Gradient-Flow SDEs Have Unique Transient Population Dynamics

Vincent Guan
University of British Columbia

Joseph Janssen
University of British Columbia

Nicolas Lanzetti
ETH Zürich

Antonio Terpin
ETH Zürich

Geoffrey Schiebinger
University of British Columbia

Elina Robeva
University of British Columbia

Abstract

Identifying the drift and diffusion of an SDE from its population dynamics is a notoriously challenging task. Researchers in machine learning and single-cell biology have only been able to prove a partial identifiability result: for potential-driven SDEs, the gradient-flow drift can be identified from temporal marginals if the Brownian diffusivity is already known. Existing methods therefore assume that the diffusivity is known a priori, despite it being unknown in practice. We dispel the need for this assumption by providing a complete characterization of identifiability: the gradient-flow drift and Brownian diffusivity are jointly identifiable from temporal marginals if and only if the process is observed outside of equilibrium. Given this fundamental result, we propose `nn-APPEX`, the first Schrödinger Bridge–based inference method that can simultaneously learn the drift and diffusion of a gradient-flow SDE solely from observed marginals. Extensive experiments show that `nn-APPEX`'s ability to adjust its diffusion estimate enables accurate inference, while previous Schrödinger Bridge methods obtain biased drift estimates due to their assumed, and likely incorrect, diffusion.

1 INTRODUCTION

Mathematical biologists conceptualize the evolving genetic profiles of cells, or *developmental trajectories*, with Waddington's epigenetic landscape, which likens differentiating cells to marbles rolling along a surface (Waddington, 1935). Cell development is therefore thought to be driven by the gradient of this unknown landscape. More broadly, gradient-driven dynamics describe a host of real-world processes, such as contaminant flow within groundwater systems, electric current within an electromagnetic field, and molecular dynamics driven by interatomic potential energy. Since these processes exhibit stochasticity, they are often modeled by *gradient-flow* stochastic differential equations (SDEs), such that the drift is given by $-\nabla\Psi(X_t)$, and stochasticity is introduced by Brownian motion W_t, with diffusivity $\sigma^2 > 0$ (Weinreb et al., 2018; Lavenant et al., 2024; Lelievre and Stoltz, 2016):

$$\mathrm{d}X_t = -\nabla\Psi(X_t)\mathrm{d}t + \sigma\mathrm{d}W_t. \tag{1}$$

While an SDE's drift and diffusion can generally be inferred from trajectories (Nielsen et al., 2000; Bishwal, 2007; Browning et al., 2020; Wang et al., 2024), it is often only possible to observe population dynamics. For example, hydrogeochemical sensors can only detect plumes, rather than particle trajectories, and scRNA sequencing technologies destroy cells upon measurement (Trapnell et al., 2014). This limits observations to temporal snapshots of the marginals $\big(p(\cdot,t)\big)_{t\in[0,T]}$, and raises the fundamental question:

> Q: *Under what conditions are* $-\nabla\Psi$ *and* σ^2 *identifiable from marginals* $\big(p(\cdot,t)\big)_{t\in[0,T]}$*?*

Despite receiving significant attention from researchers in mathematical biology and machine learning (Hashimoto et al., 2016; Weinreb et al., 2018; Neklyudov et al., 2023; Lavenant et al., 2024; Yeo et al., 2021; Zhang et al., 2021; Chizat et al., 2022; Bunne et al., 2022; Shen et al., 2025; Terpin et al., 2024), the most comprehensive identifiability results only prove that the drift $-\nabla\Psi$ can be identified from $\big(p(\cdot,t)\big)_{t\in[0,T]}$ if the diffusivity σ^2 is already known:

$$\text{observe } \big(p(\cdot,t)\big)_{t\in[0,T]} \text{ and } \sigma^2 \text{ is known} \implies -\nabla\Psi \text{ is identifiable.} \tag{2}$$

Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300.

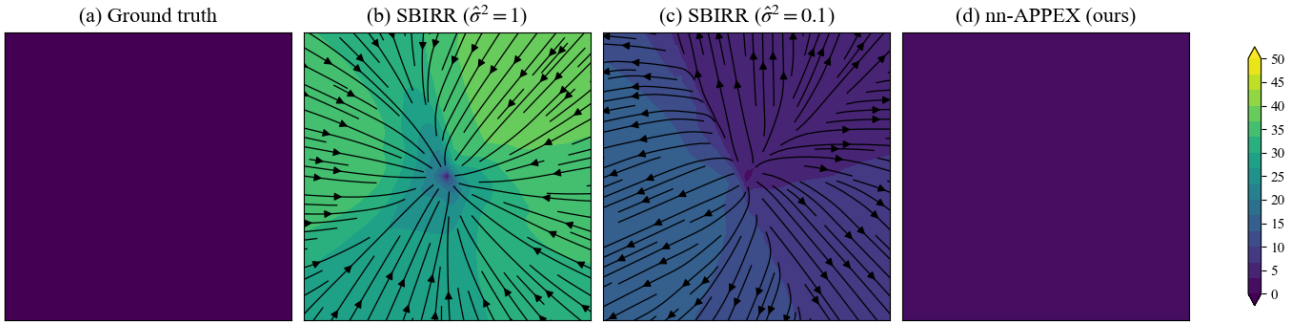


Figure 1: The true drift field (a) and estimated drift fields (b)-(d) are shown for the simple example of a Brownian motion, $\mathrm{d}X_t = \sqrt{0.2}\mathrm{d}W_t$. The current state-of-the-art Schrödinger Bridge method `SBIRR` (Shen et al., 2025; Zhang, 2024) presumes prior knowledge ($\hat{\sigma}^2$) of the diffusivity σ^2 instead of inferring it from data. Figure 1(b) shows that it may wrongly infer a compressive drift force if $\hat{\sigma}^2 > \sigma^2$, while Figure 1(c) shows that it may wrongly infer an expanding drift force if $\hat{\sigma}^2 < \sigma^2$. Figure 1(d) shows that by iteratively learning the diffusion as well as the drift, our method `nn-APPEX` can accurately infer drift without knowing diffusion a priori.

This result, popularized in the mathematical biology community by Weinreb et al. (2018) and in the machine learning community by Hashimoto et al. (2016), has informed the predominant view that the only way to achieve principled inference is to assume that the diffusion is already known. As a result of this theory, virtually all inference methods have been developed to estimate $-\nabla\Psi$ given a known diffusivity σ^2 (Hashimoto et al., 2016; Neklyudov et al., 2023; Lavenant et al., 2024; Yeo et al., 2021; Zhang et al., 2021; Chizat et al., 2022; Bunne et al., 2022; Shen et al., 2025; Zhang, 2024). While the drift is often the principal object of interest, the diffusion is also typically unknown, and contains important insights in its own right, such as the sensitivity of cell fates to initial conditions (Forrow, 2024). Perhaps most importantly, misspecified diffusion can significantly bias drift estimation. We provide a simple example in Figure 1, where the true landscape is flat, but overestimated diffusion introduces a sink and underestimated diffusion introduces a source.

Contributions. It has remained an open question whether it is possible and, if so, under which circumstances, *both* the drift and diffusion of a gradient-flow SDE (1) can be identified from its marginals. Our first theoretical contribution is a complete solution to this problem. Theorem 3.2 provides the *necessary and sufficient conditions* for identifying the gradient flow drift and diffusivity from marginals:

$$\text{observe non-stationary } \big(p(\cdot,t)\big)_{t\in[0,T]} \iff -\nabla\Psi \text{ and } \sigma^2 \text{ are identifiable.} \tag{3}$$

We also extend this result beyond the setting of continuous observation $t \in [0,T]$, by proving in Corollary 3.3 that observing three distinct marginals identifies the true SDE from any countable set of candidates, with probability 1. Since our identifiability conditions are commonly observed in practice, our results provide:

1. A theoretical foundation for jointly inferring drift and diffusion directly from marginals.
2. A call to action to move beyond the conventional wisdom that diffusion should be specified a priori for principled inference from population dynamics.

With our identifiability theory guaranteeing that joint inference is well-posed, we propose `nn-APPEX`, the first Schrödinger Bridge (SB) based method capable of estimating both the gradient flow drift $-\nabla\Psi$ and the diffusivity σ^2. In particular, `nn-APPEX` introduces a diffusion estimation step at each iteration while utilizing a neural network to flexibly infer the drift field. To assess the importance of `nn-APPEX`'s diffusion estimation step, we compare its performance against previous SB methods on simulated benchmark data (Terpin et al., 2024; Persiianov et al., 2025) and on a single-cell dataset from human embryonic stem cells (Chu et al., 2016; Shen et al., 2025). The results corroborate our identifiability theory, while also highlighting the importance of learning the correct diffusion for accurate drift estimation, as `nn-APPEX` markedly outperforms previous SB methods.

2 MATHEMATICAL SETUP

As is standard in stochastic analysis (Shen et al., 2025; Berlinghieri et al., 2025), we ensure that the gradient-flow SDE (1) is well-defined by assuming that the drift $-\nabla\Psi$ satisfies Lipschitz continuity and linear growth, $\|\nabla\Psi(x)\| \leq K(1 + \|x\|)$ for some $K > 0$ (Oksendal, 2013, Theorem 5.5). Given these conditions, and an

initial distribution $p(\cdot, 0)$ with finite second moments, the population dynamics are defined by the Fokker-Planck equation. For all $x \in \mathbb{R}^d, t \geq 0$,

$$\begin{aligned} \frac{\partial p(x,t)}{\partial t} &= \mathcal{L}^*_{-\nabla\Psi,\sigma^2}(p(\cdot,t))(x) \\ &\coloneqq \nabla \cdot (p(x,t)\nabla\Psi(x)) + \frac{\sigma^2}{2}\Delta p(x,t). \end{aligned} \tag{4}$$

While the Fokker-Planck equation (4) may not be defined pointwise if the arguments are not sufficiently smooth, it can be defined in the weak distributional sense, such that $\mathcal{L}^*_{-\nabla\Psi,\sigma^2}$ is an operator on probability distributions (see Appendix A.1). We therefore adopt the weak formulation and use standard notation by omitting x (Bogachev et al., 2022), e.g., $\mathcal{L}^*_{-\nabla\Psi,\sigma^2}\big(p(\cdot,t)\big)$ rather than $\mathcal{L}^*_{-\nabla\Psi,\sigma^2}\big(p(\cdot,t)\big)(x)$.

We now rigorously define identifiability of gradient-flow SDEs, given continuous observation of their marginals. Intuitively, an SDE with parameters $(-\nabla\Psi_1, \sigma_1^2)$ is identifiable from its marginals if and only if no other SDE, with distinct parameters $(-\nabla\Psi_2, \sigma_2^2) \neq (-\nabla\Psi_1, \sigma_1^2)$, can produce the same probability flow $\frac{\partial p}{\partial t}$.

Definition 2.1 (Identifiability). *The SDE (1) with parameters $(-\nabla\Psi_1, \sigma_1^2)$ is identifiable from its marginals $\big(p(\cdot,t)\big)_{t\in[0,T]}$ if and only if, $\forall(-\nabla\Psi_2, \sigma_2^2)$,*

$$\begin{aligned} &\mathcal{L}^*_{-\nabla\Psi_1,\sigma_1^2}\big(p(\cdot,t)\big) = \mathcal{L}^*_{-\nabla\Psi_2,\sigma_2^2}\big(p(\cdot,t)\big) \qquad \forall t \in [0,T] \\ &\implies (-\nabla\Psi_1, \sigma_1^2) = (-\nabla\Psi_2, \sigma_2^2), \end{aligned}$$

where the first equality holds in the distributional sense and the second equality holds pointwise in $\mathbb{R}^d \times \mathbb{R}^+$.

The following example (Lavenant et al., 2024; Guan et al., 2024) shows that gradient-flow SDEs (1) are not always identifiable from their marginals $\big(p(\cdot,t)\big)_{t\in[0,T]}$.

Example 2.2 (Non-identifiability at the stationary distribution). *Consider two SDEs with quadratic potential, i.e., Ornstein-Uhlenbeck processes,*

$$\mathrm{d}X_t = -X_t\mathrm{d}t + \mathrm{d}W_t \qquad X_0 \sim p_0 \tag{5}$$

$$\mathrm{d}Y_t = -10Y_t\mathrm{d}t + \sqrt{10}\mathrm{d}W_t \qquad Y_0 \sim p_0, \tag{6}$$

with $p_0 = \mathcal{N}(0, \frac{1}{2})$, i.e., Gaussian with mean 0 and variance $\frac{1}{2}$. Since p_0 is the stationary distribution for both SDEs, then $X_t, Y_t \sim p(\cdot,t) = p_0$ for all $t \geq 0$, which makes the SDEs non-identifiable from marginals.

3 IDENTIFIABILITY RESULTS

In this section, we present our main result in Theorem 3.2, which states that the gradient-flow SDE (1) is identifiable from its marginals, $\big(p(\cdot,t)\big)_{t\in[0,T]}$, *if and only if* it is observed outside of equilibrium. We then show in Corollary 3.3 that three distinct marginals identify the true SDE with probability 1 from any countable set of candidate SDEs.

We first prove that non-stationarity is a necessary condition for identifiability, by extending Example 2.2 for arbitrary potentials.

Proposition 3.1 (Non-identifiability from equilibrium). *If p_{eq} is a stationary distribution for the SDE (1), then it is also a stationary distribution for the "rescaled" SDE*

$$\mathrm{d}X_t = -\alpha\nabla\Psi(X_t)\mathrm{d}t + \sqrt{\alpha}\sigma\mathrm{d}W_t, \tag{7}$$

for any $\alpha > 0$.

Proof. If p_{eq} is stationary for $\mathcal{L}^*_{-\nabla\Psi,\sigma^2}$, then

$$\begin{aligned} 0 &= \mathcal{L}^*_{-\nabla\Psi,\sigma^2}(p_{\mathrm{eq}}) \\ &= \nabla \cdot (p_{\mathrm{eq}}\nabla\Psi) + \frac{\sigma^2}{2}\Delta p_{\mathrm{eq}} \\ \Leftrightarrow 0 &= \alpha\left(\nabla \cdot (p_{\mathrm{eq}}\nabla\Psi + \frac{\sigma^2}{2}\Delta p_{\mathrm{eq}})\right) \\ &= \nabla \cdot (p_{\mathrm{eq}}\nabla(\alpha\Psi)) + \frac{(\sqrt{\alpha}\sigma)^2}{2}\Delta p_{\mathrm{eq}} \\ &= \mathcal{L}^*_{-\nabla(\alpha\Psi),\alpha\sigma^2}(p_{\mathrm{eq}}). \end{aligned}$$

Thus, given the initialization $p(\cdot,0) = p_{\mathrm{eq}}$, we would observe $p(\cdot,t) = p_{\mathrm{eq}}\ \forall t \geq 0$ for both processes. □

This shows that the SDE (1) is non-identifiable from its marginals $\big(p(\cdot,t)\big)_{t\in[0,T]}$ if we observe complete stationarity, $p(\cdot,t) = p(\cdot,0)\ \forall t \geq 0$. We now prove that this is the only source of non-identifiability.

Theorem 3.2 (Identifiability of gradient-flow SDEs). *The SDE (1) is non-identifiable from its marginals $\big(p(\cdot,t)\big)_{t\in[0,T]}$ if and only if $p(\cdot,t) = p(\cdot,0)\ \forall t \geq 0$.*

Proof. Proposition 3.1 proves that if $p(\cdot,t) = p(\cdot,0)$ $\forall t \geq 0$ then the SDE is non-identifiable.

Now, we prove that non-identifiability implies that $p(\cdot,t) = p(\cdot,0)\ \forall t \geq 0$. Suppose that SDEs with parameters $(-\nabla\Psi_1, \sigma_1^2)$ and $(-\nabla\Psi_2, \sigma_2^2)$ produce the same marginals $\big(p(\cdot,t)\big)_{t\in[0,T]}$. Without loss of generality, we can assume that $\sigma_1^2 > \sigma_2^2$, because the previous partial result (2) states that if $\sigma_1^2 = \sigma_2^2$, then $\nabla\Psi_1 = \nabla\Psi_2$ also follows (Lavenant et al., 2024, Theorem 2.1). Then, from Definition 2.1, for all observed marginals $p(\cdot,t)$,

$$\mathcal{L}^*_{-\nabla\Psi_1,\sigma_1^2}(p(\cdot,t)) = \mathcal{L}^*_{-\nabla\Psi_2,\sigma_2^2}(p(\cdot,t)). \tag{8}$$

By the linearity of the Fokker-Planck operator, we have

$$0 = (\mathcal{L}^*_{-\nabla\Psi_1,\sigma_1^2} - \mathcal{L}^*_{-\nabla\Psi_2,\sigma_2^2})(p(\cdot,t)) \tag{9}$$

$$= \nabla \cdot \big(p(\cdot,t)\nabla(\Psi_1 - \Psi_2)\big) + \frac{\sigma_1^2 - \sigma_2^2}{2}\Delta p(\cdot,t). \tag{10}$$

We note that the PDE (10) is the Fokker-Planck equation, parametrized by the "residual" drift and diffusion. Then, to prove that $p(\cdot,t) = p(\cdot,0)\ \forall t \geq 0$, it suffices to show that there is at most one probability distribution, $p(\cdot,t) = \mu$, which solves (10), since each marginal would then coincide. Since the residual diffusivity $\sigma_1^2 - \sigma_2^2 > 0$ is nondegenerate, and the residual drift $-\nabla(\Psi_1 - \Psi_2)$ is Lipschitz and obeys linear growth, it follows that the residual Fokker-Planck equation (10) has at most one stationary distribution (Bogachev et al., 2022, Theorem 4.1.6, Example 4.1.8). □

By reducing non-identifiability to the uniqueness of solutions to an elliptic PDE, our proof provides a fundamental insight about gradient-flow SDEs: *observing non-identifiability between two processes is equivalent to observing a residual process at equilibrium.* If the SDE parameters are time-homogeneous, then each marginal $p(\cdot,t)$ coincides with the unique stationary distribution. Even if the parameters are time-inhomogeneous, non-identifiability is still characterized by equilibrium. However, equilibrium would not imply a static set of marginals, since equilibrium itself may change over time. We provide an example in Example B.3, where the stationary distributions of the residual process coincide with Brownian marginals $\mathcal{N}(0,t)$. In Appendix B, we also prove sufficient conditions for identifiability in the time-inhomogeneous case, when the SDE parameters change at discrete times, a setting commonly observed in cell development (Monnier et al., 2012).

While Proposition 3.1 shows that either the diffusivity σ^2 or the potential Ψ needs to be known to disambiguate the true SDE from other SDEs with the same stationary distribution, Theorem 3.2 shows that this assumption is not required if the marginals are transient. This is important for real-world inference since the condition $p(\cdot,t) = p(\cdot,0)\ \forall t \geq 0$ is easily verifiable, and it is much more common to observe transient behaviour than it is to know an accurate a priori estimate of the diffusivity coefficient (Forrow, 2024). However, Theorem 3.2 still relies on continuous observation, while we only observe a finite number of marginals in practice. We address this by showing that observing three distinct marginals is enough to identify the true SDE from any countable set of candidates, e.g. the set of SDEs that can be represented on a computer. To facilitate the proof, we assume that $\Psi \in C^\infty(\mathbb{R}^d)$ (see details in Appendix A.2).

Corollary 3.3 (Identifiability from three marginals)**.** *Let $\mathcal{S}$ be a* countable *set of gradient-flow SDEs with smooth potentials, which all share the same initial distribution $p(\cdot,0)$. If we observe distinct marginals $\{p(\cdot,t_i)\}_{i=1}^3$ from an SDE in $\mathcal{S}$, such that the measurement times $\{t_i\}_{i=1}^3$ are uniformly sampled, i.e., $t_i \sim \mathrm{Unif}[T_i, T_{i+1}]$, for some $0 < T_1 < T_2 < T_3 < T_4$, then, with probability* 1*, this is the only SDE in $\mathcal{S}$ with marginals $\{p(\cdot,t_i)\}_{i=1}^3$.*

Proof. By the countability of $\mathcal{S}$, it suffices to prove that the event that any two distinct SDEs in $\mathcal{S}$ share the same marginals at each of the times $\{t_i\}_{i=1}^3$ has probability 0. Denote their parameters by $(-\nabla\Psi_1, \sigma_1^2) \neq (-\nabla\Psi_2, \sigma_2^2)$ and their marginals by $\big(p(\cdot,t)\big)_{t\geq 0}$ and $\big(q(\cdot,t)\big)_{t\geq 0}$. Suppose for contradiction that if we sample $t_i \sim \mathrm{Unif}[T_i, T_{i+1}]$, then we observe $p(\cdot,t_i) = q(\cdot,t_i)$ $\forall i \in \{1,2,3\}$, with some nonzero probability.

Since we sample $t_i \sim \mathrm{Unif}[T_i, T_{i+1}]$, then the set of coincidence times in each subinterval, $\mathcal{I}_i = \{t \in [T_i, T_{i+1}] \mid p(\cdot,t) = q(\cdot,t)\}$, must be infinite to ensure nonzero probability. By Bolzano-Weierstrass, we can construct the convergent sequence $t_1^{(n)} \to t_1^* \in [T_1, T_2]$, using times $t_1^{(n)} \in \mathcal{I}_1$, and the convergent sequence $t_2^{(n)} \to t_2^* \in [T_2, T_3]$, using times $t_2^{(n)} \in \mathcal{I}_2$. At the limit points, t_1^* and t_2^*, we have $p(\cdot,t_i^*) = q(\cdot,t_i^*)$ and $\frac{\partial}{\partial t}p(\cdot,t_i^*) = \frac{\partial}{\partial t}q(\cdot,t_i^*)$ (see Lemma A.2 and the proof of Hashimoto et al. (2016, Corollary 2)). Hence,

$$\mathcal{L}^*_{-\nabla\Psi_1,\sigma_1^2}(p(\cdot,t_1^*)) = \mathcal{L}^*_{-\nabla\Psi_2,\sigma_2^2}(p(\cdot,t_1^*)) \tag{11}$$

$$\mathcal{L}^*_{-\nabla\Psi_1,\sigma_1^2}(p(\cdot,t_2^*)) = \mathcal{L}^*_{-\nabla\Psi_2,\sigma_2^2}(p(\cdot,t_2^*)) \tag{12}$$

It then follows from Theorem 3.2 that (11) and (12) are both solved by at most one distribution μ. Thus, $\mu = p(\cdot,t_1^*) = p(\cdot,t_2^*)$. However, by Proposition A.3, the only marginal that can repeat at distinct times is the stationary Gibbs distribution p_{eq}. This follows from a variation of Boltzmann's H-theorem, which states that the marginals of a gradient-flow SDE (1) decrease free energy, which is uniquely minimized at p_{eq} (Jordan et al., 1998; Chafaï, 2015). Both processes must therefore reach p_{eq} by t_1^*, such that $p(\cdot,t) = q(\cdot,t) = p_{\mathrm{eq}}$ for all $t \geq t_1^*$. In particular, since t_1^* is in the first subinterval, we have $p(\cdot,t_2) = p(\cdot,t_3) = p_{\mathrm{eq}}$, which contradicts the assumption that the marginals $\{p(\cdot,t_i)\}_{i=1}^3$ are distinct. □

4 NN-APPEX: THREE-STAGE SB REFINEMENT

We have shown in Section 3 that the full gradient-flow SDE is identifiable from $N \geq 3$ distinct marginals, $\big(p(\cdot,t_i)\big)_{i=0}^{N-1}$. Building on this, we propose Neural Network-based Alternating Projection Parameter Estimation from X_0 (`nn-APPEX`). `APPEX` searches for the true parameters $(-\nabla\Psi, \sigma^2)$ by alternating between trajectory inference, drift estimation, and diffusion estimation. `nn-APPEX` improves upon `APPEX` (Guan et al., 2024) by replacing the linearly parameterized drift estimation step with a neural network, such that

it can infer SDEs with nonlinear drift, e.g. arbitrary gradient-flow $-\nabla\Psi$.

The three-stage optimization proceeds as follows. First, given a reference SDE Q, we solve a multi-marginal Schrödinger Bridge problem to infer a law on paths P. In particular, P minimizes $\mathrm{KL}(P\|Q)$, while obeying the multi-marginal constraints, i.e., $P \in \Pi\big(p(\cdot,t_i)_{i=0}^{N-1}\big)$. Second, from these inferred trajectories, we perform a maximum likelihood estimation (MLE) of the drift $-\nabla\Psi$. For generality and tractability, we optimize over neural network parameters θ. Third, from the inferred trajectories and the estimated drift, we estimate the diffusivity σ^2 with a closed form MLE expression. We then use the new drift and diffusion estimates to update the reference SDE Q for the next iteration. Formally, at iteration k, `nn-APPEX` performs the steps

$$P_k = \underset{P\in\Pi\left(p(\cdot,t_i)_{i=0}^{N-1}\right)}{\arg\min} \mathrm{KL}(P\|Q_{-\nabla\Psi_{\theta_{k-1}},\sigma^2_{k-1}}), \tag{13}$$

$$\theta_k = \underset{\theta}{\arg\min}\,\mathrm{KL}(P_k\|Q_{-\nabla\Psi_{\theta},\sigma^2_{k-1}}), \tag{14}$$

$$\sigma^2_k = \underset{\sigma^2}{\arg\min}\,\mathrm{KL}(P_k\|Q_{-\nabla\Psi_{\theta_k},\sigma^2}), \tag{15}$$

and we iterate until convergence, or until a maximum number of iterations is reached. We emphasize that `nn-APPEX` does not require prior knowledge, since any gradient-flow SDE (1) can be used as the initial reference $Q_{-\nabla\Psi_{\theta_0},\sigma_0^2}$. While it is known that the KL divergence on the space $C([0,T],\mathbb{R}^d)$ is infinite between SDEs with different diffusions (Shen et al., 2025), we note that given finite observation times, $\{t_i\}_{i=0}^{N-1}$, we can instead evaluate the KL divergence on $C(\{t_i\}_{i=0}^{N-1},\mathbb{R}^d)$, computed over the observed couplings $p_{t_i,t_{i+1}}$. As shown in Guan et al. (2024, Section 4) and Appendix D, this objective is finite for gradient-flow SDEs (1), and thus enables principled three-stage optimization. We present pseudocode in Algorithm 1.

The SB problem (13) is equivalent to entropic optimal transport with an adjusted cost, and can thus be solved with a variety of iterative proportional fitting algorithms, such as Sinkhorn's algorithm (Peyré et al., 2019). By default, we use the multi-marginal Sinkhorn algorithm (Marino and Gerolin, 2020). For MLE parameter estimation on the inferred paths, we use the Euler-Maruyama approximation, and present derivations in Appendix E. In particular, the optimal drift parameters $\hat\theta$ are independent of σ^2, since minimizing the negative log-likelihood with respect to θ is equivalent to minimizing

$$\ell(\theta) = \sum_{m=1}^{M}\sum_{i=0}^{N-2}\left\|x_{i+1}^{(m)} - x_i^{(m)} + \nabla\Psi_\theta(x_i^{(m)})\Delta t\right\|_2^2, \tag{16}$$

where M is the number of paths, N is the number of time steps, and Δt is the time step. We therefore estimate $-\nabla\Psi_{\hat\theta}$ by fitting a neural network whose parameters minimize $\ell(\theta)$, as done in Shen et al. (2025). The diffusion MLE, $\hat\sigma^2$, is then derived from the quadratic variation, conditioned on the estimated drift $\nabla\Psi_{\hat\theta}$,

$$\hat\sigma^2 = \frac{1}{d\,M\,(N-1)\,\Delta t}\ell(\hat\theta). \tag{17}$$

Algorithm 1 `nn-APPEX`

1: **Input:** Observed marginals $\hat p_{t_i}$, $i=0,\dots,N-1$, number of iterations K, time step Δt
2: **Initialize:** $\nabla\hat\Psi \leftarrow 0$, $\hat\sigma^2 \leftarrow 1$
3: **for** $k=1,\dots,K$ **do**
4: $\quad\{\mathcal{K}_i(x,y)\}_{i=0}^{N-2} \leftarrow \exp\Big(-\frac{\|y-x+\nabla\hat\Psi(x)\,\Delta t\|^2}{2\hat\sigma^2\Delta t}\Big)$
5: $\quad\{\pi_i\}_{i=0}^{N-2} \leftarrow \textsc{MultiSinkhorn}(\{\hat p_{t_i}\},\{\mathcal{K}_i\})$
6: $\quad\{\hat\tau^{(s)}\}_{s=1}^{S} \leftarrow \textsc{SamplePaths}(\{\pi_i\})$
7: $\quad\nabla\hat\Psi \leftarrow \textsc{NNTrain}\big(\{\hat\tau^{(s)}\}\big)$
8: $\quad\hat\sigma^2 \leftarrow \textsc{MLEDiffusivity}\Big(\{\hat\tau^{(s)}\},\nabla\hat\Psi\Big)$
9: **end for**
10: **return** $\nabla\hat\Psi$, $\hat\sigma^2$

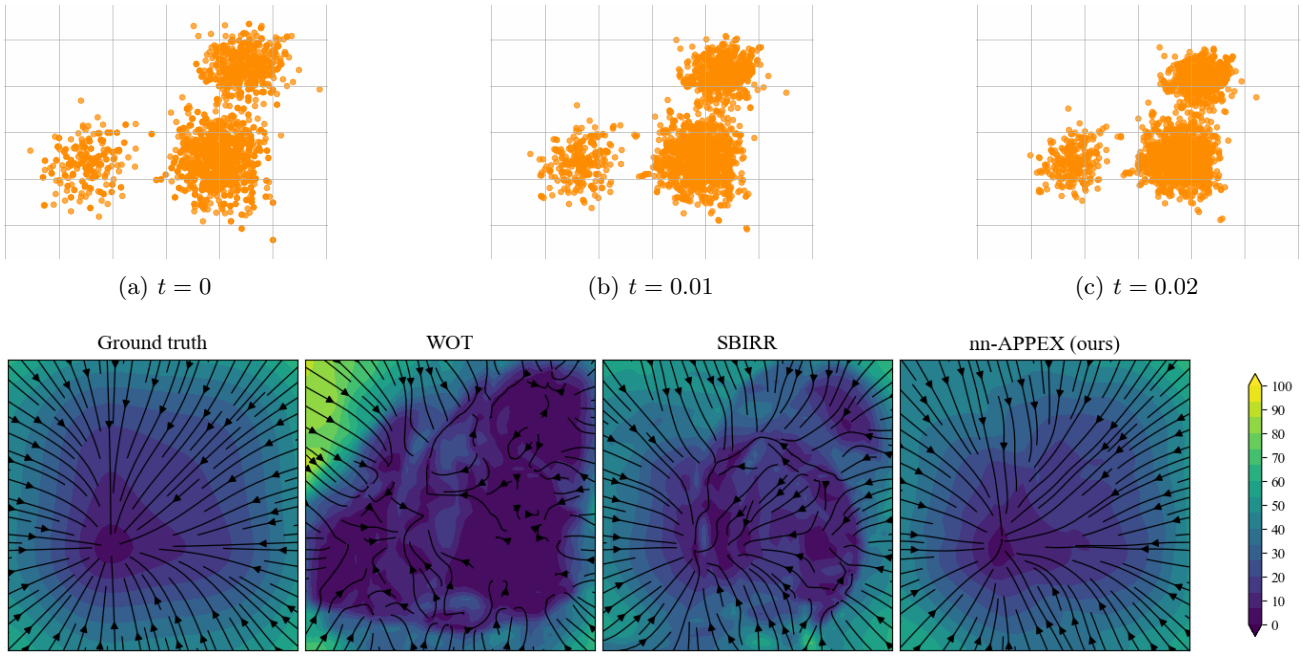
By alternating between these three procedures, `nn-APPEX` iteratively reduces the divergence between *reconstructed paths*, $P \in \Pi\left(p(\cdot,t_i)_{i=0}^{N-1}\right)$, which obey the observed marginals, and *paths from the estimated gradient-flow SDE* with law Q:

$$\begin{aligned}\mathrm{KL}(P_{k+1}\|Q_{-\nabla\Psi_{k+1},\sigma^2_{k+1}}) &\le \mathrm{KL}(P_{k+1}\|\,Q_{-\nabla\Psi_{k+1},\sigma^2_k})\\ &\le \mathrm{KL}(P_{k+1}\|Q_{-\nabla\Psi_k,\sigma^2_k}) \le \mathrm{KL}(P_k\|Q_{-\nabla\Psi_k,\sigma^2_k}).\end{aligned} \tag{18}$$

By decreasing the relative divergence, `nn-APPEX` iteratively approaches the unique true solution, though it is unclear whether the algorithm can stagnate at a different set of parameters. We conjecture that convergence to the true solution holds given infinite data, and leave the proof for future work.

To highlight distinctions with existing SB methods, we note relative differences in algorithmic design:

- Waddington-OT (`WOT`) (Schiebinger et al., 2019) fixes the reference Q to be a Brownian motion with diffusivity σ^2 and stops after (14), i.e., fix $K=1$ in Algorithm 1.
- `SBIRR` (Shen et al., 2025; Zhang, 2024) fixes σ^2 and iteratively performs steps (13) and (14), i.e., exclude line 8 in Algorithm 1.
- `APPEX` (Guan et al., 2024) iteratively performs all three steps, but only considers linear drift, i.e., replace line 7 in Algorithm 1 with a closed form MLE for linear SDEs.



(a) $t = 0$ (b) $t = 0.01$ (c) $t = 0.02$

(d) True and estimated landscapes

Figure 2: We simulate gradient-flow SDEs from a variety of potentials and provide inference methods with samples from three distinct marginals, initialized from a random Gaussian mixture model. Data for one seed is plotted for the Oakley–O'Hagan potential in (a)–(c), along with the true and estimated landscapes in (d).

Each method reconstructs paths by solving a Schrödinger Bridge problem. However, these paths will be incorrect unless the reference SDE matches the true SDE. While `nn-APPEX` has the flexibility to re-estimate the gradient-flow drift as well as the diffusivity, previous methods cannot reliably determine a suitable reference SDE. `WOT`, and related methods (Lavenant et al., 2024; Forrow and Schiebinger, 2021; Chizat et al., 2022), fix a pure Brownian reference, which is incompatible with any nonzero potential. Although `SBIRR` adjusts its reference drift, if σ^2 is misspecified, then it would also misspecify the reference SDE, hindering its ability to learn the drift. While `APPEX` adjusts drift and diffusion, it estimates a misspecified SDE, unless the ground-truth potential Ψ happens to be quadratic.

5 SIMULATED EXPERIMENTS

In this section, we perform extensive experiments on simulated data to evaluate `nn-APPEX` against previous Schrödinger Bridge methods, `WOT` and `SBIRR`. The code repository is available on GitHub: `https://github.com/guanton/identifying-gradient-flow-sdes`.

5.1 Experimental Setup

We explain details about the data, method implementation, and performance metrics below.

Data generation. To simulate gradient-flow SDEs, we fix the diffusivity $\sigma^2 = 0.2$ and consider five potentials that are commonly used as benchmarks in the literature (Terpin et al., 2024; Persiianov et al., 2025). Potentials were chosen to ensure the existence of a stationary distribution $p_{\text{eq}} = \frac{1}{Z}\exp\left(-\frac{\Psi}{2\sigma^2}\right)$. See Appendix F for the list of potentials. For our main experiments, we initialize the first marginal as a random Gaussian mixture model (GMM), due to their universal approximation properties (McLachlan and Rathnayake, 2014) and their applicability for clustering scRNA data into different cell types (Yu et al., 2021). We uniformly sample the number of components between 1 and 10, their means within $[-3, 3]^2$, and their variances between 0.5 and 1.0. We then mirror the "three marginals" identifiability setting from Corollary 3.3, by sampling $N = 2000$ points from the GMM initial distribution, and then forward simulating these points using the Euler-Maruyama scheme for an additional 2 time steps of size $\Delta t = 0.01$. Experiments are repeated across 10 seeds for each SDE to ensure replicability. Marginals for one seed are pictured for the SDE driven by the Oakley-O'Hagan potential in Figure 2(a)-(c).

Methods. To perform a systematic ablation that isolates the effects of iterative reference refinement and adaptive diffusion on inference, all methods are given

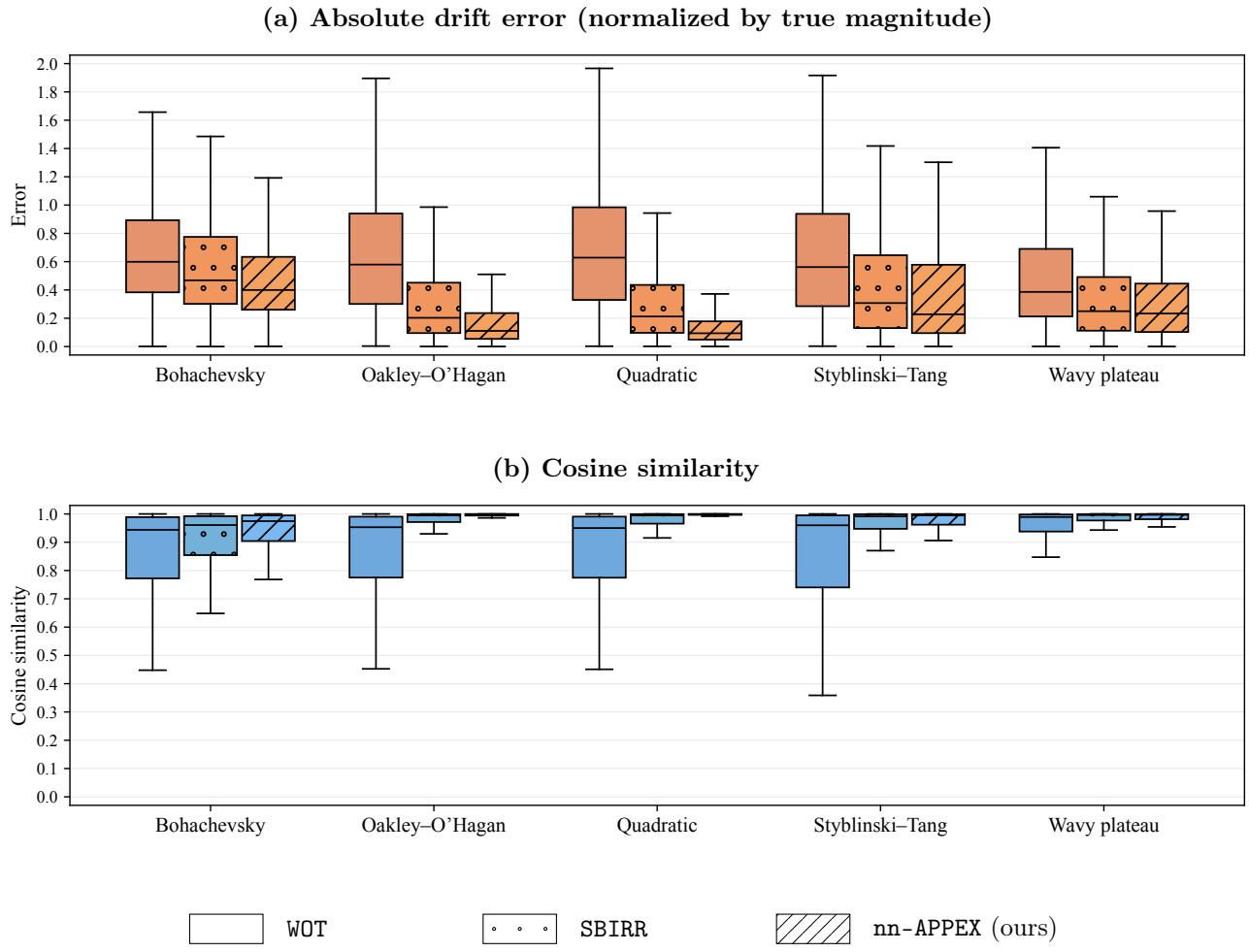


Figure 3: The ability of different Schrödinger Bridge methods to infer the gradient-flow drift is evaluated across five different potentials using (a) normalized absolute error (lower is better) and (b) cosine similarity (higher is better). Methods are given samples from three distinct marginals, such that the initial distribution is a Gaussian mixture model with randomly initialized components. The box-and-whisker plots aggregated from 10 seeds show that our method, `nn-APPEX`, performs the best across all potentials.

the same data and initialization, and any shared subprocedure is implemented identically across methods. To emulate the realistic scenario where the practitioner can only estimate the true diffusivity σ^2 up to an order of magnitude (Guan et al., 2024), we sample the diffusivity prior as $\sigma^2_{\text{prior}} \sim \sigma^2 \times 10^{\text{Unif}[-1,1]}$, for each of the 10 seeds. This setting could be considered favourable towards methods like `WOT` and `SBIRR`, since real-world diffusion estimates may in fact be off by two orders of magnitude (Forrow, 2024). For all methods, the SB step (13) is solved using up to 200 iterations of a multi-marginal iterative proportional fitting procedure (IPFP) (Marino and Gerolin, 2020, Section 4.3), which rescales the per-time slice scalings until the marginal constraints are met up to an L^∞ error of 10^{-5}. 2000 trajectories are then sampled from the inferred law on paths. Each method performs the drift MLE step (14) by fitting a 2-layer (128 neurons per layer) multi-layer perceptron with `SiLU` activation, trained to minimize (16) via the Adam optimizer (epochs $= 500$ and learning_rate $= 3 \times 10^{-3}$) (Kingma and Ba, 2014). Only `SBIRR` and `nn-APPEX` are iterative, and we use 30 iterations for each method.

Metrics. We assess `WOT`, `SBIRR`, and `nn-APPEX`, by evaluating both the scale and shape of their resulting landscapes $\nabla\Psi_{\hat{\theta}}$ with respect to the true landscape $\nabla\Psi$. We use the following metrics, evaluated on 2601 points (51×51 point grid) within $[-4, 4]^2$:

- *Absolute error (normalized)*: $\dfrac{|\nabla\Psi_{\hat{\theta}}(x) - \nabla\Psi(x)|}{\|\nabla\Psi(x)\|}$
- *Cosine similarity*: $\dfrac{\langle \nabla\Psi_{\hat{\theta}}(x),\ \nabla\Psi(x)\rangle}{\|\nabla\Psi_{\hat{\theta}}(x)\|\|\nabla\Psi(x)\|}$

5.2 Results

The results are aggregated over 10 seeds for each gradient-flow SDE, and shown in Figure 3. We see that `nn-APPEX` yields the best and most robust estimates for the drift landscapes. On each of the five SDEs, its estimates have the lowest absolute error, the highest cosine similarity, and the lowest variance. While `SBIRR` significantly outperforms `WOT`, due to iterative drift refinement, its inability to update its diffusion prior prevents it from learning the drift field to the same fidelity as `nn-APPEX` (see also Table 4 and Table 5 in the Appendix). In particular, `WOT` and `SBIRR` regularly orient flow lines incorrectly in low potential zones, due to misspecified diffusivity (see Figure 2 and Table 4).

We also replicate the experiment such that the initial distribution is either uniform, $p_0 \sim \text{Unif}[-4, 4]^2$, as done in previous work (Terpin et al., 2024), or given by the stationary Gibbs distribution $p_0 \sim p_{\text{eq}}$. The results are summarized in Figure 4 and Figure 5. For the uniform initialization, `nn-APPEX` continues to perform the best on all SDEs, and all methods perform relatively better, due to higher observability in the evaluation region $[-4, 4]^2$. For the Gibbs initialization, all methods fail at inference, which is consistent with identifiability theory (Proposition 3.1). Finally, we evaluate `nn-APPEX` against the state-of-the-art variational method `JKOnet*` (Terpin et al., 2024) given GMM initializations. Results are plotted in Figure 6 and show that `nn-APPEX` achieves more accurate drift and diffusion estimates.

6 BIOLOGICAL EXPERIMENTS

We now test `nn-APPEX`'s ability to perform trajectory inference on real biological data. As done in Shen et al. (2025), we use single-cell data from human embryonic stem cells (hESC) (Chu et al., 2016).

6.1 Experimental Setup

For consistency, we use the same data preprocessing, method implementation, and performance evaluation as Shen et al. (2025). Our code builds on the `SBIRR` public repository, such that the only modifications were to randomize the initial diffusivity, and to add the diffusion update step to implement `nn-APPEX`.

Dataset. The hESC dataset (Chu et al., 2016) comprises 5 time marginals (0h, 12h, 24h, 36h, 72h), and the observed number of cells per marginal are: 92, 102, 66, 172, 138. 5 PCA components are used.

Method implementation. For this experiment, the SB solver for each method is implemented with the forward-backward iterative proportional maximum likelihood algorithm (Vargas et al., 2021). `SBIRR` and `nn-APPEX` both use the same gradient field MLE for re-estimating the reference drift (see Shen et al. (2025, Section D.6.2) for details), and we add the closed form diffusivity MLE (15) for re-estimating the reference diffusion for `nn-APPEX`. We also implement a time-varying version of `nn-APPEX`, which computes (15) between each pair of marginals to estimate time-varying diffusivity. All iterative methods are run for 10 iterations.

Evaluation. We train each method on the first (0h), third (24h), and fifth (72h) marginals, and then use the learned dynamics to predict the second (12h) and fourth (36h) marginals. We then evaluate performance using the earth mover's distance between the estimated and true holdout marginals. We average results over 10 random seeds, and for each seed, we sample a random initial diffusivity within an order of magnitude of $\sigma^2 = 0.1$, the default considered in Shen et al. (2025).

6.2 Results

We report results in Table 1. We observe that `SBIRR`'s iterative drift refinement enables better performance compared to the simplest method `WOT`. Adding the diffusion update then provides another small but noticeable improvement, such that `nn-APPEX` (with time-varying diffusivity) achieves the best overall performance on both holdout marginals.

This experiment enhances our point that diffusion should be learned from the data, rather than assumed beforehand. Indeed, for real data, the diffusion is unknown, and it is often not obvious how it should be specified. Methods that refine both drift and diffusion also predict unseen marginals with higher accuracy, which suggests that they learn a better model.

Table 1: The average earth mover's distance (± standard error) across each method and each hold out time on the hESC dataset. The best results are displayed in bold and the second best results are underlined.

Method	Time	Avg. EMD
`WOT`	t_2	0.803 ± 0.039
	t_4	1.492 ± 0.081
`SBIRR`	t_2	0.694 ± 0.021
	t_4	1.473 ± 0.033
`nn-APPEX`	t_2	<u>0.683</u> ± 0.021
	t_4	<u>1.470</u> ± 0.040
`nn-APPEX` (time-varying)	t_2	**0.678** ± 0.022
	t_4	**1.454** ± 0.033

7 RELATED WORKS

Our work contributes novel theory and methodology for identifying the true SDE from observed marginals. We review relevant literature below.

Identifiability theory for population dynamics. To the best of our knowledge, conditions for jointly identifying an SDE's drift and diffusion from marginals have only been proven for linear additive noise SDEs (Guan et al., 2024), which cannot capture the multistable dynamics of cell differentiation. For the identifiability of gradient-flow SDEs, the partial identifiability result (2) was first proven by Hashimoto et al. (2016). Lavenant et al. (2024) extended the result for time-inhomogeneous drift $-\nabla\Psi(x,t)$, and showed that a single sample per marginal suffices, if the measurement times are dense. Finally, Neklyudov et al. (2023) proved that the result also holds if the known diffusivity $\sigma(t)^2$ is time-inhomogeneous.

Schrödinger Bridge based inference methods for population dynamics. A popular approach for marginals-based inference is to reconstruct trajectories with entropic optimal transport, which is equivalent to solving a Schrödinger Bridge (SB) problem (Peyré et al., 2019, Prop. 4.2). The first such method was Waddington-OT (`WOT`) (Schiebinger et al., 2019), which performs trajectory inference by solving a single SB problem with respect to a Brownian motion reference with a prescribed diffusivity σ^2. `WOT` spawned many variations, which incorporate additional information, like cell lineages (Forrow and Schiebinger, 2021; Ventre et al., 2023), or improve robustness given limited samples (Zhang et al., 2021; Lavenant et al., 2024; Chizat et al., 2022). In the last year, SB methods were improved with the introduction of iterative reference refinement, i.e. `SBIRR` (Shen et al., 2025; Zhang, 2024). Rather than fixing Brownian motion as a reference, `SBIRR` fixes the diffusivity σ^2 and iteratively re-estimates the drift of the reference process. In contrast to `SBIRR`, the method that we introduce in this work, `nn-APPEX`, additionally re-estimates the diffusivity at each iteration. `nn-APPEX` builds upon the three-stage iterative scheme of `APPEX` (Guan et al., 2024).

Other inference methods for population dynamics. In addition to SB methods, variational energy-based methods (Hashimoto et al., 2016; Bunne et al., 2022; Neklyudov et al., 2023; Terpin et al., 2024; Persiianov et al., 2025) have been developed for SDE inference from population dynamics. These methods leverage the fact that the marginals of potential-driven SDEs minimize a corresponding energy, and use discrete numerical schemes to approximate the underlying parameters. Of these, only `JKOnet*` (Terpin et al., 2024) jointly estimates drift and diffusion. We note that `JKOnet*` and `nn-APPEX` consider distinct optimization approaches (the former leverages the JKO scheme (Jordan et al., 1998) with a joint variational step, whereas `nn-APPEX` performs three-stage SB refinement) and we compare their performance in Appendix G. We also note that Terpin et al. (2024) did not study identifiability theory, and therefore lacked guarantees for principled inference. Recent work presents another framework for joint inference via maximum mean discrepancy minimization (Berlinghieri et al., 2025).

8 CONCLUSIONS

This work provides a call to action for researchers in ML and single-cell biology to move beyond the current paradigm of assuming known diffusion for marginals-based SDE inference, as we show that this is not only unnecessary, but also problematic in practice. Our theoretical contributions resolve a longstanding identifiability problem by proving that the gradient-flow drift and the diffusivity are jointly identifiable if and only if we observe transient marginals. This is significant for practical applications, since transience is a common and verifiable condition. We further expand the practicality of our result by proving that only three distinct marginals are needed for identifiability. To translate this theory into practice, we introduced the first Schrödinger Bridge-based method capable of inferring arbitrary gradient-flow drift and diffusivity. Extensive experiments demonstrate that our method, `nn-APPEX`, outperforms previous SB methods, in the realistic scenario where diffusivity is unknown.

Limitations and future work. While our work contributes novel theory and methodology for SDE inference from marginals, we note several limitations, which point to important directions for future work. First, the gradient-flow SDE model (1) is prevalent in many fields, but it does not capture some important dynamics. In single-cell dynamics, one may expect rotational dynamics from non-conservative drift, due to genetic feedback loops. As another example, in hydrology, material heterogeneity requires modified models for drift and diffusion. Second, our theory was proven given exact observation of the marginals $p(\cdot,t)$. Rigorous asymptotic theory quantifying identifiability with finite samples and observational noise would be useful to the community. Finally, there is still a gap in showing optimality of joint inference methods. While experiments show positive results, proving that `nn-APPEX` converges to the true parameters under basic conditions would mark a major advance.

Acknowledgments. We thank NSERC for their support for Vincent Guan and Joseph Janssen. Nicolas Lanzetti was supported by the NCCR Automation, a National Centre of Competence in Research, funded by the Swiss National Science Foundation (grant number 51NF40_225155). Elina Robeva was supported by a Canada CIFAR AI Chair and an NSERC Discovery Grant (DGECR-2020-00338). The authors would also like to thank United Therapeutics for supporting this research.

References

Jean-David Benamou, Guillaume Carlier, Simone Di Marino, and Luca Nenna. An entropy minimization approach to second-order variational mean-field games. *Mathematical Models and Methods in Applied Sciences*, 29(08):1553–1583, 2019.

Renato Berlinghieri, Yunyi Shen, Jialong Jiang, and Tamara Broderick. Oh snapmmd! forecasting stochastic dynamics beyond the schrödinger bridge's end. *arXiv preprint arXiv:2505.16082*, 2025.

Jaya PN Bishwal. *Parameter estimation in stochastic differential equations*. Springer, 2007.

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

Alexander P Browning, David J Warne, Kevin Burrage, Ruth E Baker, and Matthew J Simpson. Identifiability analysis for stochastic differential equation models in systems biology. *Journal of the Royal Society Interface*, 17(173):20200652, 2020.

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

Djalil Chafaï. From boltzmann to random matrices and beyond. In *Annales de la Faculté des sciences de Toulouse: Mathématiques*, volume 24, pages 641–689, 2015. no. 4.

Lénaïc Chizat, Stephen Zhang, Matthieu Heitz, and Geoffrey Schiebinger. Trajectory inference via mean-field Langevin in path space. *Advances in Neural Information Processing Systems*, 35:16731–16742, 2022.

Li-Fang Chu, Ning Leng, Jue Zhang, Zhonggang Hou, Daniel Mamott, David T Vereide, Jeea Choi, Christina Kendziorski, Ron Stewart, and James A Thomson. Single-cell rna-seq reveals novel regulators of human embryonic stem cell differentiation to definitive endoderm. *Genome biology*, 17(1):173, 2016.

Rick Durrett. *Probability: theory and examples*, volume 49. Cambridge university press, 2019.

Aden Forrow. Consistent diffusion matrix estimation from population time series. *arXiv preprint arXiv:2408.14408*, 2024.

Aden Forrow and Geoffrey Schiebinger. Lineageot is a unified framework for lineage tracing and trajectory inference. *Nature communications*, 12(1):4940, 2021.

Vincent Guan, Joseph Janssen, Hossein Rahmani, Andrew Warren, Stephen Zhang, Elina Robeva, and Geoffrey Schiebinger. Identifying drift, diffusion, and causal structure from temporal snapshots. *arXiv preprint arXiv:2410.22729*, 2024.

Wiebke Günther, Oana-Iuliana Popescu, Martin Rabel, Urmi Ninad, Andreas Gerhardus, and Jakob Runge. Causal discovery with endogenous context variables. *Advances in Neural Information Processing Systems*, 37:36243–36284, 2024.

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

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.

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

Hugo Lavenant, Stephen Zhang, Young-Heon Kim, and Geoffrey Schiebinger. Toward a mathematical theory of trajectory inference. *Ann. Appl. Probab.*, 34(1A), Feb 2024.

Tony Lelievre and Gabriel Stoltz. Partial differential equations and stochastic methods in molecular dynamics. *Acta Numerica*, 25:681–880, 2016.

Fadji Zaouna Maina and Erica R Siirila-Woodburn. Watersheds dynamics following wildfires: Nonlinear feedbacks and implications on hydrologic responses. *Hydrological Processes*, 34(1):33–50, 2020.

Simone Di Marino and Augusto Gerolin. An optimal transport approach for the schrödinger bridge problem and convergence of sinkhorn algorithm. *Journal of Scientific Computing*, 85(2):27, 2020.

Geoffrey J McLachlan and Suren Rathnayake. On the number of components in a gaussian mixture model. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 4(5):341–355, 2014.

Stéphane Menozzi, Antonello Pesce, and Xicheng Zhang. Density and gradient estimates for non degenerate brownian sdes with unbounded measurable

drift. *Journal of Differential Equations*, 272:330–369, 2021.

Nilah Monnier, Syuan-Ming Guo, Masashi Mori, Jun He, Péter Lénárt, and Mark Bathe. Bayesian approach to msd-based analysis of particle motion in live cells. *Biophysical journal*, 103(3):616–626, 2012.

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

Jan Nygaard Nielsen, Henrik Madsen, and Peter C Young. Parameter estimation in stochastic differential equations: an overview. *Annual Reviews in Control*, 24:83–94, 2000.

Bernt Oksendal. *Stochastic differential equations: an introduction with applications*. Springer Science & Business Media, 2013.

Grigorios A Pavliotis. Stochastic processes and applications. *Texts in Applied Mathematics*, 60, 2014.

Asger Roer Pedersen. A new approach to maximum likelihood estimation for stochastic differential equations based on discrete observations. *Scandinavian journal of statistics*, pages 55–71, 1995.

Mikhail Persiianov, Jiawei Chen, Petr Mokrov, Alexander Tyurin, Evgeny Burnaev, and Alexander Korotin. Learning of population dynamics: Inverse optimization meets jko scheme. *arXiv preprint arXiv:2506.01502*, 2025.

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.

Hannes Risken. Fokker-planck equation. In *The Fokker-Planck equation: methods of solution and applications*, pages 63–95. Springer, 1989.

Jakob Runge, Sebastian Bathiany, Erik Bollt, Gustau Camps-Valls, Dim Coumou, Ethan Deyle, Clark Glymour, Marlene Kretschmer, Miguel D Mahecha, Jordi Muñoz-Marí, et al. Inferring causation from time series in earth system sciences. *Nature communications*, 10(1):2553, 2019.

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

Jan Seibert, Jeffrey J McDonnell, and Richard D Woodsmith. Effects of wildfire on catchment runoff response: a modelling approach to detect changes in snow-dominated forested catchments. *Hydrology research*, 41(5):378–390, 2010.

Yunyi Shen, Renato Berlinghieri, and Tamara Broderick. Multi-marginal schrödinger bridges with iterative reference refinement. In *International Conference on Artificial Intelligence and Statistics*, pages 3817–3825. PMLR, 2025.

Daniel W. Stroock. *Partial Differential Equations for Probabilists*. Cambridge Studies in Advanced Mathematics. Cambridge University Press, 2008.

Antonio Terpin, Nicolas Lanzetti, Martín Gadea, and Florian Dörfler. Learning diffusion at lightspeed. *Advances in Neural Information Processing Systems*, 37:6797–6832, 2024.

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

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

Elias Ventre, Aden Forrow, Nitya Gadhiwala, Parijat Chakraborty, Omer Angel, and Geoffrey Schiebinger. Trajectory inference for a branching sde model of cell differentiation. *arXiv preprint arXiv:2307.07687*, 2023.

CH Waddington. How animal develop, 1935.

Yuanyuan Wang, Xi Geng, Wei Huang, Biwei Huang, and Mingming Gong. Generator identification for linear sdes with additive and multiplicative noise. *Advances in Neural Information Processing Systems*, 36, 2024.

Caleb Weinreb, Samuel Wolock, Betsabeh K Tusi, Merav Socolovsky, and Allon M Klein. Fundamental limits on dynamic inference from single-cell snapshots. *Proceedings of the National Academy of Sciences*, 115(10):E2467–E2476, 2018.

Grace Hui Ting Yeo, Sachit D Saksena, and David K Gifford. Generative modeling of single-cell time series with prescient enables prediction of cell trajectories with interventions. *Nature communications*, 12(1): 3222, 2021.

Bin Yu, Chen Chen, Ren Qi, Ruiqing Zheng, Patrick J Skillman-Lawrence, Xiaolin Wang, Anjun Ma, and Haiming Gu. scgmai: a gaussian mixture model for clustering single-cell rna-seq data based on deep autoencoder. *Briefings in bioinformatics*, 22(4): bbaa316, 2021.

Stephen Zhang, Anton Afanassiev, Laura Greenstreet, Tetsuya Matsumoto, and Geoffrey Schiebinger. Optimal transport analysis reveals trajectories in steady-state systems. *PLoS computational biology*, 17(12): e1009466, 2021.

Stephen Y Zhang. Joint trajectory and network inference via reference fitting. In *Machine Learning in Computational Biology*, pages 72–85. PMLR, 2024.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [**Yes**/No/Not Applicable]
   The data-generating model is defined in (1) and precise mathematical details and blanket assumptions are given in Section 2. Our inference algorithm is defined in Section 4 and further details are given in Appendix D
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [**Yes**/No/Not Applicable]
   The mathematical properties of our algorithm are discussed in Section 4, with related derivations in Appendix D and Appendix E. Relevant numerical details are provided in Section 5, and runtimes are reported in Table 3.
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [**Yes**/No/Not Applicable]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [**Yes**/No/Not Applicable]
   All assumptions, outside the blanket assumptions (two textbook assumptions required for SDE existence and uniqueness), are contained within the statements of the theoretical result.
   (b) Complete proofs of all theoretical results. [**Yes**/No/Not Applicable]
   Given the importance of our theory and the relative brevity of the proofs, we include full proofs in Section 3, under Theorem 3.2 and Corollary 3.3 respectively. Technical details related to our inference algorithm are derived in Appendix D and Appendix E.
   (c) Clear explanations of any assumptions. [**Yes**/No/Not Applicable]
   We discuss assumptions, their relevance for practical inference, and comparison with assumptions from previous literature throughout the paper. For example, see the discussion after the proof of Theorem 3.2.

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [**Yes**/No/Not Applicable]
   A zip file of the code repository with scripts for running the experiments, and a detailed README file are included in supplemental material.
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [**Yes**/No/Not Applicable] *We report implementation details in Section 5 and emphasize that we use exact implementations for any shared procedures used by baselines.*
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [**Yes**/No/Not Applicable]
   We specify our metrics in Section 5 and state that we report results over 10 *random seeds.*
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [**Yes**/No/Not Applicable]
   Experiments were run locally on a laptop. Details are given at the start of Appendix G.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [**Yes**/No/Not Applicable]
   We declare dependencies on two previous repositories in Appendix G and cite the original works.
   (b) The license information of the assets, if applicable. [**Yes**/No/Not Applicable]
   We report the license in Appendix G and also in our source code.
   (c) New assets either in the supplemental material or as a URL, if applicable. [**Yes**/No/Not Applicable]
   We provide the full code repository as supplemental material.
   (d) Information about consent from data providers/curators. [Yes/No/**Not Applicable**]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Yes/No/**Not Applicable**]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Yes/No/**Not Applicable**]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Yes/No/**Not Applicable**]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Yes/No/**Not Applicable**]

A THEORETICAL BACKGROUND

A.1 The Distributional Definition of the Fokker-Planck Equation

The Fokker-Planck equation defines the evolution of an SDE's marginals. For an overdamped Langevin SDE (1), its Fokker-Planck equation is given by

$$\frac{\partial p(x,t)}{\partial t} = \mathcal{L}^*_{-\nabla\Psi,\sigma^2}(p(\cdot,t))(x), \tag{19}$$

where the Fokker-Planck operator $\mathcal{L}^*_{-\nabla\Psi,\sigma^2}$ is defined as

$$(\mathcal{L}^*_{-\nabla\Psi,\sigma^2}p)(x) = \nabla\cdot(p(x)\nabla\Psi(x)) + \frac{\sigma^2}{2}\Delta p(x).$$

However, the equation (19) cannot be interpreted as a strong pointwise equality unless p is a twice differentiable probability density and $\Psi \in C^2(\mathbb{R}^d)$, which is a stronger condition than Lipschitz continuity of $\nabla\Psi$. To show that the evolution of marginals remains well-defined when we consider a general probability measure μ with finite second moments, we consider the weak distributional formulation of the Fokker-Planck equation (19) through its adjoint operator,

$$(\mathcal{L}_{-\nabla\Psi,\sigma^2}f)(x) = -\nabla\Psi(x)\cdot\nabla f(x) + \frac{\sigma^2}{2}\Delta f(x).$$

Consider now a family of Borel locally finite measures on $\mathbb{R}^d \times (0,T)$, which we denote by $(\mu_t)_{t\in(0,T)}$. To define weak solutions, we follow (Bogachev et al., 2022, Definition 6.1.1, Proposition 6.1.2(ii)) and say that $(\mu_t)_{t\in(0,T)}$ solves the Fokker-Planck equation for the initial condition $\mu|_{t=0} = \nu$ in the weak sense if for all test functions $\varphi \in C_0^\infty(\mathbb{R}^d)$, we have that

$$\int_{\mathbb{R}^d}\varphi(x)\mathrm{d}\mu_t(x) - \int_{\mathbb{R}^d}\varphi(x)\mathrm{d}\nu(x) = \lim_{\tau\to 0+}\int_\tau^t\int_{\mathbb{R}^d}\mathcal{L}_{-\nabla\Psi,\sigma^2}\varphi(x)\mathrm{d}\mu_s(x)\mathrm{d}s \tag{20}$$

for almost all $t \in [0,T]$. This can be denoted using the compact notation

$$\partial_t\mu = \mathcal{L}^*_{-\nabla\Psi,\sigma^2}\mu. \tag{21}$$

In this setting, the existence and uniqueness of a weak solution follows from (Bogachev et al., 2022, Theorem 9.4.8, Example 9.4.7), which applies because σ^2 is constant and we assumed that $\nabla\Psi$ is Lipschitz and obeys the growth condition, $\|\nabla\Psi(x)\| \le K(1+\|x\|)$. This result is also given in Stroock (2008, Theorem 1.1.9), which additionally shows that the family $(\mu_t)_{t\in(0,T)}$ is continuous under the topology of weak convergence. Thus, the evolution of marginals can be identified with a continuous transition probability function, even under the weak formulation.

A.2 Additional Proofs

The following results are used to complete the proof of Corollary 3.3. Since the observed marginals in Corollary 3.3 are sampled after some time $T_1 > 0$, we note that it suffices to consider times $t \ge T_1 > 0$. Also recall that for this result, we assume that $\Psi \in C^\infty(\mathbb{R}^d)$. This facilitates the proof, by ensuring that any marginal $p(\cdot,t)$ of a candidate gradient-flow SDE exhibits nice regularity and decay. Indeed, by elliptic regularity theory, the Fokker-Planck operator $\mathcal{L}_{-\nabla\Psi,\sigma^2}$ smooths marginals within any finite time.

Lemma A.1 (Finite-time smoothing of marginals)**.** *Suppose that the potential of the gradient-flow SDE (1) is smooth, i.e. $\Psi \in C^\infty(\mathbb{R}^d)$. Then, for any positive time $t > 0$, it follows that the marginal $p(\cdot,t)$ admits a twice differentiable density, which obeys the Gaussian decay estimates*

$$p(x,t) \le Kt^{\frac{-d+2}{2}}\exp\big(-\frac{\delta}{2t}\|x\|^2\big)$$
$$\|\nabla p(x,t)\| \le Kt^{\frac{-d+2}{2}}\exp\big(-\frac{\delta}{2t}\|x\|^2\big)$$

for some $K, \delta > 0$.

Proof. This result is given by Pavliotis (2014, Theorem 4.1). However, we note that in the setup of this theorem, the initial distribution $p(\cdot, 0)$ is assumed to have a density. Thus, we first apply the density estimates from Menozzi et al. (2021, Theorem 1.2) to show that for any positive time $t/2 > 0$, the marginal $p(\cdot, t/2)$ admits a density, even if $p(\cdot, 0)$ is only a probability measure μ_0. Indeed, by Menozzi et al. (2021, Theorem 1.2), the SDE admits a transition density function, $p(x, t|y, 0)$. Therefore, $p(\cdot, t) = \mu_0 * p(\cdot, t \mid \cdot, 0)$ admits a density, since it is the convolution of a probability measure with a probability density (Durrett, 2019, Theorem 2.1.16). We may thus apply Pavliotis (2014, Theorem 4.1) after setting the initial density as the density of $p(\cdot, t/2)$. By combining these two PDE regularity theorems, we obtain the desired regularity and decay properties. □

Lemma A.2 (Equivalence of marginals and of flows at accumulation point)**.** *As defined in Corollary 3.3, let* $\mathcal{I}_i = \{t \in [T_i, T_{i+1}] \mid p(\cdot, t) = q(\cdot, t)\}$*, and let* $\{t_n\}_{n\geq 0}$ *be a sequence of times in* $\mathcal{I}_i$*, which converges to* t_i^**. Then,* $p(\cdot, t_i^*) = q(\cdot, t_i^*)$ *and* $\frac{\partial}{\partial_t} p(\cdot, t_i^*) = \frac{\partial}{\partial_t} q(\cdot, t_i^*)$*.*

Proof. First, we define the map $f(t) : t \in [T_i, T_{i+1}] \to p(\cdot, t) - q(\cdot, t)$, where we recall that $T_i > 0$. Then, by Lemma A.1, for any $t \geq T_i$, $p(\cdot, t)$ and $q(\cdot, t)$ each admit a smooth density. Since we also have $\Psi \in C^\infty(\mathbb{R}^d)$, it follows that the Fokker-Planck equation (4) is defined pointwise, which implies the existence of the time derivatives $\frac{\partial p(x,t)}{\partial t}$ and $\frac{\partial q(x,t)}{\partial t}$ (Pavliotis, 2014, Theorem 4.1). Thus, $f(t) = p(\cdot, t) - q(\cdot, t)$ is also a strongly differentiable map in $[T_i, T_{i+1}]$. From this, we deduce that $\mathcal{I}_i$ is closed, since $\mathcal{I}_i = f^{-1}(\{0\})$, which is the preimage of a closed set under a continuous map. Since t_i^* is the limit point of a sequence $\{t_n\}_{n\in\mathbb{N}}$ in $\mathcal{I}_i$, it follows that t_i^* is also in the coincidence set $\mathcal{I}_i$, which proves that $p(\cdot, t_i^*) = q(\cdot, t_i^*)$. In fact, since they each admit densities, this is equivalent to the pointwise equality, $p(x, t_i^*) = q(x, t_i^*)\ \forall x \in \mathbb{R}^d$. Then, to prove that $\frac{\partial}{\partial_t} p(\cdot, t_i^*) = \frac{\partial}{\partial_t} q(\cdot, t_i^*)$, we recall that $p(\cdot, t)$ and $q(\cdot, t)$ are strongly differentiable in time, such that, for all $x \in \mathbb{R}^d$,

$$\frac{\partial}{\partial t} p(x, t_i^*) = \lim_{t_n \to t_i^*} \frac{p(x, t_n) - p(x, t_i^*)}{t_n - t_i^*}$$
$$\frac{\partial}{\partial t} q(x, t_i^*) = \lim_{t_n \to t_i^*} \frac{q(x, t_n) - q(x, t_i^*)}{t_n - t_i^*}.$$

Then, by convergence of the sequence $t_n \to t_i^*$, the fact that $p(x, t_n) = q(x, t_n)$ (by construction of the sequence) and $p(x, t_i^*) = q(x, t_i^*)$, it follows that $\frac{\partial}{\partial t} p(x, t_i^*) = \frac{\partial}{\partial t} q(x, t_i^*)$ for all $x \in \mathbb{R}^d$. □

Proposition A.3 (H-Theorem: marginals of gradient-flow SDEs monotonically decrease free energy)**.** *Let* $(p(\cdot, t))_{t\geq T_1}$ *be the marginals of a Langevin SDE (1) from the earliest possible observation time in the setting of Corollary 3.3, and suppose that* $\Psi \in C^\infty(\mathbb{R}^d)$*. Then, the free energy* $J(\rho) = \int_{\mathbb{R}^d} \Psi(x)\rho(x)\mathrm{d}x + \beta \int_{\mathbb{R}^d} \rho(x) \log \rho(x)\mathrm{d}x$*, such that* $\beta = \frac{\sigma^2}{2}$*, strictly decreases in time, unless the process is at its stationary Gibbs distribution,* p_{eq}*. Thus,* p_{eq} *is the only possible marginal that can repeat at distinct times.*

Proof. We apply Lemma A.1, which implies that for all $t \geq T_1$, $p(\cdot, t)$ admits a twice differentiable density, with the exponential decay on both the density $p(x, t) \leq K t^{\frac{-d+2}{2}} \exp\left(-\frac{\delta}{2t}\|x\|^2\right)$ and the gradient $\|\nabla p(x, t)\| \leq K t^{\frac{-d+2}{2}} \exp\left(-\frac{\delta}{2t}\|x\|^2\right)$, for some $K, \delta > 0$. Thus, we may use chain rule, substitute the Fokker-Planck equation, and integrate by parts (with all boundary terms vanishing) to obtain

$$\frac{\mathrm{d}}{\mathrm{d}t} J(p(\cdot, t)) = \int_{\mathbb{R}^d} \left(\Psi(x) + \beta(\log p(x, t) + 1)\right) \frac{\partial}{\partial t} p(x, t)\, \mathrm{d}x \tag{22}$$
$$= \int_{\mathbb{R}^d} \left(\Psi(x) + \beta(\log p(x, t) + 1)\right) \left[\nabla \cdot (p(x, t)\nabla\Psi(x)) + \beta\Delta p(x, t)\right] \mathrm{d}x \tag{23}$$
$$= -\int_{\mathbb{R}^d} |\nabla\Psi(x)|^2 p(x, t)\mathrm{d}x - 2\beta \int \nabla p(x, t) \cdot \nabla\Psi(x)\mathrm{d}x - \beta^2 \int \frac{|\nabla p(x, t)|^2}{p(x, t)}\mathrm{d}x \tag{24}$$
$$= -\int_{\mathbb{R}^d} p(x, t) \left\|\nabla\Psi(x) + \beta\nabla \log p(x, t)\right\|^2 \mathrm{d}x \leq 0, \tag{25}$$

which has equality only if $\log p(x, t) + \frac{\Psi(x)}{\beta} = K$ for some constant K. It follows that the Gibbs distribution

$$p_{\mathrm{eq}}(x) = \frac{1}{Z_{\Psi,\sigma^2}} \exp\left(-\frac{2\Psi(x)}{\sigma^2}\right),$$

is the only possible marginal, observed from $t \geq T_1$, which induces no change in free energy $J(p(\cdot,t))$. Moreover, by Lemma A.4, we note that $J(p(\cdot,t)) < \infty$. Each transient marginal in the evolution $\big(p(\cdot,t)\big)_{t \geq T_1}$ therefore has a distinct well-defined free energy, which strictly decreases in time, unless we are at $p_{\text{eq}}(x)$. We conclude that the only possible re-occurring marginal in the observation period $t \geq T_1$ is the Gibbs distribution p_{eq}, which is the unique stationary distribution as long as it is integrable (Bogachev et al., 2022)[Theorem 4.1.11]. □

Lemma A.4 (Finite free energy). *Let $p(\cdot,t)$ satisfy $p(x,t) \leq Kt^{\frac{-d+2}{2}} \exp\big(-\frac{\delta}{2t}\|x\|^2\big)$ for some $K, \delta > 0$. Then, $J(p(\cdot,t)) = \int_{\mathbb{R}^d} \Psi(x)p(x,t)\mathrm{d}x + \beta \int_{\mathbb{R}^d} p(x,t)\log(p(x,t))\mathrm{d}x < \infty$, with $\beta = \frac{\sigma^2}{2}$.*

Proof. By the fundamental theorem of calculus along the segment $t \mapsto tx$,

$$\Psi(x) - \Psi(0) = \int_0^1 \nabla\Psi(tx) \cdot x \,\mathrm{d}t.$$

Then, since we have the linear growth condition, $\|\nabla\Psi(x)\| \leq K(1 + \|x\|)$ for all $x \in \mathbb{R}^d$, we have

$$|\Psi(x)| \leq |\Psi(0)| + \|x\| \int_0^1 \|\nabla\Psi(tx)\| \,\mathrm{d}t \leq |\Psi(0)| + \|x\| \int_0^1 K(1 + t\|x\|) \,\mathrm{d}t = |\Psi(0)| + K\|x\| + \tfrac{K}{2}\|x\|^2.$$

This bound shows that if $p(\cdot,t)$ has finite second moments, then the potential energy term $\int_{\mathbb{R}^d} \Psi(x)p(x,t)\mathrm{d}x$ in the free energy is bounded. Since we always assume that the initial distribution has finite second moments, then $p(\cdot,t)$ also has finite second moments (see (Stroock, 2008, Theorem 1.1.9)).

We now control the second term in the free energy, by showing that $p(\cdot,t)$ has finite differential entropy. It suffices to apply the Gaussian decay estimate $p(x,t) \leq Kt^{\frac{-d+2}{2}} \exp\big(-\frac{\delta}{2t}\|x\|^2\big)$. By direct computation,

$$\begin{aligned}
\int_{\mathbb{R}^d} p(x,t)\log p(x,t)\,\mathrm{d}x &\leq \int_{\mathbb{R}^d} p(x,t)\left[\log K + \tfrac{-d+2}{2}\log t - \tfrac{\delta}{2t}\|x\|^2\right]\mathrm{d}x \\
&\leq \log K \int p(x,t)\,\mathrm{d}x + \tfrac{-d+2}{2}\log t \int p(x,t)\,\mathrm{d}x - \tfrac{\delta}{2t}\int \|x\|^2 p(x,t)\,\mathrm{d}x \\
&= \log K + \tfrac{-d+2}{2}\log t - \tfrac{\delta}{2t}\mathbb{E}_{p(\cdot,t)}[\|x\|^2] < \infty,
\end{aligned}$$

since $p(\cdot,t)$ is a probability density with finite second moments. We therefore conclude that the free energy

$$J(p(\cdot,t)) = \int_{\mathbb{R}^d} \Psi(x)\,p(x,t)\,\mathrm{d}x + \beta \int_{\mathbb{R}^d} p(x,t)(x)\log p(x,t)\,\mathrm{d}x < \infty$$

□

We emphasize that Proposition A.3 is a classical result, which is well known in the stochastic analysis and statistical mechanics literature, where it is commonly referred to as an H-theorem, due to its connection to Boltzmann's second law of thermodynamics (Chafaï, 2015). For example, (Jordan et al., 1998) states that given that $\Psi \in C^\infty(\mathbb{R}^d)$ and $p(\cdot,0)$ is a density with finite free energy, then it is known in the literature (e.g. (Risken, 1989) and attached references) that $J(\rho)$ is a monotone functional on the marginals of a gradient-flow SDE (1), and that $J(\rho)$ is uniquely minimized at the stationary Gibbs distribution. While we have derived a proof of this result, we believe that the statement holds under more relaxed conditions. In particular, we use the assumption $\Psi \in C^\infty(\mathbb{R}^d)$ to derive the regularity and decay estimates from Lemma A.1, which we frequently use for our proof. However, we believe that existing results in parabolic elliptic PDE theory should yield the same estimates under milder assumptions. We note that a proof is given in (Chafaï, 2015, Sec. 1.7), provided that the Gibbs distribution is integrable, and (Jordan et al., 1998) notes that the monotonicity holds, even if the Gibbs distribution is not integrable, but it is unclear what conditions are required.

B IDENTIFIABILITY IN THE TIME-INHOMOGENEOUS CASE

Theorem 3.2 assumes that the drift and diffusion terms are time-homogeneous. However, many physical systems have time-inhomogeneous dynamics. For example, in hydrology, extreme events, such as droughts or fires, can

cause "regime shifts" in the underlying functional relationships (Runge et al., 2019; Günther et al., 2024; Pedersen, 1995; Seibert et al., 2010; Maina and Siirila-Woodburn, 2020). We may define this notion for gradient-flow SDEs, such that the drift and diffusion terms change due to regime shifts at specific times $\{t_i\}_{i=0}^{n-1}$ (with the drift remaining smooth and satisfying the growth condition in each regime to ensure well-posedness).

Definition B.1 (Time-inhomogeneous gradient-flow SDE with discrete regimes)**.** *A time-inhomogeneous Langevin SDE with discrete regimes, marked by events taking place at times $\{t_i\}_{i=0}^{n-1}$, is given by*

$$\mathrm{d}X_t = -\nabla\Psi(X_t, t_i)\mathrm{d}t + \sigma(t_i)\mathrm{d}W_t, \tag{26}$$

such that $\nabla\Psi(X_t, t)$ and $\sigma(t)^2$ are time-homogeneous within each regime $t \in [t_i, t_{i+1})$.

Under these time-inhomogeneous dynamics, identifiability, which is defined analogously to the time-homogeneous case, is still guaranteed by transient marginals.

Corollary B.2 (Identifiability for time-inhomogeneous gradient-flow SDEs with discrete regimes)**.** *The time-inhomogeneous gradient-flow SDE with discrete regimes (26) is identifiable from its marginals $\big(p(\cdot,t)\big)_{t\in[0,T]}$ if and only if $p(\cdot, t_i)$ is not a stationary distribution of $\mathrm{d}X_t = \nabla\Psi(X_t, t_i)\mathrm{d}t + \sigma(t_i)\mathrm{d}W_t$ for each t_i.*

Proof. We apply Theorem 3.2 on each regime $[t_i, t_{i+1})$. Since the regimes form a partition of $[0, T]$, this guarantees that there is a unique SDE with the observed marginals $\big(p(\cdot,t)\big)_{t\in[0,T]}$. □

Intuitively, the drift and diffusion within a given regime $[t_i, t_{i+1})$ are identifiable if and only if the regime's initial marginal, $p(\cdot, t_i)$, is not a stationary distribution. However, we note that if the drift or diffusion terms change continuously in time, then it no longer follows that transience completely characterizes identifiability. Indeed, the process could be in an equilibrium state that itself is changing in time, as the next example shows. In particular, distinct time-inhomogeneous SDEs may share the same marginals $\big(p(\cdot,t)\big)_{t\geq 0}$, as long as for each $t \geq 0$, the marginal $p(\cdot, t)$ is precisely the stationary distribution of the time-inhomogeneous residual process, evaluated at that time.

Example B.3 (Continuously evolving equilibrium due to time-inhomogeneous potential)**.** *The SDEs*

$$\mathrm{d}X_t = \mathrm{d}W_t, \tag{27}$$

$$\mathrm{d}Y_t = \begin{cases} \mathrm{d}W_t, & t = 0, \\ -\dfrac{Y_t}{t}\,\mathrm{d}t + \sqrt{3}\,\mathrm{d}W_t, & t > 0. \end{cases} \tag{28}$$

produce the same marginals $p(\cdot, t) = \mathcal{N}(0, t)$ when initialized with $X_0 = Y_0 = 0$, i.e., $p_0 = \delta(x)$.

Proof. Let $t > 0$. The respective Fokker-Planck operators for both SDEs are

$$\mathcal{L}^*_{0,1}\big(p(\cdot,t)\big)(x) = \frac{1}{2}\frac{\partial^2}{\partial x^2}p(x,t)$$
$$\mathcal{L}^*_{-\frac{x}{t},3}\big(p(\cdot,t)\big)(x) = \frac{\partial}{\partial x}\left(p(x,t)\frac{x}{t}\right) + \frac{3}{2}\frac{\partial^2}{\partial x^2}p(x,t)$$

Then, suppose that $p(\cdot, t) = \mathcal{N}(0, t)$, such that $p(x,t) = \frac{1}{\sqrt{2\pi t}}\exp\left(-\frac{x^2}{2t}\right)$. Then,

$$\frac{\partial}{\partial x}p(x,t) = -\frac{x}{t}p(x,t)$$
$$\frac{\partial^2}{\partial x^2}p(x,t) = -\frac{p(x,t)}{t} + \frac{x^2}{t^2}p(x,t)$$

It follows that

$$\begin{aligned}
\mathcal{L}^*_{0,1}\big(p(\cdot,t)\big)(x) &= -\frac{p(x,t)}{2t} + \frac{x^2}{2t^2}p(x,t) \\
&= \frac{p(x,t)}{2t^2}\big(x^2 - t\big) \\
\mathcal{L}^*_{-\frac{x}{t},3}\big(p(\cdot,t)\big)(x) &= -\frac{x^2}{t^2}p(x,t) + \frac{p(x,t)}{t} + \frac{3x^2}{2t^2}p(x,t) - \frac{3p(x,t)}{2t} \\
&= \frac{p(x,t)}{2t^2}\big(-2x^2 + 2t + 3x^2 - 3t\big) \\
&= \frac{p(x,t)}{2t^2}\big(x^2 - t\big).
\end{aligned}$$

Thus, the Fokker-Planck operators are equivalent for $p(\cdot,t) = \mathcal{N}(0,t)$, which are the marginals of Brownian motion. Equivalently, we observe that $\mathcal{N}(0,t)$ is the stationary distribution of the residual Fokker-Planck equation at time t:

$$0 = \mathcal{L}^*_{-\frac{x}{t},2}\big(p_{\text{eq}}\big)(x) = \frac{\partial}{\partial x}\left(p_{\text{eq}}(x)\frac{x}{t}\right) + \frac{\partial^2}{\partial x^2}p_{\text{eq}}(x).$$

Given $t > 0$, $p_{\text{eq}}(x) = \frac{1}{\sqrt{2\pi t}}\exp\left(-\frac{x^2}{2t}\right)$ solves the above equation. This yields the interpretation that the marginals are at equilibrium (for the residual process $\mathrm{d}Z_t = -\frac{Z_t}{t}\mathrm{d}t + \sqrt{2}\mathrm{d}W_t$), which itself is time-varying. □

C ADDITIONAL NON-IDENTIFIABILITY EXAMPLES

For gradient-flow SDEs, we have proven that non-identifiability from marginals arises if and only if the observed marginals also correspond to the marginals of a residual process in equilibrium. An example for the time-homogeneous case was given in Example 2.2, such that the equilibrium marginals are constant, and an example for the time-inhomogenous case was given in Example B.3, such that the equilibrium distribution of the residual process continuously changes in time.

However, for other forms of SDEs, other types of non-identifiability have been documented in the literature, including: undetectable rotations (Weinreb et al., 2018; Shen et al., 2025; Guan et al., 2024), rank-degenerate trajectories (Wang et al., 2024), and sharing the same stationary distribution (Lavenant et al., 2024). We overview examples below:

Example C.1 (Gaussian pancake (Guan et al., 2024))**.** *Let $a, b, c \in \mathbb{R}$, with $a \leq b$, and let X_0 be defined as a 2d Gaussian pancake, such that for each fixed $x_0^{(2)} \in [a,b]$, the first coordinate $x_0^{(1)} \sim \mathcal{N}(0,1)$:*

$$\mathrm{d}X_t = \begin{bmatrix}-1 & 0\\ 0 & b\end{bmatrix}\mathrm{d}t + \begin{bmatrix}1 & 0\\ 0 & c\end{bmatrix}\mathrm{d}W_t, \qquad X_0 \sim \mathcal{N}(0,1)\times[a,b] \tag{29}$$

$$\mathrm{d}Y_t = \begin{bmatrix}-10 & 0\\ 0 & b\end{bmatrix}\mathrm{d}t + \begin{bmatrix}\sqrt{10} & 0\\ 0 & c\end{bmatrix}\mathrm{d}W_t. \qquad Y_0 \sim \mathcal{N}(0,1)\times[a,b] \tag{30}$$

Example C.2 (Rotation (Shen et al., 2025; Hashimoto et al., 2016; Weinreb et al., 2018; Guan et al., 2024))**.**

$$\begin{aligned}
\mathrm{d}X_t &= \mathrm{d}W_t, & X_0 &\sim \mathcal{N}(0, I)\\
\mathrm{d}Y_t &= \begin{bmatrix}0 & 1\\ -1 & 0\end{bmatrix}Y_t\,\mathrm{d}t + \mathrm{d}W_t. & Y_0 &\sim \mathcal{N}(0, I)
\end{aligned}$$

Example C.3 (Degenerate rank (Wang et al., 2024; Guan et al., 2024))**.**

$$\begin{aligned}
\mathrm{d}X_t &= \begin{bmatrix}1 & 2\\ 1 & 0\end{bmatrix}X_t\,\mathrm{d}t + \begin{bmatrix}1 & 2\\ -1 & -2\end{bmatrix}\mathrm{d}W_t, & X_0 &= \begin{bmatrix}1\\ -1\end{bmatrix}\\
\mathrm{d}Y_t &= \begin{bmatrix}1/3 & 4/3\\ 2/3 & -1/3\end{bmatrix}Y_t\,\mathrm{d}t + \begin{bmatrix}1 & 2\\ -1 & -2\end{bmatrix}\mathrm{d}W_t. & Y_0 &= \begin{bmatrix}1\\ -1\end{bmatrix}
\end{aligned}$$

D ADDITIONAL NOTES ON NN-APPEX

D.1 Convergence

To discuss convergence of `nn-APPEX`'s tri-level iterative scheme (13)-(15), we first rigorously justify why we may re-estimate diffusion while ensuring that the objective remains finite. In particular, we must show that we have finite KL divergence between the reconstructed paths from the step (13) and the paths following MLE parameter estimation of a gradient-flow SDE from the steps (14) and (15).

Given continuously observed marginals of a d-dimensional process, recall that the KL divergence between two laws on paths P and Q, taken over the path space $\Omega = C([0,T],\mathbb{R}^d)$, is given by

$$\mathrm{KL}(Q\|P) = \int_\Omega \log\left(\frac{dQ}{dP}(\omega)\right) dQ(\omega). \tag{31}$$

By Girsanov's theorem, (31) is only finite for Q and P, if their underlying SDEs share the same diffusion (Vargas et al., 2021). Besides non-identifiability, this is another cited reason for why previous Schrödinger Bridge methods assume that diffusivity must be fixed (Shen et al., 2025). However, in the practical setting, we only observe a finite number of marginals, so we should instead optimize KL divergence over the discretized path space $\Omega_N = C(\{t_i\}_{i=0}^{N-1},\mathbb{R}^d)$. Then, for each law on paths, we only observe the couplings, $Q_{t_i,t_{i+1}}$ and $P_{t_i,t_{i+1}}$, between consecutive times. Let Q^N and P^N denote the concatenations of these couplings from t_0 to t_{N-1}. Then, by Benamou et al. (2019)[Lemma 3.4], evaluating the KL divergence over $\Omega_N = C(\{t_i\}_{i=0}^{N-1},\mathbb{R}^d)$ yields

$$\mathrm{KL}(P^N\|Q^N) = \sum_{i=0}^{N-2} \mathrm{KL}(P_{t_i,t_{i+1}}\|Q_{t_i,t_{i+1}}) - \sum_{i=1}^{N-2} \mathrm{KL}(P_{t_i}\|Q_{t_i}). \tag{32}$$

First note that by the data processing inequality, $\mathrm{KL}(P_{t_i}\|Q_{t_i}) \le \mathrm{KL}(P_{t_i,t_{i+1}}\|Q_{t_i,t_{i+1}})$. Hence, the expression will be finite as long as $\mathrm{KL}(P_{t_i,t_{i+1}}\|Q_{t_i,t_{i+1}}) = \int_{\mathbb{R}^d\times\mathbb{R}^d} \log\left(\frac{dP_{t_i,t_{i+1}}}{dQ_{t_i,t_{i+1}}}\right) P_{t_i,t_{i+1}}(x)dx < \infty$. Since the transition density of any nondegenerate diffusion process is absolutely continuous with respect to the Lebesgue measure for any positive time, it follows that the expression is finite for all gradient-flow SDEs (1).

As described in Section 4, `nn-APPEX` will alternate between finding a law on paths P that minimizes KL in the first argument (trajectory inference) and finding a law on paths Q that minimizes KL in the second argument (MLE parameter estimation). Note that each optimization is subject to distinct hard constraints, which respectively stipulate that $P \in \Pi(p(\cdot,t_i)_{i=0}^{N-1})$ adheres to the produced marginals and that Q is the law of a gradient-flow SDE. The algorithm therefore performs alternating projections, onto distinct spaces. In fact, by Corollary 3.3, only one gradient-flow SDE $Q_{-\nabla\Psi,\sigma^2}$ obeys the marginal constraints, which implies that the intersection of the two projection spaces is a singleton. In other words,

$$\inf_{P\in\Pi\left(p(\cdot,t_i)_{i=0}^{N-1}\right)} \mathrm{KL}(P\|Q_{-\nabla\Psi,\sigma^2}) = 0 \iff P = Q_{-\nabla\Psi,\sigma^2}.$$

By construction, the iterates will produce a monotonically decreasing sequence of KL divergences between reconstructed laws and MLE estimated laws. With the above argument, we have that the KL divergences are bounded above by some $M > 0$, and from below by 0 (attained by the true gradient-flow SDE parameters $(-\nabla\Psi,\sigma^2)$). Thus, the sequence of KL divergences must converge. However, as noted in Shen et al. (2025), this does not imply that the arguments necessarily converge, since multiple pairs (P,Q) may produce the same divergence, such that the iterative algorithm gets stuck in a cycle. While this has not been observed empirically, and is impossible if the divergence is 0 (since there is a unique gradient-flow SDE satisfying the marginal constraints), it is an open question whether the iterative scheme will converge to the unique optimal parameters. Indeed, while convergence of the iterative SB refinement algorithm has been proven in the case where the family of admissible probability transition densities is convex (Shen et al., 2025, Proposition 1), this condition does not generally hold, and is false for the family of gradient-flow SDEs.

D.2 Runtime

The runtime of `nn-APPEX` is split across its three subprocedures:

1. **Trajectory inference** (13): solving the multi-marginal SB problem with respect to the current reference SDE to infer and sample from a law on paths.

2. **MLE drift estimation** (14): training a neural network whose parameters minimize the objective (16) over the inferred paths.

3. **MLE diffusion estimation** (15): computing the quadratic variation (17) over the inferred paths, conditioned on the drift estimate.

We note that `nn-APPEX`'s runtime will vary for different implementations of these subprocedures. In particular, there are multiple approaches for solving the multi-marginal SB problem. One approach is to consider all contiguous pairs of marginals, and then to apply iterative proportional fitting (Sinkhorn's algorithm) on each pair to determine the distribution over couplings (Shen et al., 2025; Guan et al., 2024). To enforce consistency across transitions, we instead apply a multi-marginal iterative proportional fitting across all marginals, see (Marino and Gerolin, 2020)[(4.9)]. This algorithm is analogous to the two-marginals setting, but it instead rescales per-time slices for each marginal, rather than just the endpoints. We observed better accuracy for this approach, given the same stopping criteria and maximum number of iterations, and were thus able to reduce runtime with relatively fewer iterations. We similarly note that MLE drift estimation can be implemented using different neural network architectures and hyperparameters. In all experiments we use 128 neurons per layer, but this can be changed with some alterations to both the flexibility and the runtime (Table 2).

Table 2: Average runtime of drift estimation step (14) using different neural network widths.

NN Width	Avg Time (s)
16	1.32
32	1.85
64	3.20
128	6.57
256	13.20

In contrast to the drift estimation, the diffusion MLE estimate is a closed formula, and is thus cheap to compute. We report the average runtime (in seconds) of each of the three subprocedures for a single iteration, run on our main experiments in Table 3.

Table 3: Average runtime per iteration (seconds) for each subprocedure, aggregated across all potentials.

	Trajectory inference	Drift MLE	Diffusion MLE	Total
Average	3.842 ± 0.046	6.176 ± 0.061	0.011 ± 0.002	10.030 ± 0.065

Since we used 30 iterations of `nn-APPEX`, each SDE inference for our main experiments was approximately 5 minutes. Since the diffusion MLE is computationally negligible, `nn-APPEX` has virtually the same runtime as an analogous `SBIRR` algorithm. In contrast, `WOT` only comprises a single iteration, and thus has the fastest runtime.

Stopping criteria, such as thresholds on the drift and diffusion estimates of the iterates, or approximations of the KL objective based on the W^1,W^2 metrics or maximum mean discrepancy (MMD), can be enforced for better performance and faster runtime. However, we note that these evaluations tend to be computationally intensive themselves, since they either require evaluations on a discretized grid or an approximation of relative entropy.

E MAXIMUM LIKELIHOOD ESTIMATION FOR DRIFT AND DIFFUSION

Let the parameters of a gradient-flow SDE (1) be given by drift $-\nabla\Psi_\theta(x)$ and diffusivity σ^2. Then, if we apply the first-order Euler-Maruyama linear approximation, we have that $X_{t+\Delta t}|X_t \sim \mathcal{N}(X_t - \nabla\Psi_\theta(X_t)\Delta t, \sigma^2\Delta t)$.

Hence, for a single observation of a trajectory $x_t \to x_{t+\Delta t}$,

$$p(x_{t+\Delta t}|x_t) = \frac{1}{(2\pi\sigma^2\Delta t)^{d/2}} \exp\left(-\frac{\|x_{t+\Delta t} - x_t + \nabla\Psi_\theta(x_t)\Delta t\|^2}{2\sigma^2\Delta t}\right).$$

The full negative log-likelihood function is therefore given by

$$l(\theta, \sigma^2|x_t, x_{t+\Delta t}) = -\log\left(p(x_{t+\Delta t}|x_t)\right) = \frac{d}{2}\log(2\pi\Delta t) + \frac{d}{2}\log\big(\sigma^2\big) + \frac{\|x_{t+\Delta t} - x_t + \nabla\Psi_\theta(x_t)\Delta t\|^2}{2\sigma^2\Delta t}. \tag{33}$$

The MLE parameters $(\hat{\theta}, \hat{\sigma}^2)$ minimize $l(\theta, \sigma^2|x_t, x_{t+\Delta t})$. However, note that θ only appears in the numerator of the last term. It follows that $\hat{\theta}$ minimizes the squared error of the finite difference, which we denote

$$\ell(\theta) = \|x_{t+\Delta t} - x_t + \nabla\Psi_\theta(x_t)\Delta t\|^2. \tag{34}$$

If we observe M independent trajectories over $N-1$ time steps (N observed times), let $X = \big\{x^{(m)}_{i\Delta t}, x^{(m)}_{(i+1)\Delta t} : m \in \{1, \cdots, M\}, i \in \{0, \cdots, N-2\}\big\}$. We similarly observe that the negative log-likelihood function is given by

$$l\left(\theta, \sigma^2|X\right) = -\log\left(\prod_{m=1}^{M}\prod_{i=0}^{N-2} p(x^{(m)}_{(i+1)\Delta t}|x^{(m)}_{i\Delta t})\right) = -\sum_{m=1}^{M}\sum_{i=0}^{N-2}\log\Big(p(x^{(m)}_{(i+1)\Delta t}|x^{(m)}_{i\Delta t})\Big),$$

and it follows that $\hat{\theta}$ would minimize the mean squared error

$$\ell(\theta) = \sum_{m=1}^{M}\sum_{i=0}^{N-2}\left\|x^{(m)}_{(i+1)\Delta t} - x^{(m)}_{i\Delta t} + \nabla\Psi_\theta(x^{(m)}_{i\Delta t})\Delta t\right\|_2^2. \tag{35}$$

To derive the diffusion MLE estimator, $\hat{\sigma}^2$, we first fix $\theta = \hat{\theta}$, since the drift parameter can be independently optimized. Then, we solve $\frac{\partial}{\partial\sigma^2} l\left(\hat{\theta}, \sigma^2|X\right) = 0$, and obtain

$$\begin{aligned}
0 &= \sum_{m=1}^{M}\sum_{i=0}^{N-2}\frac{d}{2\sigma^2} - \frac{1}{2\sigma^4\Delta t}\sum_{m=1}^{M}\sum_{i=0}^{N-2}\left\|x^{(m)}_{(i+1)\Delta t} - x^{(m)}_{i\Delta t} + \nabla\Psi_{\hat{\theta}}(x^{(m)}_{i\Delta t})\Delta t\right\|_2^2 \\
0 &= dM(N-1) - \frac{1}{\sigma^2\Delta t}\ell(\hat{\theta}) \\
\sigma^2 &= \frac{1}{d\,M\,(N-1)\,\Delta t}\ell(\hat{\theta})
\end{aligned}$$

F POTENTIALS

Our experiments consider the following potentials for simulating gradient-flow SDEs. In particular, these are the potentials from Terpin et al. (2024), which admit a valid stationary Gibbs distribution $p_{\text{eq}} = \frac{1}{Z}\exp\left(-\frac{\Psi}{2\sigma^2}\right)$.

Bohachevsky $$\Psi(x) = 10\big(x_1^2 + 2x_2^2 - 0.3\cos(3\pi x_1) - 0.4\cos(4\pi x_2)\big) \tag{36}$$

Oakley–O'Hagan $$\Psi(x) = 5\sum_{i=1}^{2}\big(\sin(x_i) + \cos(x_i) + x_i^2 + x_i\big) \tag{37}$$

Quadratic $$\Psi(x) = 5\,\|x\|^2 \tag{38}$$

Styblinski–Tang $$\Psi(x) = \tfrac{1}{2}\sum_{i=1}^{2}\big(x_i^4 - 16x_i^2 + 5x_i\big) \tag{39}$$

Wavy plateau $$\Psi(x) = \sum_{i=1}^{2}\big(\cos(\pi x_i) + \tfrac{1}{2}x_i^4 - 3x_i^2 + 1\big) \tag{40}$$

Each of these potentials is smoothly differentiable, and we note that they each lead to well-defined strong solutions to the SDE (1). Indeed, in order to satisfy the linear growth condition, $\|\nabla\Psi(x)\| \leq K(1 + \|x\|)$, we can multiply potentials by a smooth cutoff function, which is 1 in $\|x\| \leq M$ and 0 in $\|x\| > 2M$. Choosing $M = 100$ for our experiments does not alter data generation compared to the raw setting. Similarly, since the cutoff function ensures that Ψ is supported in a compact subset of $\mathbb{R}^d$, it follows that smoothness of Ψ implies that the drift $-\nabla\Psi$ is Lipschitz.

G ADDITIONAL EXPERIMENTS

Software and hardware details. The code in this paper was adapted from two public code repositories (`APPEX`: https://github.com/guanton/APPEX and `JKOnet*`: https://github.com/antonioterpin/jkonet-star). All computations are performed on a 2024 MacBook Pro with 16GB RAM and an Apple M4 chip. The code was adapted to design the experiments, visualize the data, and interpret the results. It is available in the supplementary material.

G.1 Experimental Setup

Since the objective of our main experiments was to evaluate the inferential power of different Schrödinger Bridge methods, we considered the methods `WOT`, `SBIRR`, and `nn-APPEX` (ours), and we simulated data from a range of population dynamics for each SDE, by randomizing the initial distribution over Gaussian mixtures. We perform two additional experiments, such that we use the same experimental setup, but consider two particular initial distributions: the uniform distribution over the region of interest $[-4, 4]^2$ and the stationary Gibbs distribution p_{eq}. In order to sample from the Gibbs distributions $\frac{1}{Z}\exp\left(-\frac{2\Psi(x)}{\sigma^2}\right)$, we generate trajectories from $p_0 = \text{Unif}([-4, 4]^2)$ and run 200 steps of the SDE with step size $\Delta t = 0.01$. Indeed, each of the potentials (36)-(40) satisfies the Poincaré inequality, which ensures that marginals converge exponentially quickly to the stationary Gibbs distribution (Pavliotis, 2014, Theorem 4.4).

We also perform additional experiments comparing our method, `nn-APPEX`, against the state-of-the-art variational method, `JKOnet*` (Terpin et al., 2024). We implement `JKOnet*` with default hyperparameters, and use 100 training epochs, noting rapid convergence of the objective. We verified in a separate experiment that running with $10,000$ epochs did not meaningfully change performance. The experimental setting is the main setting over random GMM initializations, in order to test on different population dynamics. Since `JKOnet*` also jointly estimates drift and diffusion for the gradient-flow SDE (1), we additionally report diffusivity MAE for this experiment.

G.2 Results

The uniform initial distribution $p_0 \sim \text{Unif}[-4, 4]^2$ represents an idealized benchmark (Terpin et al., 2024) while the stationary distribution $p_0 \sim p_{\text{eq}}$ represents the non-identifiable setting. Results comparing the three SB methods, `WOT`, `SBIRR`, and `nn-APPEX`, are summarized for the uniform initialization in Figure 4 and for the non-identifiable stationary initialization in Figure 5.

We also present results from our main experimental setting (random GMM initializations), which compare our method `nn-APPEX` against `JKOnet*` in Figure 6. The results show that `nn-APPEX` accurately infers both the drift and diffusivity to high precision for all SDEs, and outperforms `JKOnet*`except for the diffusivity estimate of the Bohachevsky potential. We note that the Bohachevsky potential is particularly difficult to estimate, as it has the highest magnitude, and converges to its Gibbs distribution p_{eq} at a much faster rate than the other SDEs, all while exhibiting highly periodic and nonlinear dependencies. For this reason, the MLE diffusion estimate by `nn-APPEX` likely achieves biased estimates, since the adjusted quadratic variation (17) is impacted by poor drift estimation. This is also consistent with previous results (Terpin et al., 2024), which showed that the Bohachevsky potential is particularly difficult to estimate in dimension $d = 2$.

Tables 4 and 5 provide further details about our simulated experimental results from the main text. Table 4 shows that the performance gap between `SBIRR` and `nn-APPEX` is highest in low potential areas, i.e. regions where the magnitude of the gradient field $\nabla\Psi$ is small. Further, most potentials from this benchmark have high magnitudes in a large portion of the evaluation grid, while `SBIRR` particularly struggles in low potential areas (see Figure 1 and Figure 2). Intuitively, if the potential is strong enough, then it can still be approximated even if diffusion

is misspecified. To emphasize that `nn-APPEX` is especially useful for inference in low potential zones, Table 4 stratifies the results based on the magnitude of the true gradient at each evaluation point. All results show that `nn-APPEX` outperforms `SBIRR`, especially in low potential zones, as one would expect from Figures 1 and 2. This is likely due to diffusion having an outsized noisy impact on the data in these regions. Moreover, Table 5 shows that the median (used in the main text) is actually a favourable evaluation metric for `SBIRR` compared to the mean (not shown in main text). Indeed, across all potentials, the gap in drift estimation performance between `SBIRR` and `nn-APPEX` is larger for the mean compared to the median. This occurs since `SBIRR` performs fairly well most of the time, but can still frequently catastrophically fail due to poor initial diffusion estimates.

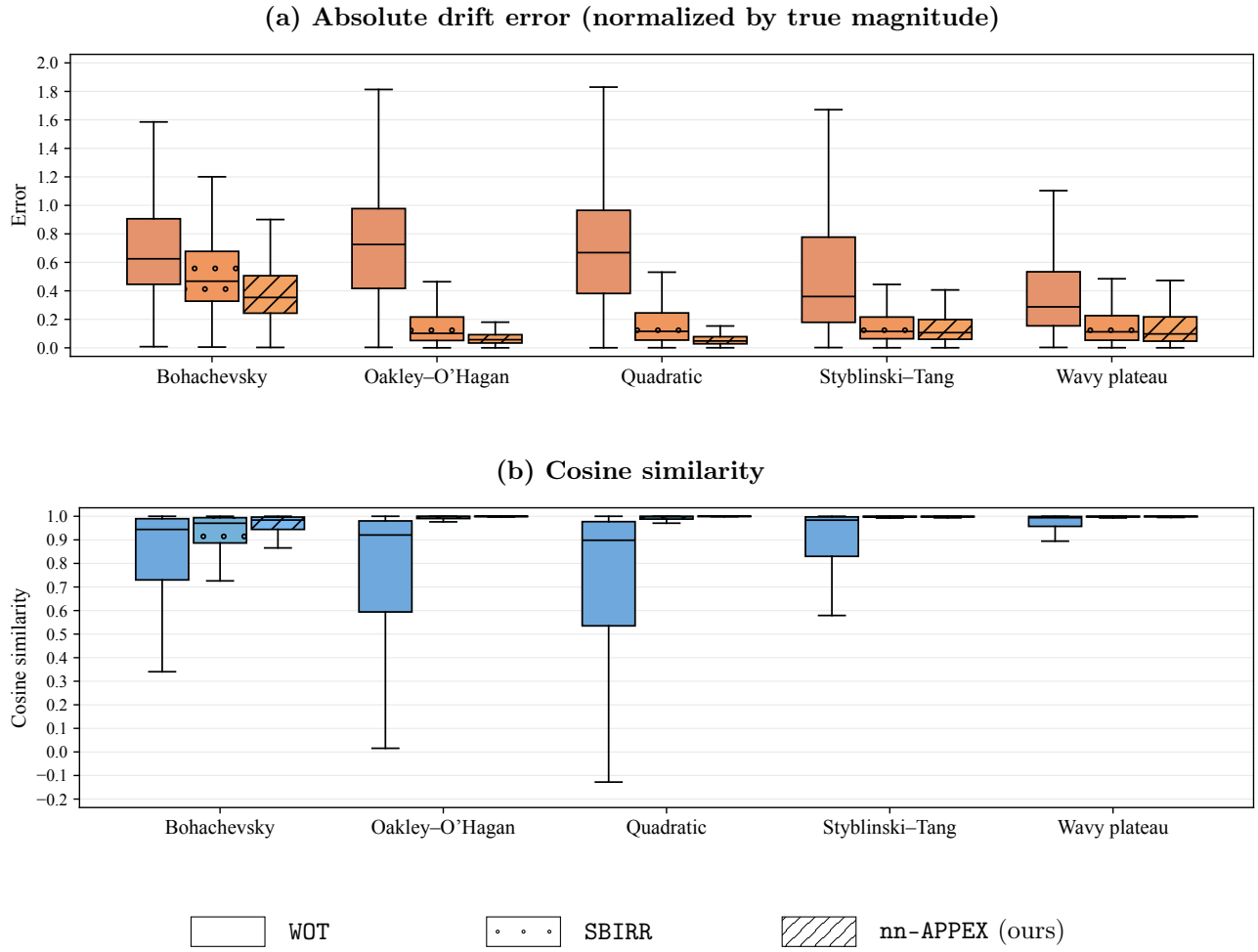


Figure 4: The ability of different Schrödinger Bridge methods to infer the gradient-flow drift is evaluated across five potentials using (a) normalized absolute error and (b) cosine similarity. Here, methods observe three marginals with a uniform initial distribution in the region of interest, i.e. $p_0 \sim \mathrm{Unif}[-4, 4]^2$. Box-and-whisker plots over 10 seeds show `nn-APPEX` performs best across all potentials.

G.3 Comparing `APPEX` and `nn-APPEX`

As a final experiment, we compare `APPEX` and `nn-APPEX` on a linear SDE problem in order to ensure `nn-APPEX` is fully expressive and can compete with `APPEX` in the linear domain for which `APPEX` was designed. Results are displayed in Table 6. As expected, `APPEX` does quite a bit better than `nn-APPEX` in terms of drift estimation since we have a closed form MLE estimate. `APPEX`'s normalized MAE score is significantly better, while their cosine similarities are comparable, pointing to the fact that `nn-APPEX` may do poorly in low drift regions. Interestingly, `nn-APPEX` does slightly better in terms of diffusion estimation (0.055 vs 0.074).

Table 4: Mean difference in normalized absolute drift error (and cosine similarity) between `nn-APPEX` and `SBIRR`. Results are displayed for each potential and aggregated with respect to low potential zones (gradient magnitude is less than median) and high potential zones (gradient magnitude is higher than median). Higher numbers indicate a greater outperformance of `nn-APPEX` over `SBIRR`.

	Gradient Magnitude	
Potential	**Low**	**High**
Bohachevsky	0.293 (0.074)	0.184 (0.018)
Oakley-O'Hagan	0.612 (0.100)	0.253 (0.042)
Quadratric	1.397 (0.098)	0.654 (0.054)
Styblinski-Tang	0.128 (0.031)	0.075 (0.005)
Wavy Plateau	2.199 (0.035)	0.185 (0.012)

Table 5: Performance comparison of `SBIRR` and `nn-APPEX` methods across various benchmark potential functions, reporting mean and median error metrics.

Potential	**Method**	**Mean**	**Median**
Bohachevsky	`SBIRR`	0.9135	0.5804
	`nn-APPEX`	0.6749	0.4561
Oakley-O'Hagan	`SBIRR`	0.5783	0.1786
	`nn-APPEX`	0.1463	0.0839
Quadratic	`SBIRR`	1.1515	0.1911
	`nn-APPEX`	0.1261	0.0739
Styblinski-Tang	`SBIRR`	0.5141	0.2245
	`nn-APPEX`	0.4126	0.1509
Wavy Plateau	`SBIRR`	2.1761	0.2107
	`nn-APPEX`	0.9842	0.1960

Table 6: Inference results for `nn-APPEX` and `APPEX` given the quadratic potential setting (see Appendix F). Both methods are given three marginals with Gaussian mixture model initialization.

Method	**Normalized MAE**	**Cosine Similarity**	**Diffusivity error**
`nn-APPEX`	0.1550	0.9925	**0.055**
`APPEX` (linear)	**0.0135**	**0.9996**	0.074

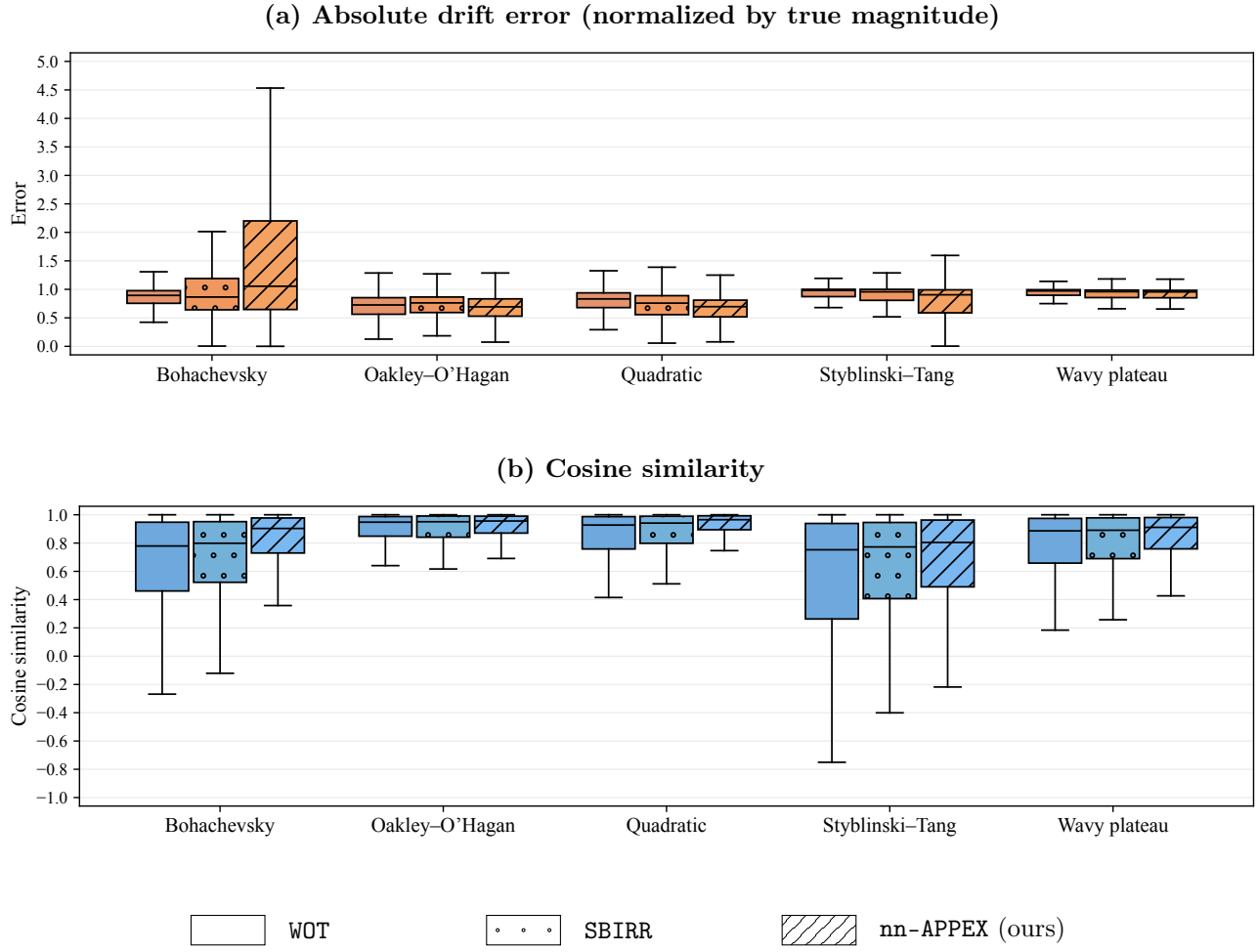


Figure 5: The ability of different Schrödinger Bridge methods to infer the gradient-flow drift is evaluated across five potentials using (a) normalized absolute error (lower is better) and (b) cosine similarity (higher is better). Methods observe three marginals with the initial distribution equal to the SDE's stationary Gibbs distribution, $p_0 \sim p_{\mathrm{eq}}$. Aggregated box-and-whisker plots over 10 seeds show that all methods perform similarly poorly, corroborating Proposition 3.1: the true gradient-flow drift is not identifiable without knowing the diffusivity.

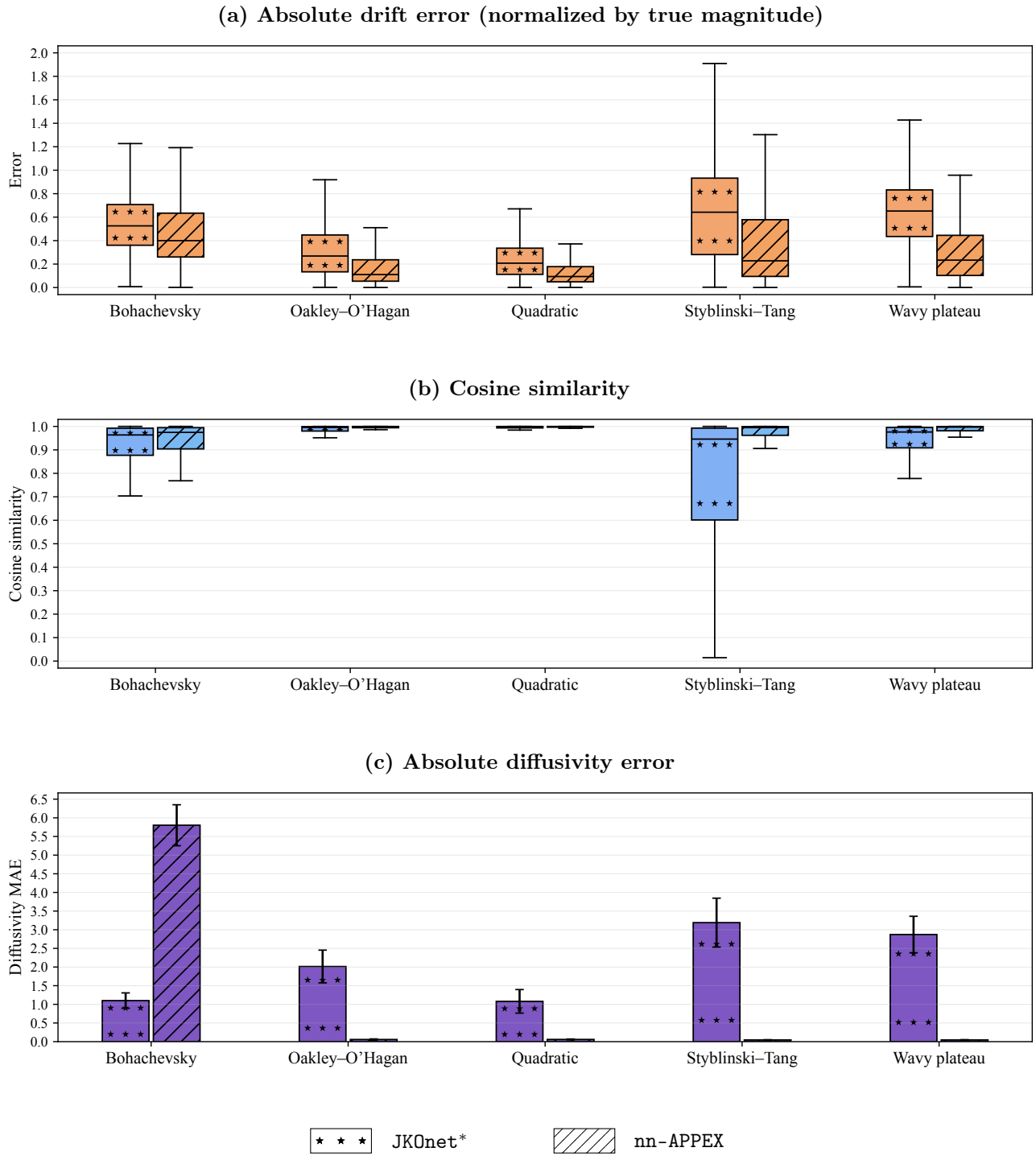


Figure 6: The ability of the variational method `JKOnet*` and `nn-APPEX` (ours) to infer the gradient-flow drift is evaluated across five potentials using (a) normalized absolute error (lower is better) and (b) cosine similarity (higher is better), and their ability to infer the diffusivity is evaluated using (c) absolute error (lower is better). Both methods are given samples from three distinct marginals, such that the initial distribution is a Gaussian mixture model with randomly initialized components. The plots aggregated from 10 seeds show that our method, `nn-APPEX`, achieves better estimation of the drift and diffusion in this setting, except for the Bohachevsky potential.