haoming Jiang, Zhehui Chen, Minshuo Chen, Feng Liu, Dingding Wang, and Tuo Zhao. On computation and generalization of gans with spectrum control. arXiv preprint arXiv:1812.10912, 2018.

Richard Jordan, David Kinderlehrer, and Felix Otto. The variational formulation of the fokker–planck equation. SIAM journal on mathematical analysis, 29(1):1–17, 1998.

Ansgar Jüngel, Stefan Portisch, and Antoine Zurek. A convergent finite-volume scheme for nonlocal cross-diffusion systems for multi-species populations. ESAIM: Mathematical Modelling and Numerical Analysis, 58(2):759–792, 2024.

Jorge Justiniano, Martin Rumpf, and Matthias Erbar. Approximation of splines in wasserstein spaces. ESAIM: Control, Optimisation and Calculus of Variations, 30:64, 2024.

Leonid V Kantorovich. On the translocation of masses. In Dokl. Akad. Nauk. USSR (NS), volume 37, pp. 199–201, 1942.

Kacper Kapusniak, Peter Potaptchik, Teodora Reu, Leo Zhang, Alexander Tong, Michael Bronstein, Joey Bose, and Francesco Di Giovanni. Metric flow matching for smooth interpolations on the data manifold. Advances in Neural Information Processing Systems, 37:135011–135042, 2024.

Keisuke Kawano, Takuro Kutsuna, Naoki Hayashi, Yasushi Esaki, and Hidenori Tanaka. Ct-ot flow: Estimating continuous-time dynamics from discrete temporal snapshots. arXiv preprint arXiv:2505.17354, 2025.

Evelyn F Keller and Lee A Segel. Model for chemotaxis. Journal of theoretical biology, 30(2): 225–234, 1971.

Diederik P Kingma. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014.

Dominik Klein, Théo Uscidda, Fabian Theis, and Marco Cuturi. Genot: Entropic (gromov) wasserstein flow matching with applications to single-cell genomics. Advances in Neural Information Processing Systems, 37:103897–103944, 2024.

Nikita Kornilov, Petr Mokrov, Alexander Gasnikov, and Aleksandr Korotin. Optimal flow matching: Learning straight trajectories in just one step. Advances in Neural Information Processing Systems, 37:104180–104204, 2024.

Alexander Korotin, Vage Egiazarian, Arip Asadulaev, Alexander Safin, and Evgeny Burnaev. Wasserstein-2 generative networks. In International Conference on Learning Representations, 2021a. URL https://openreview.net/forum?id=bEoxzW_EXsa.

Alexander Korotin, Lingxiao Li, Aude Genevay, Justin M Solomon, Alexander Filippov, and Evgeny Burnaev. Do neural optimal transport solvers work? a continuous wasserstein-2 benchmark. Advances in neural information processing systems, 34:14593–14605, 2021b.

Alexander Korotin, Lingxiao Li, Justin Solomon, and Evgeny Burnaev. Continuous wasserstein-2 barycenter estimation without minimax optimization. In International Conference on Learning Representations, 2021c.

Alexander Korotin, Daniil Selikhanovych, and Evgeny Burnaev. Neural optimal transport. In The Eleventh International Conference on Learning Representations, 2023. URL https://openreview.net/forum?id=d8CBRlWNkqH.

Takeshi Koshizuka and Issei Sato. Neural lagrangian schr\"{o}dinger bridge: Diffusion modeling for population dynamics. In The Eleventh International Conference on Learning Representations, 2023.

Chrysoula Kosma, Giannis Nikolentzos, George Panagopoulos, Jean-Marc Steyaert, and Michalis Vazirgiannis. Neural ordinary differential equations for modeling epidemic spreading. Transactions on Machine Learning Research, 2023.

Leonenko Kozachenko. Sample estimate of the entropy of a random vector. Probl. Pered. Inform., 23:9, 1987.

Rahul Krishnan, Uri Shalit, and David Sontag. Structured inference networks for nonlinear state space models. In Proceedings of the AAAI conference on artificial intelligence, volume 31, 2017.

Marc Lambert, Sinho Chewi, Francis Bach, Silvère Bonnabel, and Philippe Rigollet. Variational inference via wasserstein gradient flows. Advances in Neural Information Processing Systems, 35:14434–14447, 2022.

Jeffrey M Lane and Richard F Riesenfeld. A theoretical development for the computer generation and display of piecewise polynomial surfaces. IEEE Transactions on Pattern Analysis and Machine Intelligence, (1):35–46, 1980.

Hugo Lavenant, Stephen Zhang, Young-Heon Kim, Geoffrey Schiebinger, et al. Toward a mathematical theory of trajectory inference. The Annals of Applied Probability, 34(1A):428–500, 2024.

Wonjun Lee, Li Wang, and Wuchen Li. Deep jko: time-implicit particle methods for general nonlinear gradient flows. Journal of Computational Physics, 514:113187, 2024.

Ruikun Li, Jiazhen Liu, Huandong Wang, Qingmin Liao, and Yong Li. Weightflow: Learning stochastic dynamics via evolving weight of neural network. arXiv preprint arXiv:2508.00451, 2025a.

Xuechen Li, Ting-Kam Leonard Wong, Ricky TQ Chen, and David Duvenaud. Scalable gradients for stochastic differential equations. In International Conference on Artificial Intelligence and Statistics, pp. 3870–3882. PMLR, 2020.

Zihao Li, Zhichen Zeng, Xiao Lin, Feihao Fang, Yanru Qu, Zhe Xu, Zhining Liu, Xuying Ning, Tianxin Wei, Ge Liu, et al. Flow matching meets biology and life science: A survey. arXiv e-prints, pp. arXiv–2507, 2025b.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In 11th International Conference on Learning Representations, ICLR 2023, 2023.

Guan-Horng Liu, Yaron Lipman, Maximilian Nickel, Brian Karrer, Evangelos Theodorou, and Ricky T. Q. Chen. Generalized schrödinger bridge matching. In The Twelfth International Conference on Learning Representations, 2024. URL https://openreview.net/forum?id=SoismgeX7z.

Qiang Liu and Dilin Wang. Stein variational gradient descent: A general purpose bayesian inference algorithm. Advances in neural information processing systems, 29, 2016.

Xingchao Liu and Lemeng Wu. Learning diffusion bridges on constrained domains. In international conference on learning representations (ICLR), 2023.

Xingchao Liu, Chengyue Gong, and qiang liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In The Eleventh International Conference on Learning Representations, 2023. URL https://openreview.net/forum?id=XVjTT1nw5z.

Zewen Liu, Xiaoda Wang, Bohan Wang, Zijie Huang, Carl Yang, and Wei Jin. Graph odes and beyond: A comprehensive survey on integrating differential equations with graph neural networks. In Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2, pp. 6118–6128, 2025.

Evan Z Macosko, Anindita Basu, Rahul Satija, James Nemesh, Karthik Shekhar, Melissa Goldman, Itay Tirosh, Allison R Bialas, Nolan Kamitaki, Emily M Martersteck, et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. Cell, 161(5):1202–1214, 2015.

Bertrand Maury, Aude Roudneff-Chupin, and Filippo Santambrogio. A macroscopic crowd motion model of gradient flow type. Mathematical Models and Methods in Applied Sciences, 20(10): 1787–1821, 2010.

Bertrand Maury, Aude Roudneff-Chupin, Filippo Santambrogio, and Juliette Venel. Handling congestion in crowd motion modeling. Networks and Heterogeneous Media, 6(3):485–519, 2011.

Robert J McCann. Existence and uniqueness of monotone measure-preserving maps. Duke Mathematical Journal, 80(2):309, 1995.

Boris Medina-Salgado, Eddy Sánchez-DelaCruz, Pilar Pozos-Parra, and Javier E Sierra. Urban traffic flow prediction techniques: A review. Sustainable Computing: Informatics and Systems, 35:100739, 2022.

Nicholas Metropolis and Stanislaw Ulam. The monte carlo method. Journal of the American statistical association, 44(247):335–341, 1949.

Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. In International Conference on Learning Representations, 2018. URL https://openreview.net/forum?id=B1QRgziT-.

Petr Mokrov, Alexander Korotin, Lingxiao Li, Aude Genevay, Justin M Solomon, and Evgeny Burnaev. Large-scale wasserstein gradient flows. Advances in Neural Information Processing Systems, 34:15243–15256, 2021.

Gaspard Monge. Mémoire sur la théorie des déblais et des remblais. Mem. Math. Phys. Acad. Royale Sci., pp. 666–704, 1781.

Kevin R Moon, David Van Dijk, Zheng Wang, Scott Gigante, Daniel B Burkhardt, William S Chen, Kristina Yim, Antonia van den Elzen, Matthew J Hirn, Ronald R Coifman, et al. Visualizing structure and transitions in high-dimensional biological data. Nature biotechnology, 37(12):1482–1492, 2019.

Alex Morehead, Lazar Atanackovic, Akshata Hegde, Yanli Wang, Frimpong Boadu, Joel Selvaraj, Alexander Tong, Aditi Krishnapriyan, and Jianlin Cheng. How to go with the flow: flow matching in bioinformatics and computational biology. Authorea Preprints, August 2025. doi: 10.22541/au.175382408.89466370/v2. URL http://dx.doi.org/10.22541/au.175382408.89466370/v2.

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

Kirill Neklyudov, Rob Brekelmans, Alexander Tong, Lazar Atanackovic, Qiang Liu, and Alireza Makhzani. A computational framework for solving wasserstein lagrangian flows. In Proceedings of the 41st International Conference on Machine Learning, pp. 37461–37485, 2024a.

Kirill Neklyudov, Rob Brekelmans, Alexander Tong, Lazar Atanackovic, qiang liu, and Alireza Makhzani. A computational framework for solving wasserstein lagrangian flows. In Forty-first International Conference on Machine Learning, 2024b. URL https://openreview.net/forum?id=wwItuHdus6.

Atsushi Nitanda. Improved particle approximation error for mean field neural networks. In The Thirty-eighth Annual Conference on Neural Information Processing Systems, 2024. URL https://openreview.net/forum?id=76NKidadct.

Ivan V Oseledets. Tensor-train decomposition. SIAM Journal on Scientific Computing, 33(5): 2295–2317, 2011.

F Otto. The geometry of dissipative evolution equations: the porous medium equation. Comm. Partial Differential Equations, 26:101–174, 2001.

Neal Parikh, Stephen Boyd, et al. Proximal algorithms. Foundations and trends® in Optimization, 1(3):127–239, 2014.

Min Sue Park, Cheolhyeong Kim, Hwijae Son, and Hyung Ju Hwang. The deep minimizing movement scheme. Journal of Computational Physics, 494:112518, 2023.

A Paszke. Pytorch: An imperative style, high-performance deep learning library. arXiv preprint arXiv:1912.01703, 2019.

Stefano Peluchetti. Diffusion bridge mixture transports, schrödinger bridge problems and generative modeling. Journal of Machine Learning Research, 24(374):1–51, 2023a.

Stefano Peluchetti. Non-denoising forward-time diffusions. arXiv preprint arXiv:2312.14589, 2023b.

Gabriel Peyré, Marco Cuturi, et al. Computational optimal transport: With applications to data science. Foundations and Trends® in Machine Learning, 11(5-6):355–607, 2019.

Simon Plazotta and Jonathan Zinsl. High-frequency limit of non-autonomous gradient flows of functionals with time-periodic forcing. Journal of Differential Equations, 261(12):6806–6855, 2016.

Jonathan R Potts. Aggregation-diffusion in heterogeneous environments. arXiv preprint arXiv:2409.10147, 2024.

Jonathan R Potts and Kevin J Painter. Distinguishing between long-transient and asymptotic states in a biological aggregation model. Bulletin of Mathematical Biology, 86(3):28, 2024.

Ilan Price, Alvaro Sanchez-Gonzalez, Ferran Alet, Tom R Andersson, Andrew El-Kadi, Dominic Masters, Timo Ewalds, Jacklynn Stott, Shakir Mohamed, Peter Battaglia, et al. Probabilistic weather forecasting with machine learning. Nature, 637(8044):84–90, 2025.

Hannes Risken and Till Frank. The Fokker-Planck Equation: Methods of Solution and Applications, volume 18. Springer Science & Business Media, 1996.

R. Tyrrell Rockafellar. Convex analysis. Princeton Mathematical Series. Princeton University Press, Princeton, N. J., 1970.

Martin Rohbeck, Charlotte Bunne, Edward De Brouwer, Jan-Christian Huetter, Anne Biton, Kelvin Y. Chen, Aviv Regev, and Romain Lopez. Modeling complex system dynamics with flow matching across time and conditions. In The Thirteenth International Conference on Learning Representations, 2025. URL https://openreview.net/forum?id=hwnObmOTrV.

Litu Rout, Alexander Korotin, and Evgeny Burnaev. Generative modeling with optimal transport maps. In Proceedings of the 10th International Conference on Learning Representations (ICLR), 2022. URL https://openreview.net/forum?id=5JdLZg346Lw.

Yossi Rubner, Carlo Tomasi, and Leonidas J Guibas. A metric for distributions with applications to image databases. In Sixth international conference on computer vision (IEEE Cat. No. 98CH36271), pp. 59–66. IEEE, 1998.

Adil Salim, Anna Korba, and Giulia Luise. The wasserstein proximal gradient algorithm. Advances in Neural Information Processing Systems, 33:12356–12366, 2020.

Filippo Santambrogio. Optimal transport for applied mathematicians, volume 87. Springer, 2015.

Filippo Santambrogio. {Euclidean, metric, and Wasserstein} gradient flows: an overview. Bulletin of Mathematical Sciences, 7:87–154, 2017.

Daniel Saragih, Deyu Cao, and Tejas Balaji. Flows and diffusions on the neural manifold. arXiv e-prints, pp. arXiv–2507, 2025.

Simo Särkkä and Arno Solin. Applied stochastic differential equations, volume 10. Cambridge University Press, 2019.

Christopher Scarvelis and Justin Solomon. Riemannian metric learning via optimal transport. In The Eleventh International Conference on Learning Representations, 2023. URL https://openreview.net/forum?id=v3y68gz-WEz.

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

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

Yunyi Shen, Renato Berlinghieri, and Tamara Broderick. Multi-marginal schrödinger bridges with iterative reference refinement. In The 28th International Conference on Artificial Intelligence and Statistics, 2025. URL https://openreview.net/forum?id=VcwZ3gtYFY.

Yuyang Shi, Valentin De Bortoli, Andrew Campbell, and Arnaud Doucet. Diffusion schrödinger bridge matching. Advances in Neural Information Processing Systems, 36:62183–62223, 2023.

Fabio Sigrist, Hans R Künsch, and Werner A Stahel. Stochastic partial differential equation based modelling of large space–time data sets. Journal of the Royal Statistical Society Series B: Statistical Methodology, 77(1):3–33, 2015.

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

Yuhao Sun, Zhenyi Zhang, Zihan Wang, Tiejun Li, and Peijie Zhou. Variational regularized unbalanced optimal transport: Single network, least action. arXiv preprint arXiv:2505.11823, 2025.

Taiji Suzuki, Atsushi Nitanda, and Denny Wu. Uniform-in-time propagation of chaos for the mean-field gradient langevin dynamics. In The Eleventh International Conference on Learning Representations, 2023. URL https://openreview.net/forum?id=_JScUk9TBUn.

Chandan Tankala, Dheeraj M Nagaraj, and Anant Raj. Beyond propagation of chaos: A stochastic algorithm for mean field optimization. arXiv preprint arXiv:2503.13115, 2025.

Antonio Terpin, Nicolas Lanzetti, Martín Gadea, and Florian Dorfler. Learning diffusion at light-speed. Advances in Neural Information Processing Systems, 37:6797–6832, 2024.

Gerald Teschl. Ordinary differential equations and dynamical systems, volume 140. American Mathematical Soc., 2012.

Panagiotis Theodoropoulos, Augustinos D Saravanos, Evangelos A Theodorou, and Guan-Horng Liu. Momentum multi-marginal schr\" odinger bridge matching. arXiv preprint arXiv:2506.10168, 2025.

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

Alexander Tong, Manik Kuchroo, Shabarni Gupta, Aarthi Venkat, Beatriz P San Juan, Laura Rangel, Brandon Zhu, John G Lock, Christine L Chaffer, and Smita Krishnaswamy. Learning transcriptional and regulatory dynamics driving cancer cell plasticity using neural ode-based optimal transport. bioRxiv, pp. 2023–03, 2023.

Alexander Tong, Kilian Fatras, Nikolay Malkin, Guillaume Huguet, Yanlei Zhang, Jarrid Rector-Brooks, Guy Wolf, and Yoshua Bengio. Improving and generalizing flow-based generative models with minibatch optimal transport. Transactions on Machine Learning Research, pp. 1–34, 2024a.

Alexander Tong, Nikolay Malkin, Kilian Fatras, Lazar Atanackovic, Yanlei Zhang, Guillaume Huguet, Guy Wolf, and Yoshua Bengio. Simulation-free schrödinger bridges via score and flow matching. In The 27th International Conference on Artificial Intelligence and Statistics, pp. 1279–1287. Journal of Machine Learning Research-Proceedings Track, 2024b.

Théo Uscidda and Marco Cuturi. The monge gap: A regularizer to learn all transport maps. In International Conference on Machine Learning, pp. 34709–34733. PMLR, 2023.

Francisco Vargas, Pierre Thodoroff, Austen Lamacraft, and Neil Lawrence. Solving schrödinger bridges via maximum likelihood. Entropy, 23(9):1134, 2021.

Juan Luis Vázquez. Asymptotic behaviour methods for the heat equation. convergence to the gaussian. arXiv preprint arXiv:1706.10034, 2017.

Yogesh Verma, Markus Heinonen, and Vikas Garg. ClimODE: Climate and weather forecasting with physics-informed neural ODEs. In The Twelfth International Conference on Learning Representations, 2024. URL https://openreview.net/forum?id=xuY33XhEGR.

Alexander Vidal, Samy Wu Fung, Luis Tenorio, Stanley Osher, and Levon Nurbekyan. Taming hyperparameter tuning in continuous normalizing flows using the jko scheme. Scientific reports, 13(1):4501, 2023.

Cédric Villani et al. Optimal transport: old and new, volume 338. Springer, 2008.

Wei Wan, Yuejin Zhang, Chenglong Bao, Bin Dong, and Zuoqiang Shi. A scalable deep learning approach for solving high-dimensional dynamic optimal transport. SIAM Journal on Scientific Computing, 2023.

Dongyi Wang, Yuanwei Jiang, Zhenyi Zhang, Xiang Gu, Peijie Zhou, and Jian Sun. Joint velocity-growth flow matching for single-cell dynamics modeling. arXiv preprint arXiv:2505.13413, 2025.

Xiaomei Wang, Qi An, Zilong He, and Wei Fang. A literature review of social network analysis in epidemic prevention and control. Complexity, 2021(1):3816221, 2021.

E Weinan, Tiejun Li, and Eric Vanden-Eijnden. Applied stochastic analysis, volume 199. American Mathematical Soc., 2021.

Shang Wu and Yazhen Wang. Computational and statistical asymptotic analysis of the jko scheme for iterative algorithms to update distributions. arXiv preprint arXiv:2501.06408, 2025.

Chen Xu, Xiuyuan Cheng, and Yao Xie. Normalizing flow neural networks by jko scheme. Advances in Neural Information Processing Systems, 36:47379–47405, 2023.

Luxuan Yang, Ting Gao, Yubin Lu, Jinqiao Duan, and Tao Liu. Neural network stochastic differential equation models with applications to financial data forecasting. Applied Mathematical Modelling, 115:279–299, 2023.

Stephen Zhang, Suryanarayana Maddu, Xiaojie Qiu, and Victor Chardès. Inferring stochastic dynamics with growth from cross-sectional data. arXiv preprint arXiv:2505.13197, 2025a.

Yanbo Zhang and Michael Levin. Equilibrium flow: From snapshots to dynamics. arXiv preprint arXiv:2509.17990, 2025.

Zhenyi Zhang, Tiejun Li, and Peijie Zhou. Learning stochastic dynamics from snapshots through regularized unbalanced optimal transport. In The Thirteenth International Conference on Learning Representations, 2025b. URL https://openreview.net/forum?id=gQlxd3Mtru.

Zhenyi Zhang, Yuhao Sun, Qiangwei Peng, Tiejun Li, and Peijie Zhou. Integrating dynamical systems modeling with spatiotemporal scrna-seq data analysis. Entropy, 27(5):453, 2025c.

Zhenyi Zhang, Zihan Wang, Yuhao Sun, Tiejun Li, and Peijie Zhou. Modeling cell dynamics and interactions with unbalanced mean field schr\" odinger bridge. arXiv preprint arXiv:2505.11197, 2025d.

Zhenyi Zhang, Zihan Wang, Yuhao Sun, Jiantao Shen, Qiangwei Peng, Tiejun Li, and Peijie Zhou. Deciphering cell-fate trajectories using spatiotemporal single-cell transcriptomic data. Authorea Preprints, September 2025e. doi: 10.22541/au.175735334.45794622/v1. URL http://dx.doi.org/10.22541/au.175735334.45794622/v1.

Qinqing Zheng, Matt Le, Neta Shaul, Yaron Lipman, Aditya Grover, and Ricky TQ Chen. Guided flows for generative modeling and decision making. arXiv preprint arXiv:2311.13443, 2023.

Huminhao Zhu, Fangyikang Wang, Chao Zhang, Hanbin Zhao, and Hui Qian. Neural sinkhorn gradient flow. arXiv preprint arXiv:2401.14069, 2024.

CONTENTS

A EXTENDED DISCUSSION

We first review related work in §A.1 and then discuss the validity of time-varying potentials proposed by Terpin et al. (2024) in §A.2 in contrast to the stationary setting considered in this paper. Finally, in §A.3, we discuss potential reasons for the performance gap between the paired and unpaired setups introduced in §5.1.

A.1 RELATED WORKS

In this section, we first provide an extended discussion of related work on learning population dynamics (also known as trajectory inference) in §A.1.1, followed by an overview of applications of the JKO scheme (Jordan et al., 1998) in deep learning in §A.1.2.

A.1.1 LEARNING POPULATION DYNAMICS

In contrast to our approach, which models population dynamics as a Wasserstein gradient flow (WGF) (Ambrosio et al., 2008) of the free-energy functional (5) (Gómez-Castro, 2024), several other fruitful approaches have been proposed in the literature. We review these alternatives below.

RNNs and Neural ODEs. One of the first works on learning population dynamics is (Hashimoto et al., 2016), which used recurrent neural networks (RNNs) to learn the SDE (7) as a WGF of a poten-

tial energy functional. Later, (Chen et al., 2018a, CNF) introduced neural ODEs as the continuous-time limit of RNNs for learning ODE-based dynamics, and (Li et al., 2020) subsequently extended this approach to handle SDEs. Building on these ideas, (Erbe et al., 2023, `RNAForecaster`) applied neural ODEs to predict future transcriptomic states of single cells (Battich et al., 2020). More recently, (Li et al., 2025a) proposed `WeightFlow`, which models stochastic population dynamics by learning the continuous evolution of neural network weights that parameterize distributions at each time point, rather than modeling the dynamics directly in the state space. Their key idea is to project the evolving distribution into the weight space of a backbone neural network trained to represent the probability distribution at each snapshot.

Static OT. Another line of work models population dynamics by directly learning K *independent* transport maps from time-snapshot data, using the *static* optimal transport formulation (1) or (2), without attempting to recover the unifying energy functionals guiding the system evolution. For example, (Schiebinger et al., 2019, `Waddington-OT`) employs discrete transport maps, whereas (Bunne et al., 2023, `CellOT`) learns *neural* transport maps, as in (Fan et al., 2023, Eq.(6)), (Rout et al., 2022, Eq.(14)), and (Korotin et al., 2023, Eq.(15)), see Appendix A.2 for more details.

Dynamic OT. (Tong et al., 2020, `TrajectoryNet`) models population dynamics using the dynamic optimal transport formulation of (Benamou & Brenier, 2000), parameterized by a regularized CNF. The work (Huguet et al., 2022, `MIOFlow`) later extended this approach by incorporating a geodesic autoencoder to better capture manifold structure. More recently, (Wan et al., 2023) proposed computational techniques that make dynamic OT scalable to higher-dimensional settings.

Flow Matching. The works in this paragraph extend Flow Matching (FM) techniques (Lipman et al., 2023; Albergo & Vanden-Eijnden, 2023; Liu et al., 2023) in several directions. Conditional Flow Matching (Tong et al., 2024a, `CFM`) introduce simulation-free objectives for learning deterministic flows. Building on the celebrated work of (Otto, 2001), the Wasserstein space can be nominally (i.e., heuristically) viewed as a Riemannian manifold. Motivated by this observation, Wasserstein Flow Matching (Haviv et al., 2025, `WFM`) applies the idea of Riemannian Flow Matching (Chen & Lipman, 2024, `RFM`) to the Wasserstein space, i.e., performing FM over a distribution of distributions. Further developments, such as Meta Flow Matching (Atanackovic et al., 2025, `Meta FM`), embed the population of samples with a Graph Neural Network (GNN) (Liu et al., 2025) and use these embeddings as conditioning inputs for the learning vector field. Finally, Multi-Marginal Flow Matching (Rohbeck et al., 2025, `MMFM`) constructs a flow using smooth spline-based interpolation across time points and conditions, and regresses it with a neural network under the classifier-free guided Flow Matching framework (Zheng et al., 2023). For a broader overview of FM for learning population dynamics, see the recent surveys (Morehead et al., 2025; Li et al., 2025b).

Action Matching. (Neklyudov et al., 2023, `AM`) introduced Action Matching (AM), a method that optimizes an action-gap objective. In contrast to Flow Matching, AM learns a vector field expressed as the gradient of an action, ∇s_t^*, which uniquely defines the velocity field that transports particles optimally in the sense of optimal transport. Building on this idea, (Berman et al., 2024) conditioned the action $s_{t,\mu}$ on physical parameters μ, used (Berman & Peherstorfer, 2024, `CoLoRa`) for efficient action parametrization, and demonstrated its utility for surrogate modeling of classical numerical solvers, enabling fast prediction of system behavior across different physics parameter settings. Most recently, (Neklyudov et al., 2024b, `WLF`) proposed Wasserstein Lagrangian Flows, a unifying framework that minimizes Lagrangian action functionals over the space of probability densities rather than the ground space, thereby encapsulating AM as a special case.

Schrödinger bridges. Two seminal works, Vargas et al. (2021, `IPF (GP)`) and De Bortoli et al. (2021, `IPF (NN)`), proposed solving the Schrödinger bridge problem using the Iterative Proportional Fitting (IPF) algorithm (Gramer, 2000) in application to generative modeling, which alternates between forward and backward processes. The first approach employed Gaussian processes as parametrization, whereas the second relied on score-based neural networks. Later, bridge matching methods was developed (Peluchetti, 2023b;a; Liu & Wu, 2023; Shi et al., 2023). Tong et al. (2024b, `[SF]`^{2}`M`) introduced simulation-free objectives for both deterministic and stochastic flows, further reducing computational overhead. Additional contributions include generalized setups for Schrödinger bridges (Koshizuka & Sato, 2023; Liu et al., 2024), multi-marginal generalizations of the problem (Chen et al., 2023, `DMSB`), (Shen et al., 2025, `MMSB`), (Hong et al., 2025), and momentum-accelerated formulations (Theodoropoulos et al., 2025, `3MSBM`).

Unbalanced OT. Several works jointly model marginal transitions and growth dynamics by minimizing the action in the Wasserstein-Fisher-Rao (WFR) metric (Chizat et al., 2018a), i.e., by solving the dynamical unbalanced optimal transport problem (Chizat et al., 2018b). (Tong et al., 2023) proposed `BEMIOflow`, an extension of (Huguet et al., 2022, `MIOflow`), that incorporates a neural network to predict cell growth and death rates continuously in time, thereby augmenting optimal transport with population-size dynamics. Later, (Chen et al., 2022b) formulated an unbalanced Schrödinger Bridge problem, which was applied in practice by (Zhang et al., 2025b, `DeepRUOT`), following the developments in (Sha et al., 2024, `TIGON`). `DeepRUOT` was further simplified into a single-network formulation, `Var-RUOT` (Sun et al., 2025), by exploiting optimality conditions; extended to include interaction modeling in (Zhang et al., 2025d, `CytoBridge`); and adapted to use two independent networks to separately model distributional drift and mass growth (Wang et al., 2025, `VGFM`), (Zhang et al., 2025a). In a different direction, (Klein et al., 2024, `GENOT`) generalized OT to simultaneously handle stochasticity, entropy regularization, and unbalanced mass transport, enabling applications to cross-modal and heterogeneous data. For a broader survey of trajectory inference methods and their applications to biology data, see (Zhang et al., 2025e;c).

The concurrent work (Andrade et al., 2025, `iJKO`) also connects inverse optimization with the JKO scheme (8) but in the unbalanced setting. However, their focus is on sample complexity, the method addresses only potential energy, and experiments are restricted to low-dimensional cases.

Riemannian perspective. (Scarvelis & Solomon, 2023) proposed learning a metric tensor $A(x)$ that minimizes the average 1-Wasserstein distance on the learned manifold between pairs of consecutive population snapshots. Furthermore, (Kapusniak et al., 2024, `MFM`) introduced Metric Flow Matching (MFM), which learns interpolants $x_{t,\eta} = (1-t)x_0 + tx_1 + t(1-t)\varphi_{t,\eta}(x_0, x_1)$, where η are the parameters of a neural network $\varphi_{t,\eta}$ providing a nonlinear "correction" to straight-line interpolants (Albergo & Vanden-Eijnden, 2023). The resulting velocity field minimizes a data-dependent Riemannian metric, assigning lower transport cost to regions with higher data density.

Splines in Wasserstein Space. In the seminal work of (Schiebinger et al., 2019), a piecewise linear OT interpolation method was proposed to infer cell trajectories. Subsequent works introduced higher-order piecewise polynomials (e.g., cubic splines) in Wasserstein space: (Chen et al., 2018b; Benamou et al., 2019) formulated a global cubic spline minimization problem, which was later extended by (Chewi et al., 2021) to use Euclidean interpolation algorithm in $\mathbb{R}^D$ after a finding optimal Monge map (2) between samples from consecutive measures, with further refinements by (Clancy & Suarez, 2022; Justiniano et al., 2024). More recently, (Banerjee et al., 2025, `WLR`) proposed the Wasserstein Lane–Riesenfeld method, leveraging the classical subdivision algorithm of (Lane & Riesenfeld, 1980). In parallel, (Dyn & Sharon, 2025) developed subdivision schemes for general metric spaces, including the Wasserstein space. (Baccou & Liandrat, 2024; Kawano et al., 2025) combines subdivision schemes with OT.

Dynamics from stationary distributions. A recent line of research aims to recover meaningful dynamical structure from static data alone. The Equilibrium Flow framework of (Zhang & Levin, 2025) addresses this problem by learning a vector field whose flow preserves the observed stationary distribution p. Their method enforces the invariance condition $\nabla \cdot [p(x,t)v(x,t)] = 0$ using score-based estimates of $\nabla \log p$, enabling the recovery of plausible dynamics even without temporal information or trajectories. Despite this minimal supervision, the learned flows can be used for inverse design, where one constructs dynamics that yield a desired target distribution.

A.1.2 JKO SCHEME MEETS DEEP LEARNING

Generative Modeling. (Vidal et al., 2023, `JKO-Flow`) and (Xu et al., 2023, `JKO-iFlow`) reinterpret the JKO scheme (8) through the framework of CNFs (Chen et al., 2018b), applying it to 2D synthetic data and image generation tasks, respectively. Subsequently, (Cheng et al., 2024) provided theoretical convergence guarantees for this approach. (Choi et al., 2024, `S-JKO`) further accelerated WGF-based generative modeling by leveraging a semi-dual unbalanced OT formulation. They constructed the WGF of an f-divergence D_f and used an (Song et al., 2021, `NCSN++`) backbone – similar to (Zhu et al., 2024, `NSGF`), who designed a generative model as the WGF with respect to the Sinkhorn divergence (Peyré et al., 2019).

Tensor Train (TT) (Oseledets, 2011). (Aksenov & Eigel, 2025) introduces a method for approximating probability distributions in Bayesian inversion by minimizing a KL-based functional via an

entropically regularized JKO scheme. The approach discretizes the resulting coupled heat equations on a high-dimensional grid and solves them using accelerated fixed-point methods combined with low-rank TT representations. Related work includes (Chertkov & Oseledets, 2021), which applies TT approximations to the Fokker-Planck equation with drift and diffusion, and (Han et al., 2025), which employs entropy-regularized proximal steps with TT-cross for particle-based evolution.

Variational Inference. (Lambert et al., 2022) and (Diao et al., 2023) study variational inference through the lens of Bures–Wasserstein gradient flows. (Cheng et al., 2023, GWG) introduce a generalized minimizing movement scheme on the space $\mathcal{P}_{c_h}(\mathbb{R}^D)$, where the transport cost is defined as $c_h(x, y) = g\big(\frac{x-y}{h}\big)\, h$ with g a continuously differentiable Young function, and apply this framework to particle-based variational inference.

Datasets/Weights Learning. (Bonet et al., 2025) propose a explicit scheme that is computationally more efficient in practice than the implicit JKO scheme on the space $\mathcal{P}_2(\mathcal{P}_2(\mathbb{R}^D))$, and apply this approach to modeling flow datasets viewed as random measures. (Saragih et al., 2025) train a meta-model, based on JKO as well as other flow- and diffusion-based approaches, that generates dataset-conditioned classifiers by producing the neural network weights in a single forward pass.

A.2 Time-varying potentials discussion

In this section, we examine the validity of using time-varying potentials, as proposed in (Terpin et al., 2024, §4.4). We demonstrate that under this formulation, the original problem of reconstructing energy functionals (Bunne et al., 2022b) via JKO Scheme reduces to learning K independent neural optimal transport maps (Fan et al., 2023, Eq.(6)), (Rout et al., 2022, Eq.(14)) between the data snapshots ρ_k and ρ_{k+1}. To illustrate this, consider the loss in (11) and assume that, instead of a single functional $\mathcal{J}$, we now assign a separate only potential energy functional $\mathcal{J}^k_{\text{PE}} = \int_{\mathcal{X}} V^k(x)\,\mathrm{d}\rho_k(x)$ for each time step k:

$$\max_{\mathcal{J}^k_{\text{PE}}} \min_{T^k} \mathcal{L}(\mathcal{J}^k, T^k) \overset{\text{def}}{=} \max_{\mathcal{J}^k_{\text{PE}}} \min_{T^k} \sum_{k=0}^{K-1} \Big[\mathcal{J}^k_{\text{PE}}(T^k \sharp \rho_k) - \mathcal{J}^k_{\text{PE}}(\rho_{k+1}) + \frac{1}{2\tau}\int_{\mathcal{X}} \|x - T^k(x)\|_2^2 \rho_k(x)\,\mathrm{d}x\Big].$$

It then becomes clear that the loss optimization decomposes into K independent terms $\mathcal{L}^k(\mathcal{J}^k_{\text{PE}}, T^k)$:

$$\mathcal{L}^k(\mathcal{J}^k_{\text{PE}}, T^k) \overset{\text{def}}{=} \mathcal{J}^k_{\text{PE}}(T^k \sharp \rho_k) - \mathcal{J}^k_{\text{PE}}(\rho_{k+1}) + \frac{1}{2\tau}\int_{\mathcal{X}} \|x - T^k(x)\|_2^2 \rho_k(x)\,\mathrm{d}x \to \max_{\mathcal{J}^k_{\text{PE}}} \min_{T^k}. \quad (18)$$

By replacing V^k with the dual potential f_η and T^k with T_θ, we recover Equation (6) from (Fan et al., 2023), which aligns with the objective proposed in (Bunne et al., 2023, Eq. (9)) for modeling single-cell dynamics. Thus, (Terpin et al., 2024) introduces a new strategy for minimizing this loss rather than directly addressing the recovery of ground-truth energy functionals (Bunne et al., 2022b).

Fortunately, the theoretical framework developed by Ferreira & Valencia-Guevara (2018) and later extended by Plazotta & Zinsl (2016) establishes the **validity** of this approach even for time-varying free energy functionals (5), not only for the potential energy case, in the limit as the time discretization tends to zero. However, this connection is not acknowledged in (Terpin et al., 2024).

A.3 Paired vs. unpaired setups

In this section, we discuss the substantial performance gap observed between the paired and unpaired setups, as illustrated in Table 3. We suggest that this discrepancy stems from the intrinsic connection between the considered solvers, `iJKOnet` and `JKOnet*`, and the underlying OT problem (2).

Both solvers are designed to recover the optimal transport mappings $T^{k,*}$ between consecutive distributions ρ_k and ρ_{k+1}, satisfying the pushforward relation $T^{k,*}_{\mathcal{J}} \sharp \rho_k = \rho_{k+1}$. In `JKOnet*`, these mappings are obtained explicitly by precomputing the OT maps between ρ_k and ρ_{k+1}, whereas in our method, they are learned implicitly via neural networks T^k_φ.

In practice, the continuous distributions ρ_k are replaced by their empirical counterparts, constructed from training samples $\{x^1_k, x^2_k, \ldots, x^N_k\} = X_k \sim \rho_k$. This is precisely where the distinction between the paired and unpaired setups becomes critical:

- **Paired.** When samples are paired, they follow the ground-truth transport mappings, i.e.,

$$T^{0,*}(x_0^i) = x_1^i, \quad T^{1,*}(x_1^i) = x_2^i, \quad \dots, \quad T^{k,*}(x_k^i) = x_{k+1}^i, \quad \dots$$

  In this case, estimating $T^{k,*}$ becomes substantially simpler and can be viewed as a supervised regression problem over the given sample pairs.
- **Unpaired.** In contrast, when the samples X_k, $k \in \{0, \dots, K\}$, are mutually independent, the quality of the estimated OT maps degenerate significantly. This degradation stems from the high sample complexity of Wasserstein-2 distances and maps (known to be relatively poor, see Hütter & Rigollet (2021)), as well as from the increased variance of the estimated transport mappings.

A deeper theoretical analysis of this gap between paired and unpaired setups remains an interesting direction for future work. To quantify how performance changes when transitioning from unpaired to paired data, we conducted the following *ablation study*.

Experimental Setup. Following Appendix B.1.1, we used the same set of potentials and focused solely on the potential parametrization. In the $2D$ setting with $K = 5$, we trained for $5K$ iterations using $N = 2K$ training samples per time step and a 40% test split. The fraction of full trajectories in the training set was gradually increased from 0.0 (unpaired) to 1.0 (paired) in steps of 0.25.

Results. Figure 3 shows the results. As the fraction of paired samples decreases, the performance of `JKOnet*` degrades more rapidly than that of our method. For the Bohachevsky potential, both methods show reduced accuracy in both paired and unpaired scenarios.

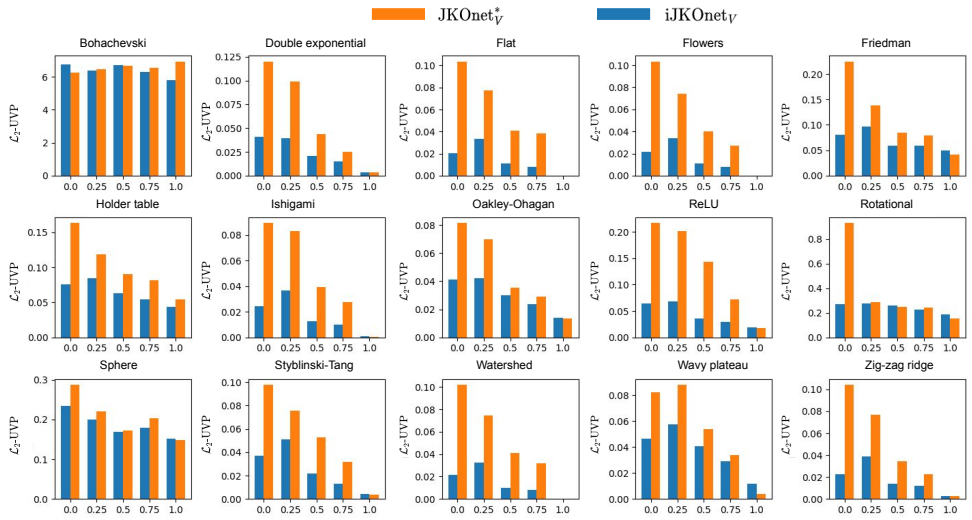


Figure 3: Ablation study comparing *paired* (1.0) and *unpaired* (0.0) setups. As the proportion of paired samples (i.e. full trajectories in the dataset) decreases, the performance of `JKOnet*` degenerate more rapidly than that of our method.

B Extended Comparisons

In this section, we first present extended synthetic experiments in §B.1, followed by results on real-world single-cell data in §B.2.

B.1 Synthetic comparisons

In this section, we further investigate our method's ability to learn potential energy (§B.1.1), interaction energy (§B.1.2), and internal energy (§B.1.3).

Since our approach builds on the JKO framework for modeling population dynamics, we adopt the experimental setup from JKOnet* and compare our method against both JKOnet Bunne et al. (2022b) and JKOnet* Terpin et al. (2024) on corresponding synthetic evaluation tasks.

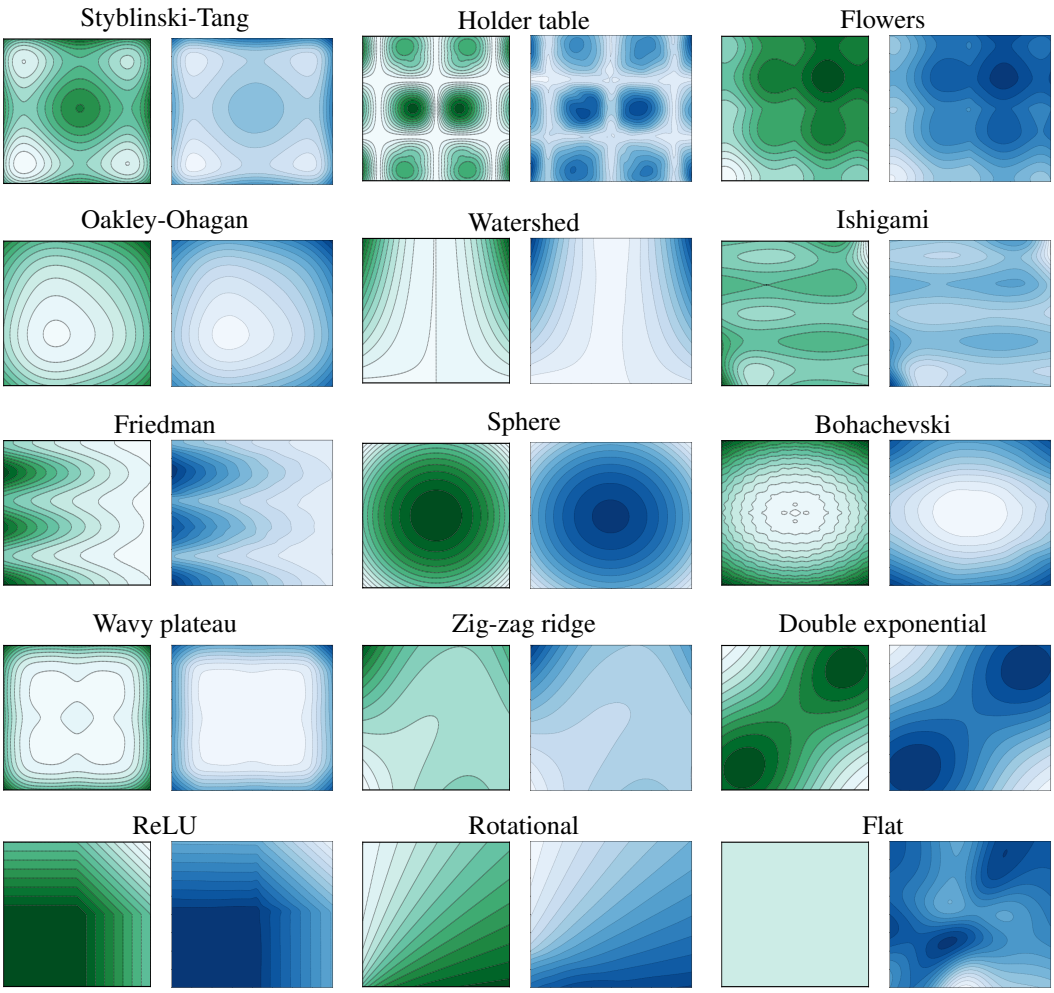


Figure 4: Level curves of the true (green) and estimated (blue) potentials for the **paired** setup, following (Terpin et al., 2024, Appendix F). These results can be directly compared with those in (Terpin et al., 2024, Figure 6). Note that for the *flat* potential, the value range is near zero, as expected for the ground-truth potential.

B.1.1 Learning potential energy

Paired Setup. Following Terpin et al. (2024, §4.1), we evaluate our method in the paired setup (see details in §5.1), which enables a direct visual comparison with JKOnet* (Terpin et al., 2024). The experiments are conducted on the synthetic dataset from Terpin et al. (2024, Appendix B), using a step size of $\tau = 0.01$, $K = 5$ time steps, and $N = 2000$ samples per step with 40% of test samples. Figure 4 presents the ground-truth potentials $V(x)$ (green), defined in Terpin et al. (2024, Appendix F, Eqs. (31)–(45)), alongside the reconstructed potentials $V_\theta(x)$ (blue) learned by our method. As shown, our approach achieves performance comparable to JKOnet* in the paired setup.

Unpaired Setup. We repeat the experiments on a corrected version of the synthetic dataset, where samples are not temporally correlated. We select the six most challenging potentials in terms of convergence. Table 3 reports the ratio of the final metrics between the unpaired and paired setups. As shown, switching to the unpaired setup significantly affects the method's performance.

Table 3: Comparison of paired and unpaired setups. Each value indicates the ratio of the final metric in the unpaired setup to that in the paired setup. Most ratios significantly exceed 1, highlighting the increased difficulty of the unpaired setting.

		iJKOnet$_V$			JKOnet$^*_V$		
Alias	Potential Name	B$d^2_{\mathbb{W}_2}$-UVP	EMD	$\mathcal{L}_2$-UVP	B$d^2_{\mathbb{W}_2}$-UVP	EMD	$\mathcal{L}_2$-UVP
FL	Flowers	8066	90	297	135325	305	5809
FN	Friedman	556	5	2	263	6	3
IG	Ishigami	2259	38	36	4276	64	210
WS	Watershed	18135	220	1509	512720	1014	102592
WP	Wavy plateau	2	2	7	1	2	29
ZR	Zigzag ridge	158	23	13	155	25	31

B.1.2 Learning interaction energy

In this section, we evaluate how accurately our method and JKOnet* (Terpin et al., 2024) can approximate the interaction energy in the unpaired setup. We exclude JKOnet from comparison, as its design does not support learning interaction energies.

We identified an inconsistency in the theoretical formulation presented in the JKOnet* paper (Terpin et al., 2024). Specifically, as discussed in §2 (see also (Santambrogio, 2015, §7.2)), the ground-truth interaction functional must be symmetric; that is, in (5), the interaction kernel should satisfy $W(z) = W(-z)$. Incorporating this correction, we conducted an experiment to assess the ability of both methods to recover the *interaction* energy in a 2D unpaired setting.

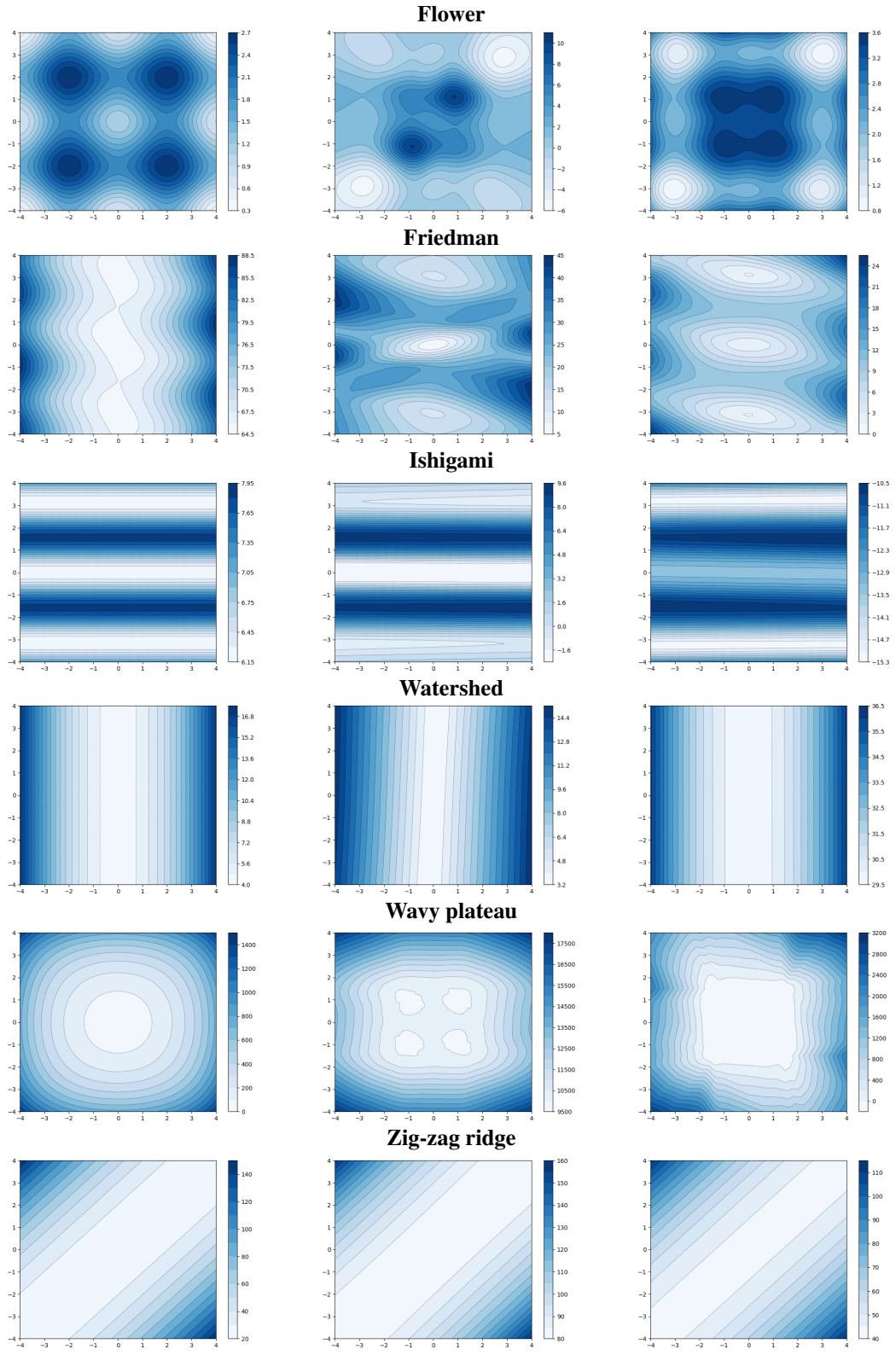
Experimental Setup. The ground-truth interaction kernels are defined as $W(z) = \frac{1}{2}\big(W_b(z) + W_b(-z)\big)$, where W_b denotes the base functionals listed in Table 3. Both methods were restricted to use only the interaction energy component $\mathcal{W}_{\theta_2}$ in the energy parameterization (12). The corresponding variants are denoted as iJKOnet$_W$ and JKOnet$^*_W$. Similar to §B.1.1, we use a step size of $\tau = 0.01$, $K = 5$ time steps, and two sample size settings: a small-scale setup with 12K total samples ($N = 2K$ per step, with 40% reserved for testing) and a large-scale setup with 60K total samples ($N = 10K$ per step, with 40% reserved for testing).

Results. Qualitative results are shown in Figure 5 for the paired setup and in Figure 6 for the unpaired setup, with quantitative $\mathcal{L}_2$-UVP metrics reported in Table 4 (see §C.1 for details). In the paired setup, most potentials are accurately restored, although Waby Plateau, Friedmann, and Flowers are not. In the unpaired setup, neither method successfully recovers the interaction energy, likely due to biases introduced by the batched estimation of the interaction term $\mathcal{W}_{\theta_2}$ in (12), which involves squaring the input. This approximation seems to cause both models to converge to batch-specific minima.

Table 4: Quantitative $\mathcal{L}_2$-UVP results for the **unpaired** setup in 2D interaction energy learning. Increasing the number of samples improves performance, though the effect is relatively limited.

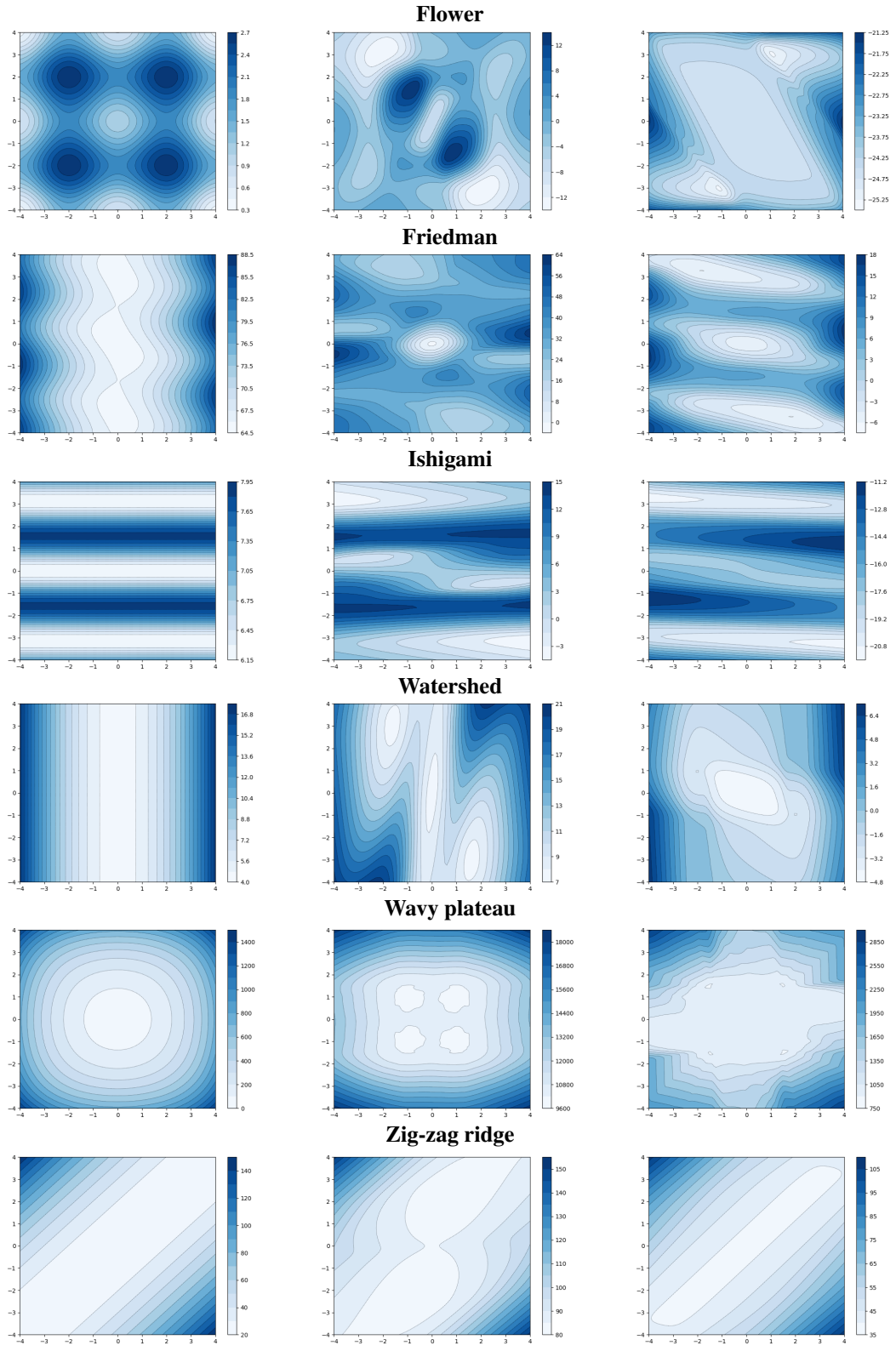
	Sample size: 2K		Sample size: 10K	
Potential	iJKOnet$_W$ **(Ours)**	JKOnet$^*_W$	iJKOnet$_W$ **(Ours)**	JKOnet$^*_W$
Flowers	0.0032	0.0101	0.0009	0.0069
Friedman	0.0087	0.0010	0.0024	0.0093
Ishigami	0.0046	0.0087	0.0008	0.0066
Watershed	0.0043	0.0086	0.0009	0.0081
Zigzag Ridge	0.0060	0.0104	0.0017	0.0034

B.1.3 Learning internal energy



(a) Ground-truth $\mathcal{W}^*$ (b) $\texttt{iJKOnet}_W$ recovery of $\mathcal{W}^*$ (c) $\texttt{JKOnet}^*_W$ recovery of $\mathcal{W}^*$

Figure 5: Qualitative results from §B.1.2 in the **paired** setup, using 10K samples per time step (with 40% reserved for testing). Compared to Figure 6, which shows the unpaired setup, the model performs noticeably better: nearly all potentials are accurately reconstructed, highlighting the relative simplicity of the paired scenario.



(a) Ground-truth $\mathcal{W}^*$ (b) $\texttt{iJKOnet}_W$ recovery of $\mathcal{W}^*$ (c) $\texttt{JKOnet}^*_W$ recovery of $\mathcal{W}^*$

Figure 6: Qualitative results from §B.1.2 in the **unpaired** setup with 10K samples in total (40% reserved for testing). Neither model is able to accurately reconstruct the ground-truth interaction energy $\mathcal{W}^*$, likely due to biases introduced in the estimation of the integral in (12).

In this section, we assess how accurately our method and JKOnet* (Terpin et al., 2024) can estimate the diffusion coefficient θ_3 in (12) corresponding to the internal energy term under the **unpaired** setup. We exclude JKOnet from comparison, as its design does not support learning internal energy.

Experimental Setup. Following the protocol of Terpin et al. (2024, §4.3), we estimate the diffusion coefficient θ_3 in 2D and 20D settings using both iJKOnet and JKOnet*, i.e., the full parameterization in (12). The same functional from Appendix C.7 is used for both potential and interaction energies. The ground-truth diffusion levels are set to $\beta^* \in \{0.0, 0.1, 0.2\}$.

Results. Figure 7 shows that iJKOnet fails to recover the ground-truth β^* values, while JKOnet* provides a closer approximation. However, all predicted θ_3 values tend to converge toward 0, deviating from the true levels $\beta^* \in \{0.0, 0.1, 0.2\}$. This indicates that accurately learning the internal energy remains challenging for both models. We assume that such behavior stems from difficulties of entropy estimation developed methods.

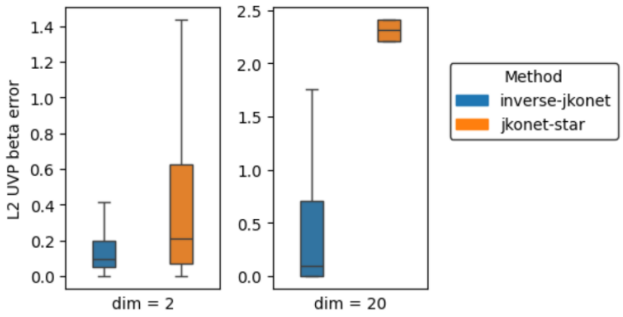


Figure 7: Estimation error of the diffusion coefficient θ_3 relative to the ground-truth values β^*. Blue bars correspond to iJKOnet and orange bars to JKOnet*. The Y-axis indicates the absolute deviation between the estimated θ_3 and true β^* values.

B.2 Learning single-cell dynamics

In this section, we present extended comparisons on real-world single-cell datasets, starting with 5D results and then considering 50D and 100D.

Datasets. Following §5.2, we consider the Embryoid Body (**EB**) dataset (Moon et al., 2019) and additionally the Multiome (**Multi**) dataset (Burkhardt et al., 2022). For both datasets, we apply the preprocessing pipeline of (Tong et al., 2020). The Multi dataset contains single-cell measurements across four time points (days 2, 3, 4, and 7), which we use for high-dimensional experiments.

5D. Following (Neklyudov et al., 2024a, Table 2) and (Terpin et al., 2024, §4.4), we conducted experiments on the **EB** dataset in 5D. At each time step, we train for 1000 epochs on 60% of the data and compute the $d_{\mathbb{W}_1}$ distance between the observed distribution ρ_k (remaining 40% of the data) and the one-step-ahead prediction $\hat{\rho}_k$. The results are shown in Table 5. Our method outperforms several previous approaches when using a time-varying potential parametrization.

50D and 100D data. Following Neklyudov et al. (2024a), we conducted experiments on the **Multi** dataset in 50D and 100D using a leave-one-out setup, averaged over three runs. Models were trained on marginals from timepoint partitions [2,4,7] and [2,3,7], and evaluated on the corresponding left-out marginals at timepoints [3] and [4]. Training was performed for 5000 epochs.

The results are presented in Table 6. Our method achieves comparable performance for the left-out [3] marginal. However, performance on the [4] marginal is poor, leading to a lower overall average. This is likely due to the longer time intervals between the learned marginals, which violate our assumption that consecutive marginals are obtained from JKO steps (see §3.1), thereby reducing the accuracy of predicting the [4] marginal.

Table 5: The results for the **EB** dataset in 5D are obtained by training on all time steps, evaluated for next-time-step prediction using the $d_{\mathbb{W}_1}$ metric. Results for non-JKO methods are taken from (Neklyudov et al., 2024a, Table 2):

Model	t_1	t_2	t_3	t_4	**Mean**
Neural SDE (Li et al., 2020)	0.69	0.91	0.85	0.81	0.82
TrajectoryNet (Tong et al., 2020)	0.73	1.06	0.90	1.01	0.93
SB-FBSDE (Chen et al., 2022a)	0.56	0.80	1.00	1.00	0.84
NLSB (Koshizuka & Sato, 2023)	0.68	0.84	0.81	0.79	0.78
OT-CFM (Tong et al., 2024a)	0.78	0.76	0.77	0.75	0.77
WLF-OT (Neklyudov et al., 2024a)	0.65	0.78	0.76	0.75	0.74
WLF-SB (Neklyudov et al., 2024a)	0.63	0.79	0.77	0.74	0.73
JKOnet (Bunne et al., 2022b)	1.53	1.27	1.13	1.41	1.34
JKOnet$^*_V$ Terpin et al. (2024)	0.99	1.11	1.06	1.30	1.12
iJKOnet$_V$ **(Ours)**	0.92	1.11	0.95	1.21	1.05
JKOnet$^*_{t,V}$ (Terpin et al., 2024)	0.69	0.77	0.69	0.78	0.73
iJKOnet$_{t,V}$ **(Ours)**	0.51	0.58	0.57	0.64	**0.58**

Table 6: The results for **Multi** dataset for direct comparison with from (Neklyudov et al., 2024a, Table 1), averaged for 3 runs for leave-one-out setup, $d_{\mathbb{W}_2}$ metric.

Dimension	**50**			**100**		
Metric	**BW-UVP**	**EMD**	**MMD**	**BW-UVP**	**EMD**	**MMD**
JKOnet$^*_V$	128.729 ± 78.307	68.406 ± 6.055	0.0003	92.512 ± 38.075	72.639 ± 3.161	0.0002
iJKOnet$_V$ **(Ours)**	38.318 ± 18.325	50.560 ± 7.187	0.0003	38.991 ± 16.918	59.216 ± 6.869	0.0002
JKOnet$^*_{t,V}$	89.264 ± 38.372	68.600 ± 9.846	0.0003	93.429 ± 34.332	78.674 ± 8.236	0.0002
JKOnet$_{t,V}$ **(Ours)**	36.156 ± 16.515	50.026 ± 7.727	0.0003	38.387 ± 18.525	59.318 ± 7.365	0.0002

C Training and Model Details

C.1 Metric computation details

In this section, we provide details on the evaluation metrics used in our experiments.

$\mathcal{L}_2$-UVP. When the ground-truth functional F^* (e.g., V or W) is available, we assess the discrepancy between it and its reconstruction $\widehat{F}$ using the *backward $\mathcal{L}_2$-based Unexplained Variance Percentage* ($\mathcal{L}_2$-UVP) metric introduced by Korotin et al. (2021a), defined as follows:

$$\mathcal{L}_2\text{-UVP}(F^*, \widehat{F}) = 100 \cdot \frac{\tau^2 \|\nabla \widehat{F} - \nabla F^*\|^2_{\rho_{k+1}}}{\text{Var}(\rho_k)}\%, \tag{19}$$

where the norm $\|\cdot\|_{\rho_{k+1}}$ is computed with respect to the ground-truth distribution ρ_{k+1}, and τ denotes the time step size. Values close to 0% indicate that $\nabla \widehat{F}$ closely approximates ∇F^*.

B$d^2_{\mathbb{W}_2}$-UVP. The *Bures-Wasserstein UVP* introduced by (Korotin et al., 2021c) is defined as

$$\text{B}d^2_{\mathbb{W}_2}\text{-UVP}(\rho_k, \hat{\rho}_k) \stackrel{\text{def}}{=} 100 \cdot \frac{\text{B}d^2_{\mathbb{W}_2}(\rho_k, \hat{\rho}_k)}{\frac{1}{2}\text{Var}(\rho_k)}\%, \tag{20}$$

where the *Bures-Wasserstein distance* is given by

$$\text{B}d^2_{\mathbb{W}_2}(\mathbb{P}, \mathbb{Q}) \stackrel{\text{def}}{=} d^2_{\mathbb{W}_2}\left(\mathcal{N}(\mu_{\mathbb{P}}, \Sigma_{\mathbb{P}}), \mathcal{N}(\mu_{\mathbb{Q}}, \Sigma_{\mathbb{Q}})\right), \tag{21}$$

with $\mu_{\mathbb{P}}$ and $\Sigma_{\mathbb{P}}$ denoting the mean and covariance of distribution $\mathbb{P}$, respectively.

EMD or $d_{\mathbb{W}_1}$. The *Earth Mover's Distance* (EMD) (Rubner et al., 1998) is defined as

$$\text{EMD}(\rho_k, \hat{\rho}_k) \equiv d_{\mathbb{W}_1}(\rho_k, \hat{\rho}_k) \stackrel{\text{def}}{=} \min_{\pi \in \Pi(\rho_k, \hat{\rho}_k)} \int_{\mathcal{X}\times\mathcal{X}} \|x - y\| \, \text{d}\pi(x, y), \tag{22}$$

where ρ_k and $\hat{\rho}_k$ denote the ground truth and predicted distributions at time step k.

MMD. *Maximum Mean Discrepancy* (MMD) (Gretton et al., 2012) between two distributions ρ_k and $\hat{\rho}_k$ is defined as

$$\mathrm{MMD}^2(\rho_k, \hat{\rho}_k) = \|\mathbb{E}_{x\sim\rho_k}[h(x)] - \mathbb{E}_{\hat{x}\sim\hat{\rho}_k}[h(\hat{x})]\|^2_{\mathcal{H}}\,, \tag{23}$$

where h is the feature map associated with the reproducing kernel $\mathbf{H}$, and $\|\cdot\|_{\mathcal{H}}$ denotes the norm in the corresponding RKHS (Ghojogh et al., 2021). Here, ρ_k and $\hat{\rho}_k$ are defined as in the EMD metric.

C.2 Training details

Energy. To ensure stability, we accumulate gradients across all time steps $k = 0, \ldots, K$, using mini-batch estimates $\widehat{\mathcal{J}}_\theta$ of the energy function $\mathcal{J}_\theta$ (see Eq. (12)). For each time step t_k, we sample a mini-batch $X_k = \{x_k^1, \ldots, x_k^B\} \sim \rho_k$, with a fixed batch size B. The energy is then estimated as:

$$\mathcal{J}_\theta(\rho_k) \approx \widehat{\mathcal{J}}_\theta(X_k) = \frac{1}{B}\sum_{i=0}^{B}\left[V_{\theta_1}(x_k) + \frac{1}{B}\sum_{j=0}^{B} W_{\theta_2}(x_i - x_j)\right] - \theta_3\widehat{\mathcal{H}}(X_k), \tag{24}$$

where $\widehat{\mathcal{H}}(X_k)$ denotes the estimated entropy of ρ_k discussed in the following section.

Map. In practice, we use two strategies to parameterize each transport map T_φ^k: for large-scale tasks, we assign a separate network to each time step; for moderate-dimensional tasks, we encode the time index k as an additional input, i.e., $T_\varphi^k(x) = T_\varphi(x, k)$, which helps prevent overfitting. This approach performs better than using a network without any time embedding. Thanks to `Optax` (DeepMind et al., 2020), the inference of all T_φ^k can be performed in parallel, treating them as a single 'generator' step. We also experimented with initializing the maps during the early training epochs by aligning them with discrete OT maps computed using the `OTT-JAX` (Cuturi et al., 2022).

Training. We aggregate gradients for $\mathcal{J}_\theta$ and T_φ^k across all time steps. We also experimented with alternative aggregation strategies, but found that equal aggregation across all time steps was the most effective. During optimization, we perform multiple update steps for the transport maps T_φ^k, parameterized by φ, while updating the energy function $\mathcal{J}_\theta$, parameterized by θ, only once. This follows standard practice in min–max optimization (Goodfellow et al., 2020). The overall procedure is summarized in the pseudocode in Algorithm 1.

Stability. The main challenge was avoiding suboptimal energy functionals that did not converge to zero. We found that the simplest setup (combined with careful hyperparameter tuning using the `Optuna` framework (Akiba et al., 2019)) was the most effective. We additionally experimented with common stabilization techniques from GAN training, including gradient penalties (Gulrajani et al., 2017), spectral normalization (Miyato et al., 2018; Jiang et al., 2018), and extragradient updates (Daskalakis et al., 2018). Overall, we perform optimization for single energy functional, this naturally acts as a regularizer, improving the stability of the training.

Scalability. Our method scales well with the number of time points. Leveraging `JAX` (Bradbury et al., 2018) and `Optax` (DeepMind et al., 2020), we avoid memory issues even for large batches and many time steps, since we do not backpropagate through time as in `JKOnet` (Bunne et al., 2022b). Each step is processed independently, and the loss is computed by averaging outputs from separate networks, avoiding large computation graphs. Training time on GPU is comparable to `JKOnet`*, but our approach eliminates the costly precomputation of OT couplings. In high-dimensional settings (50D–100D), `JKOnet`* exceeds GPU memory limits and requires CPU-based precomputations, making it substantially slower.

C.3 Entropy estimation details

Prior to training, we estimated $\widehat{\mathcal{H}}(\rho_k)$ for each $k = 0, \ldots, K$ using the Kozachenko–Leonenko nearest-neighbor estimator (Kozachenko, 1987; Berrett et al., 2019) with 5 nearest neighbors. We used the publicly available implementation from (Butakov et al., 2024)[2]. As discussed in §3.3, we

[2] `https://github.com/VanessB/mutinfo`

Algorithm 1: `iJKOnet` Training

Input:
Sequence of measures $\{\rho_k\}_{k=0}^K$ (accessible via samples);
Mapping networks $T_\varphi^k : \mathcal{X} \to \mathcal{X}$;
Potential network $V_{\theta_1} : \mathcal{X} \to \mathbb{R}$, interaction network $W_{\theta_2} : \mathcal{X} \to \mathbb{R}$;
Diffusion coefficient $\theta_3 \in \mathbb{R}^+$, time step τ, max inner iterations I_T;
Output:
Learned parameters $\theta = (\theta_1, \theta_2, \theta_3)$ and φ;
while not converged **do**
  for $i \leftarrow 1$ **to** I_T **do**
    `// Update transport maps` T_φ^k
    for $k \leftarrow 0$ **to** $K-1$ **do**
      Sample batch $X_k \sim \rho_k$, $X_{k+1} \sim \rho_{k+1}$;
      $X_{k+1}^{pred} \leftarrow T_\varphi^k(X_k)$;
      $\mathcal{L}_\varphi^k \leftarrow \mathcal{J}_\theta(X_{k+1}^{pred}) + \frac{1}{2\tau}\|X_{k+1}^{pred} - X_{k+1}\|_2^2$;
    $\varphi \leftarrow \varphi - \nabla_\varphi \sum_k \mathcal{L}_\varphi^k$;
  `// Update energy parameters` θ
  for $k \leftarrow 0$ **to** $K-1$ **do**
    Sample batch $X_k \sim \rho_k$;
    $X_{k+1}^{pred} \leftarrow T_\varphi^k(X_k)$;
    $\mathcal{L}_\theta^k \leftarrow -\mathcal{J}_\theta(X_{k+1}^{pred}) + \mathcal{J}_\theta(X_{k+1})$;
  $\theta \leftarrow \theta - \nabla_\theta \sum_k \mathcal{L}_\theta^k$;

use the following equation for estimating $\widehat{\mathcal{H}}(T_\varphi^k \sharp \rho_k)$:

$$\widehat{\mathcal{H}}(T_\varphi^k \sharp \rho_k) = \widehat{\mathcal{H}}(\rho_k) - \int_{\mathcal{X}} \log|\det \nabla_x T_\varphi^k(x)| \, \mathrm{d}\rho_k(x). \tag{25}$$

Since our map is not required to be invertible during training, as is the case for gradient-based ICNN (Amos et al., 2017) parameterizations (Bunne et al., 2022b), we compute the sign and the natural logarithm of the absolute value of the Jacobian determinant, both of which are supported by modern computational frameworks (Bradbury et al., 2018; Paszke, 2019).

C.4 Optimizer

We optimize two loss functions with respect to $\mathcal{J}_\theta$ and T_φ^k using the Adam optimizer (Kingma, 2014). For $\mathcal{J}_\theta$, we use parameters $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\varepsilon = 1 \times 10^{-8}$, and a constant learning rate of 5×10^{-4}, with gradient clipping applied using a maximum global norm of 10. For T_φ^k, we use parameters $\beta_1 = 0.5$, $\beta_2 = 0.9$, and $\varepsilon = 1 \times 10^{-8}$, with a learning rate of 1×10^{-3}. Training is performed with mini-batches of size 500.

C.5 Network architecture

The neural networks for the potential V_{θ_1} and interaction W_{θ_2} energies are multi-layer perceptrons (MLPs) with two hidden layers of size 64, using `softplus` activation functions, and a one-dimensional output layer. The neural network for the optimal transport maps T_φ^k is also an MLP with two hidden layers of size 64. The time step k is concatenated with the input x_k^i, and the network uses `selu` activations, with an output layer that matches the input dimension of x_k^i.

C.6 HARDWARE

Experiments were conducted on a CentOS Linux 7 (Core) system with NVIDIA A100 GPUs. Most of the computation time was spent on metric evaluation. Leveraging JAX's just-in-time compilation, our method completed 100 training epochs in approximately one minute.

Table 7: Training time (in hours) for **EB** (Moon et al., 2019) dataset for 5 and 100 dimensions.

Solver	**Dimension**	mean	std
iJKOnet$_V$	100	0.333	0.020
iJKOnet$_{t,V}$	100	0.338	0.023
iJKOnet$_V$	5	0.321	0.005
iJKOnet$_{t,V}$	5	0.332	0.001

C.7 FUNCTIONALS

For easier representation of our experimental results, we use abbreviations for each potential, each abbreviation being shown in parentheses and the definitions can be found in (Terpin et al., 2024, Appendix F): Wavy Plateau (**WP**), Double Exponential (**DE**), Rotational (**RO**), ReLU (**RE**), Flat (**FT**), Friedman (**FN**), Watershed (**WS**), Ishigami (**IG**), Flowers (**FL**), Bohachevsky (**BC**), Sphere (**SP**), Styblinski-Tang (**ST**), Oakley–Ohagan (**OO**), Zigzag Ridge (**ZR**), and Holder Table (**HT**).

D PROOFS

To begin with, we recall the notion of strong smoothness, see (Beck, 2017, §5.1). A functional $F : \mathcal{X} \to \mathbb{R}$ is called $\frac{1}{\beta}$ - *strongly smooth* if it is continuously differentiable on $\mathcal{X}$ and it holds:

$$\beta\|\nabla F(x) - \nabla F(y)\|_2 \leq \|x - y\|_2 \quad \forall x, y \in \mathcal{X}.$$

Now we proceed to some auxiliary results.

Lemma D.1 (Solution to the JKO problem with potential energy is unique)**.** *Let $\mathcal{J}^*(\rho) = \int_{\mathcal{X}} V^*(x)\mathrm{d}\rho(x)$, $\rho_0 \in \mathcal{P}_{2,ac}(\mathcal{X})$ and $\rho_1 = \mathrm{JKO}_{\tau\mathcal{J}^*}(\rho_0)$. Then $T_{V^*} = \arg\min_{T:\mathcal{X}\to\mathcal{X}} \mathcal{L}(V^*, T)$ is the (unique) optimal transport map between ρ_0 and ρ_1. In particular, $T_{V^*}\sharp\rho_0 = \rho_1$.*

Proof. Consider the functional $\rho \mapsto d^2_{\mathbb{W}_2}(\rho_0, \rho)$, where $\rho \in \mathcal{P}_2(\mathcal{X})$. In what follows, we prove that this functional is strictly convex.

Consider $\rho_a, \rho_b \in \mathcal{P}_2(\mathcal{X})$; $\alpha > 0$. Let $\rho = \alpha\rho_a + (1-\alpha)\rho_b$. Note that ρ_0 is absolutely continuous. By the Brenier's theorem, it holds:

$$d^2_{\mathbb{W}_2}(\rho_0, \rho_a) = \int \|x - T_a(x)\|_2^2 d\rho_0(x) \; ; \quad d^2_{\mathbb{W}_2}(\rho_0, \rho_b) = \int \|x - T_b(x)\|_2^2 d\rho_0(x)$$

for the corresponding (unique) OT maps $T_a : T_a\sharp\rho_0 = \rho_a$; $T_b : T_b\sharp\rho_0 = \rho_b$.

Now consider the weighted sum:

$$\begin{aligned}\alpha d^2_{\mathbb{W}_2}(\rho_0, \rho_a) + (1-\alpha) d^2_{\mathbb{W}_2}(\rho_0, \rho_b) &= \int_{\mathcal{X}} \big(\alpha\|x - T_a(x)\|_2^2 + (1-\alpha)\|x - T_b(x)\|_2^2\big) d\rho_0(x) \\ &= \int_{\mathcal{X}} \|x - y\|_2^2 d\pi_{ab}(x, y),\end{aligned}$$

where plan $\pi_{ab} \in \Pi(\rho_0, \rho)$ has the following conditionals:

$$\pi(y|x) = \alpha\delta\big(y = T_a(x)\big) + (1-\alpha)\delta\big(y = T_b(x)\big).$$

Since the OT map between ρ_0 and ρ is unique (and deterministic), we have

$$d^2_{\mathbb{W}_2}(\rho_0, \rho) = \min_{T\sharp\rho_0 = \rho} \int_{\mathcal{X}} \|x - T(x)\|_2^2 d\rho(x) < \int_{\mathcal{X}} \|x - y\|_2^2 d\pi_{ab}(x, y),$$

which yields the strict convexity of $\rho \mapsto d^2_{\mathbb{W}_2}(\rho_0, \rho)$.

The main statement of the Lemma follows from the strict convexity of $\rho \mapsto \int_{\mathcal{X}} V^*(x)d\rho(x) + d^2_{\mathbb{W}_2}(\rho_0, \rho)$ and Brenier's theorem. □

Now we are ready to prove our main quality bound theorem.

Proof. (Theorem 3.1).

Prelimiary Note: The facts from convex analysis (properties of convex functions and their conjugates) which we use below could be found in (Kornilov et al., 2024, Lemma 1).

At first, we simplify the expression for the JKO loss:

$$\begin{aligned}\mathcal{L}(V, T_V) &= \int V(T_V(x))d\rho_0(x) - \int V(y)d\rho_1(y) + \frac{1}{2\tau}\int \|x - T_V(x)\|_2^2 d\rho_0(x)\\ &= \frac{1}{\tau}\int V_q(T_V(x))d\rho_0(x) - \frac{1}{\tau}\int V_q(y)d\rho_1(y) - \frac{1}{\tau}\int \langle x, T_V(x)\rangle d\rho_0(x)\\ &\quad + \underbrace{\int \frac{\|y\|_2^2}{2\tau}d\rho_1(y) + \int \frac{\|x\|_2^2}{2\tau}d\rho_0(x)}_{\stackrel{\text{def}}{=} C(\rho_0,\rho_1)},\end{aligned} \tag{26}$$

where $V_q = \tau V + \frac{1}{2}\|\cdot\|_2^2$.

From (26) we note that $T_V = \arg\min_T \int \big(V_q(T_V(x)) - \langle x, T_V(x)\rangle\big)d\rho_0(x)$. From the convex analysis, it follows that $T_V(x) = \nabla \overline{V_q}(x)$ delivers the minimum; $\overline{V_q}$ is the (Fenchel) conjugate of V_q. Note that T_V is measurable since it is continuous almost surely (Rockafellar, 1970, Theorem 25.5). Substituting $\nabla\overline{V_q}$ into (26) on par with Fenchel-Young equality yields:

$$\mathcal{L}(V, T_V) = -\frac{1}{\tau}\int \overline{V}_q(x)d\rho_0(x) - \frac{1}{\tau}\int V_q(y)d\rho_1(y) + C(\rho_0, \rho_1). \tag{27}$$

Note that $T_{V^*}\sharp\rho_0 = \rho_1$ (Lemma D.1). Therefore, eq. (26) for $\mathcal{L}(V^*, T_{V^*})$ could be simplified:

$$\begin{aligned}&\mathcal{L}(V^*, T_{V^*}) =\\ &\frac{1}{\tau}\underbrace{\int V_q^*(T_{V^*}(x))d\rho_0(x)}_{=\int V_q^*(y)d\rho_1(y)} - \frac{1}{\tau}\int V_q^*(y)d\rho_1(y) - \frac{1}{\tau}\int \langle x, T_{V^*}(x)\rangle d\rho_0(x) + C(\rho_0,\rho_1) =\\ &-\frac{1}{\tau}\int \langle x, T_{V^*}(x)\rangle d\rho_0(x) + C(\rho_0,\rho_1).\end{aligned} \tag{28}$$

Now we analyze the gap $\varepsilon(V)$ between optimal and optimized JKO losses. Leveraging (27) and (28) yields:

$$\tau\varepsilon(V) = \tau\mathcal{L}(V^*, T_{V^*}) - \tau\mathcal{L}(V, T_V) = \int \overline{V_q}(x)d\rho_0(x) + \int V_q(y)d\rho_1(y) - \int \langle x, T_{V^*}(x)\rangle d\rho_0(x). \tag{29}$$

Now we note that $T_{V^*} = \nabla\overline{V_q^*} : \mathcal{X} \to \mathbb{R}^D$. From the properties of convex functions it follows that $\nabla V_q^* : \mathcal{X} \to \mathbb{R}^D$ defines the *inverse* Optimal Transport mapping. In particular (we assume V_q^* to be convex) $\nabla V_q^*\sharp\rho_1 = \rho_0$. Then, changing the variables in (29) results in:

$$\begin{aligned}\tau\varepsilon(V) &= \int \overline{V_q}(\nabla V_q^*(y))d\rho_1(y) + \int V_q(y)d\rho_1(y) - \int \langle \nabla V_q^*(y), y\rangle d\rho_1(y)\\ &= \int_{\mathcal{X}} \Big[\overline{V_q}(\nabla V_q^*(y)) + V_q(y) - \langle \nabla V_q^*(y), y\rangle\Big]d\rho_1(y). \qquad (30)\\ &= \int_{\mathcal{X}} \mathcal{D}_{\overline{V_q}}(\nabla V_q^*(y), \nabla V_q(y))d\rho_1(y), \qquad (31)\end{aligned}$$

where $\mathcal{D}_{\overline{V_q}}(\,\cdot\,,\,\cdot\,)$ is the *Bregman* divergence, see (Banerjee et al., 2005, Def. 1) for the definition and (Banerjee et al., 2005, Appendix A) for the "Dual Divergences" property used in the transition from (30) to (31).

Since V_q is $\frac{1}{\beta}$-smooth, then $\overline{V_q}$ is β-strongly convex. Therefore, by the property of strongly convex functions (we pick $y \in \text{supp}(\rho_1)$):

$$\begin{aligned}\frac{\beta}{2}\|\nabla V_q^*(y) - \nabla V_q(y)\|_2^2 &\leq \overline{V_q}(\nabla V_q^*(y)) - \overline{V_q}(\nabla V_q(y)) - \langle \nabla \overline{V_q}(\nabla V_q(y)), \nabla \overline{V_q^*}(y) - \nabla V_q(y)\rangle \\ &= \mathcal{D}_{\overline{V_q}}(\nabla V_q^*(y), \nabla V_q(y)). \end{aligned} \tag{32}$$

Combining (31) and (32) yields:

$$\tau\varepsilon(V) \geq \frac{\beta}{2}\int_{\mathcal{X}} \|\nabla V_q^*(y) - \nabla V_q(y)\|_2^2 d\rho_1(y).$$

To complete the proof, we are left to note that $\nabla V_q^*(y) - \nabla V_q(y) = \tau\big(\nabla V^*(y) - \nabla V(y)\big)$. □