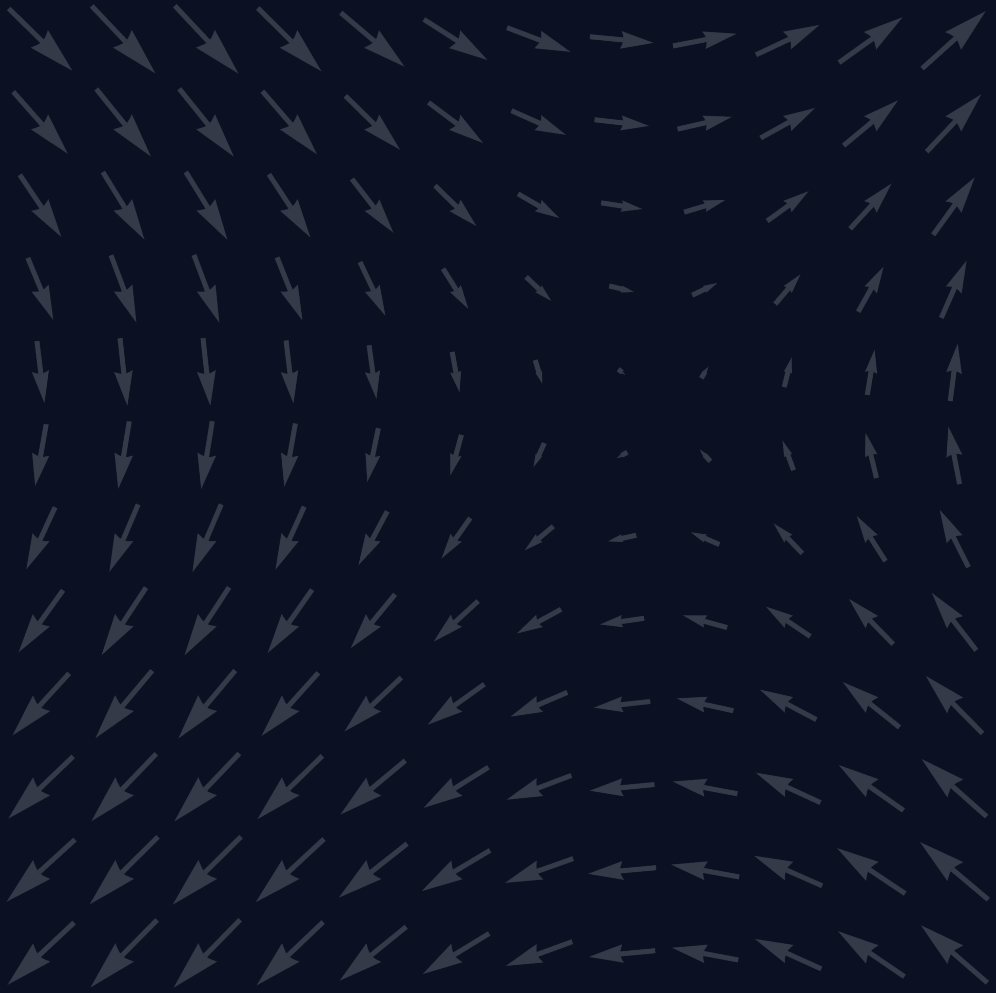
Neural Optimal Transport for Dynamical Systems

Methods and Applications in Biomedicine

Charlotte Bunne



Diss.-No. ETH 29XXX

CHARLOTTE BUNNE

NEURAL OPTIMAL TRANSPORT FOR DYNAMICAL SYSTEMS

METHODS AND APPLICATIONS IN BIOMEDICINE

DISS. ETH NO. ?

NEURAL OPTIMAL TRANSPORT FOR DYNAMICAL SYSTEMS

METHODS AND APPLICATIONS IN BIOMEDICINE

A dissertation submitted to attain the degree of

DOCTOR OF SCIENCES of ETH ZURICH
(Dr. sc. ETH Zurich)

presented by

CHARLOTTE BUNNE
M. sc. ETH Zurich

born on 29 August 1995

accepted on the recommendation of

Prof. Dr. Andreas Krause, examiner
Prof. Dr. Marco Cuturi, co-examiner
Prof. Dr. Lucas Pelkmans, co-examiner
Prof. Dr. Jure Leskovec, co-examiner

2023

Charlotte Bunne: *Neural Optimal Transport for Dynamical Systems,* © 2023

DOI: 10.3929/ethz-a-

Und was in schwankender Erscheinung schwebt,
Befestiget mit dauernden Gedanken.
— Johann Wolfgang von Goethe, *Faust I* (1808)

ABSTRACT

English abstract here.

ZUSAMMENFASSUNG

Deutsche Zusammenfassung hier.

ACKNOWLEDGEMENTS

First and foremost, I would like to express my sincere gratitude to my advisors Andreas Krause and Marco Cuturi. ... I am very grateful to Anne Carpenter and Shantanu Singh, my mentors at the Broad Institute of MIT and Harvard. ... I will be forever grateful for this. Further, I would also like to thank ... and Jure Leskovec for being on my thesis committee, and for providing me with invaluable feedback. Aviv, in particular, ...

The thesis would not be the same without the longstanding collaboration with Lucas Pelkmans and Gabriele Gut. ...

I am profoundly indebted to other co-authors with whom I worked on many projects throughout the past years, in particular, Gunnar Rätsch, Stefan Stark, Ya-Ping Hsieh, Vignesh Ram Somnath, Frederike Lübeck, Matteo Pariset, Valentin De Bortoli, Octavian Ganea, Laetitia Meng-Papaxanthos, and Philippe Schwaller. Thank you especially to Ya-Ping Hsieh and Stefan Stark for hour-long discussions on various projects. ...

Further, I am very grateful for my first academic mentors, Stefanie Jegelka, David Alvarez Melis, Roland Eils, Thomas Höfer, and Lisa Buchauer. ... I would not have started my scientific journey without the Life-Science Lab of the German Cancer Research Center. It is what sparked my interest in biology and engineering. It ...

Thanks to all members of the Learning and Adaptive Systems group for creating an excellent research environment. In particular, my office mates Parnian Kassraie, Lars Lorch, and Jonas Rothfuss. Thank you, Rita Klute, for entangling the jungle of bureaucracy and never getting tired of my administrative requests.

I am very grateful to my parents Nele and Egon and siblings Kaspar, Henriette, and Frieder for there endless love, advice, and support. Lastly, thank you to Pol: "Jeder Zustand, ja jeder Augenblick [mit dir] ist von unendlichem Wert, denn er ist der Repräsentant einer ganzen Ewigkeit[1]."

1 Johann Peter Eckermann and Johann Wolfgang von Goethe. Gespräche mit Goethe in den letzten Jahren seines Lebens. 1823-1832: 2. Vol. 2. FA Brockhaus, 1836.

CONTENTS

NOTATION

Σ_d Probability simplex of size d

ACRONYMS

BDT Black-Derman-Toy.
BM Brownian motion.

DDPM denoising diffusion probabilistic model.
DSB diffusion Schrödinger bridge.

ESC embryonic stem cell.

GSB Gaussian Schrödinger bridge.

ICNN input convex neural network.
IPF iterative proportional fitting.

JKO Jordan-Kinderlehrer-Otto.

MMD maximum-mean-discrepancy.

NF normalizing flows.

o.o.d. out-of-distribution.
OT optimal transport.
OU Ornstein-Uhlenbeck.

PCA principal component analysis.
PICNN partially input convex neural network.

SB Schrödinger bridge.
SDE stochastic differential equation.
SMLD score matching with Langevin dynamics.

VESDE variance exploding SDE.
VPSDE variance preserving SDE.

1

INTRODUCTION

The balance of nature is not a status quo; it is fluid, ever shifting, in a constant state of adjustment. Man, too, is part of this balance.
— Rachel Carson, *Silent Spring* (1962)

Biology is determined by structure, patterns, and dynamics at various scales, ranging from molecular interactions to organismal behavior. At the lower end of that scale, single-cell genomics and transcriptomics provide now a direct window, with a resolution that was deemed unthinkable two decades ago, into the molecular makeup of individual cells, capturing vividly the inner workings of cells at any point in time. Similarly, advances in imaging technology provide tools to map the spatial organization of tissues and organs at the cellular and subcellular level, improving our understanding of key physiological processes.The ability of single-cell high-throughput methods to produce routinely millions of data points holds multiple promises. They do, however, come with an important limitation: they produce data that are not *aligned*, namely, such methods are destructive assays, meaning that the same cell cannot be observed twice. This limitation is particularly acute in the field of personalized medicine, where the goal is precisely to understand the dynamic response of a patient's cells to a stimulus, and would therefore rest, in theory, on the ability to observe the same cell before and after treatment. Similarly, most single-cell technologies require the physical dissection and dissociation of tissues and organs, resulting in a loss of spatial information. These issues are well known challenges, and many technologies have tried to circumvent such destructive steps, notably through spatial-omics. The scalability of such methods does, however, lag behind that of single-cell sequencing, which calls for algorithmic solutions to this problem.

Our goal in this review is to highlight that the common thread in all of these problems is the recurring need to realign datasets, and that such problems can be solved using optimal transport (OT) theory (Villani, 2003; Santambrogio, 2015). OT theory, a major research area in pure mathematics in recent decades (with Fields medals awarded to Villani in 2010 and Figalli in 2018), has emerged as a contender to fill in that gap *in silico*. OT is best

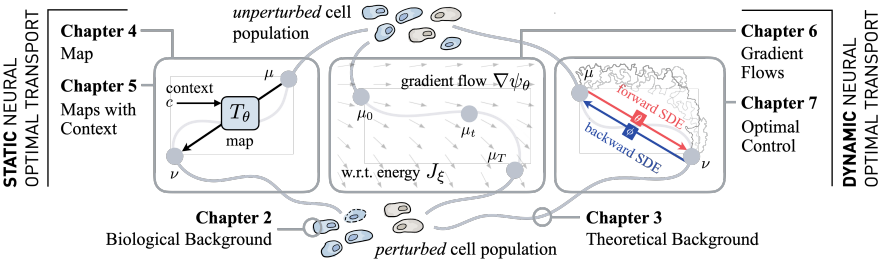


Figure 1.1: Overview on ...

described as a toolbox that allows reconstructing how a "source" population (represented as one probability distribution) can morph efficiently into another "target" population, given only source and target samples. Taking for the source distribution a sample of cells pre-stimuli, and for the target, another sample of cells post-stimuli, OT can reconstruct the unobserved process and provide an informed guess to define a "transport map" that relates these two cell populations.

Applied to the analysis and modeling of single-cell biology problems, OT has been used to infer the distributions of cells' ancestors and descendants along development (Schiebinger et al., 2019), perform trajectory inference (Bunne et al., 2022b; Forrow and Schiebinger, 2021; Bunne et al., 2023a; Lavenant et al., 2021; Schiebinger et al., 2019; Tong et al., 2020; Yang et al., 2020; Zhang et al., 2021; Chizat et al., 2022), predict perturbation responses (Bunne et al., 2023b; Yang and Uhler, 2019; Lübeck et al., 2022), integrate multi-omics data of different modalities (Demetci et al., 2022), infer cell-cell similarity (Huizing et al., 2022), and integrate across scales (e.g., morphology and molecular profiling) (Yang et al., 2021). The increasing data complexity across multiple levels of biological organization, from molecular and cellular through spatial profiling (Moriel et al., 2021) of tissues, and imaging of organs, cement further the status of OT as an indispensable framework for high-throughput, multimodal, and multi-scale molecular, cell, tissue, and organ biology. The effectiveness of OT comes, however, with drawbacks: because the theory builds on extremely sophisticated mathematics that blends optimization (Cuturi, 2013; Cuturi et al., 2022), stochasticity (Chizat et al., 2022; Bunne et al., 2023a) and partial differential equations (Bunne et al., 2022b), and, more recently, deep learning (Tong et al., 2020; Bunne et al., 2023b, 2022a; Yang and Uhler, 2019; Lübeck et al., 2022; Yang et al., 2021), its computations are challenging even by modern ML standards.

In this primer, we introduce the mathematical and computational principles of OT, with the goal of facilitating its use by researchers that wish to apply to novel applications. We provide the reader with intuitive explanations of how seemingly unrelated mathematical approaches for analyzing single-cell data can be unified through OT theory, and how that theory has triggered recent advances in deep learning. We provide an overview of the broad range of biological applications, demonstrating the successes of OT in the field, especially within the field of single-cell biology. With its rich properties, astonishing mathematical connections, and its innovative numerical implementations (Cuturi et al., 2022), OT makes for an exciting avenue of future work to make novel biological discoveries, infer personalized cancer therapies from single-cell patient samples, and push the boundaries of regenerative medicine.

2

DYNAMICAL PROCESSES IN BIOMEDICINE

The results suggest a helical structure (which must be very closely packed) containing probably 2, 3 or 4 coaxial nucleic acid chains per helical unit and having the phosphate groups near the outside.

— Rosalind Franklin, *Report* (1952)

...

3

OPTIMAL TRANSPORT FOR DYNAMICAL SYSTEMS

The power of a theory is exactly proportional to the diversity of situations it can explain.
— Elinor Ostrom, *Governing the Commons* (1990)

Optimal transport theory (Santambrogio, 2015) plays now a prominent role in the machine learning toolbox and has become within a few years the go-to framework to analyze, model, and solve an ever-increasing variety of tasks involving probability measures. This is best exemplified by its increasing importance to fitting generative models, where the goal is to learn a map (Arjovsky et al., 2017; Genevay et al., 2018; Salimans et al., 2018), or more generally a diffusion (Song et al., 2021; De Bortoli et al., 2021b) to morph a simple measure (e.g., Gaussian) onto a data distribution of interest (e.g., images). This is also apparent in the many applications that use OT to align probability measures that have since arisen, e.g., to transfer label knowledge between datasets (Flamary et al., 2016; Singh and Jaggi, 2020), to analyze sampling schemes (Dalalyan, 2017), or study population trajectories (Schiebinger et al., 2019).

We will start the tutorial by emphasizing the centrality of the optimal *assignment* problem in various ML problems, motivated both by its conceptual simplicity and practical ability to compare sets of the same size. We will work out the mathematical foundations of optimal transport as an extension from these simple principles, recall its mathematical history from Monge (1781) and Kantorovich (1942) to modern Fields medal winners Villani (2009) and Figalli (2010). We focus next on the numerical resolution of the Kantorovich problem, its statistical (Rigollet and Stromme, 2022; Genevay et al., 2019) and computational challenges, to motivate by-now classic algorithms (Cuturi, 2013; Chizat et al., 2018), and their large-scale extensions (Altschuler et al., 2019; Scetbon et al., 2021). We will also mention quadratic (Gromov) OT (Mémoli, 2011) and its efficient computation (Solomon et al., 2016; Scetbon and Cuturi, 2022).

We expect most of the audience to be familiar with the role of OT losses in the literature, e.g., for structured prediction (Frogner et al., 2015; Janati et al., 2020a) or generative model fitting (Yang and Uhler, 2019; Arjovsky et al., 2017; Salimans et al., 2018; Genevay et al., 2018). Yet, we will gradually

move away from the usual emphasis on OT distances (e.g., Wasserstein) towards a focus on the Monge *map*, which provides an actionable way to flow from one probability distribution onto another. This section will start with a complete proof of the celebrated Brenier theorem for general costs, which will require an introduction to the notion of c-concavity. Once this quintessential result of OT theory is established, we will particularize it to translation-invariant costs and bridge it to the flurry of neural approaches that have been proposed in the literature. This comprises approaches that are a direct consequence of the Brenier theorem, to model Monge maps as gradients of convex functions, parameterized through input convex neural networks (ICNN) (Amos et al., 2017; Huang et al., 2021a; Makkuva et al., 2020; Korotin et al., 2021b; Lübeck et al., 2022; Bunne et al., 2022a), via regularizers (Uscidda and Cuturi, 2023), amortized optimization (Amos, 2023; Amos et al., 2022), or entropic maps (Pooladian and Niles-Weed, 2021; Pooladian et al., 2023; Divol et al., 2022; Cuturi et al., 2023).

Benamou and Brenier (2000) showed how the dynamic point of view offers an alternate and intuitive interpretation of optimal transport with links to fluid dynamics that surprisingly leads to a convex optimization problem that can be parameterized through normalizing flows (Tong et al., 2020). We will further highlight connections of OT to PDEs such as Fokker-Planck-like equations through the Jordan, Kinderlehrer, and Otto scheme: In recent works (Bunne et al., 2022b; Alvarez-Melis et al., 2022; Mokrov et al., 2021; Benamou et al., 2016a) it has found application in inferring the evolution of populations over time, crucial in many scientific disciplines when for instance, observing a population of cells in biology. Beyond PDEs, we will explore the relation between the optimal transport problem and the Schrödinger bridge problem from stochastic control. It represents a key connection that has recently fueled the development of diffusion Schrödinger bridges (De Bortoli et al., 2021b; Chen et al., 2021b; Bunne et al., 2023a; Liu et al., 2022). Compared to classical diffusion-based generative models (Daniels et al., 2021; Song et al., 2021), these algorithms allow interpolation between complex distributions. Extended to the Riemannian geometry (Thornton et al., 2022; De Bortoli et al., 2022), it has found applications in molecular dynamics (Holdijk et al., 2022) and cell differentiation processes (Tong et al., 2023; Bunne et al., 2023a).

3.1 STATIC OPTIMAL TRANSPORT

...

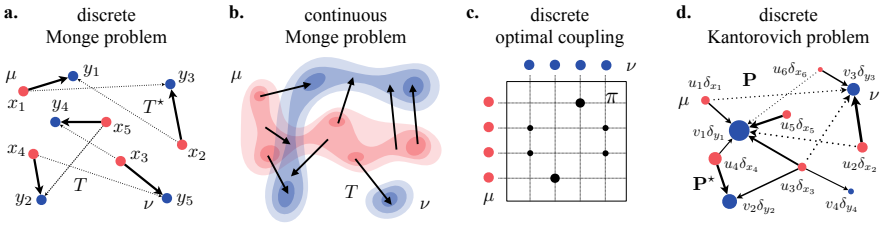


Figure 3.1: Overview on different formulations of the static OT problem for discrete and continuous measures. a. ... **b.** ... **c.** ... **d.** .... Figure adapted from Peyré and Cuturi (2019).

3.1.1 ...

For two probability measures μ, ν in $\mathcal{P}(\mathbb{R}^d)$, their squared 2-Wasserstein distance is

$$W_2^2(\mu,\nu) = \inf_{\gamma \in \Gamma(\mu,\nu)} \iint \|x-y\|_2^2 \gamma(dx,dy), \tag{3.1}$$

where $\Gamma(\mu,\nu)$ is the set of couplings on $\mathbb{R}^d \times \mathbb{R}^d$ with respective marginals μ, ν. When instantiated on finite discrete measures, such as $\mu = \sum_{i=1}^n a_i \delta_{x_i}$ and $\nu = \sum_{j=1}^m b_j \delta_{y_j}$, this problem translates to a linear program, which can be regularized using an entropy term (Cuturi, 2013; Peyré and Cuturi, 2019). For $\varepsilon \geq 0$, set

$$W_\varepsilon(\mu,\nu) := \min_{\mathbf{P} \in U(a,b)} \langle \mathbf{P}, [\|x_i - y_j\|^2]_{ij} \rangle - \varepsilon H(\mathbf{P}), \tag{3.2}$$

where $H(\mathbf{P}) := -\sum_{ij} \mathbf{P}_{ij}(\log \mathbf{P}_{ij} - 1)$ and the polytope $U(a,b)$ is the set of $n \times m$ matrices $\{\mathbf{P} \in \mathbb{R}_+^{n \times m}, \mathbf{P}\mathbf{1}_m = a, \mathbf{P}^\top \mathbf{1}_n = b\}$. Notice that the definition above reduces to the usual (squared) 2-Wasserstein distance when $\varepsilon = 0$. Setting $\varepsilon > 0$ yields a faster and differentiable proxy to approximate W_0, but introduces a bias, since $W_\varepsilon(\mu,\mu) \neq 0$ in general. In the rest of this work, we therefore use the *Sinkhorn divergence* (Ramdas et al., 2017; Genevay et al., 2019; Salimans et al., 2018; Feydy et al., 2019) as a valid non-negative discrepancy,

$$\overline{W}_\varepsilon(\mu,\nu) := W_\varepsilon(\mu,\nu) - \frac{1}{2}\left(W_\varepsilon(\mu,\mu) + W_\varepsilon(\nu,\nu)\right). \tag{3.3}$$

Optimal transport plays a pair of roles: inducing a mathematically well-characterized distance measure between distributions as well as providing a geometry-based approach to realize couplings between two probability

distributions. Let μ and ν be two measures in $\mathbb{R}^d$. The optimal transport problem by Monge (1781) is defined as

$$\arg\min_{T:T_\sharp\mu=\nu} \mathbb{E}_{X\sim\mu}\|X - T(X)\|_2^2, \tag{3.4}$$

where T corresponding to the smallest cost is the optimal transport map. Kantorovich (1942) provided a relaxation to this non-convex and difficult-to-solve problem, which reads

$$W(\mu,\nu) = \inf_{\gamma\in\Gamma(\mu,\nu)} \mathbb{E}_{(X,Y)\sim\gamma}\|X - Y\|_2^2, \tag{3.5}$$

where the polytope $\Gamma(\mu,\nu)$ is $\{\gamma \in \mathbb{R}_+^{n\times m}, \gamma\mathbf{1}_m = \mu, \gamma^\top\mathbf{1}_n = \nu\}$, describing the set of all couplings (or joint distributions) γ between μ and ν. The optimal transport plan γ thus corresponds to the coupling between two probability distributions minimizing the overall transportation cost. Computing optimal transport distances in (3.5) involves solving a linear program, and thus their computational cost is prohibitive for large-scale machine learning problems. Regularizing objective (3.5) with an entropy term results in significantly more efficient optimization (Cuturi, 2013) and differentiability w.r.t. its inputs, and thus commonly used as a loss function in machine learning applications.

$$f^\star := \arg\sup_{f \text{ convex}} \mathcal{E}_{\mu,\nu}(f) := \int_{\mathbb{R}^d} f^* d\mu + \int_{\mathbb{R}^d} f d\nu. \tag{3.6}$$

Problem (3.5) denotes the primal formulation for the Wasserstein distance. The corresponding dual introduced by Kantorovich in 1942 is a constrained concave maximization problem defined as

$$W(\mu,\nu) = \sup_{(f,g)\in\Phi_c} \mathbb{E}_\mu[f(x)] + \mathbb{E}_\nu[g(y)], \tag{3.7}$$

where the set of admissible potentials is $\Phi_c := \{(f,g) \in L^1(\mu)\times L^1(\nu) : f(x) + g(y) \le \frac{1}{2}\|x-y\|_2^2, \forall(x,y) d\mu\otimes d\nu \text{ a.e.}\}$ (Villani, 2003, Theorem 1.3). Villani (2003, Theorem 2.9) further simplifies the dual problem (3.7) over the pair of functions (f,g) to

$$W(\mu,\nu) = \underbrace{\frac{1}{2}\mathbb{E}\left[\|x\|_2^2 + \|y\|_2^2\right]}_{C_{\mu,\nu}} - \inf_{f\in\tilde{\Phi}} \mathbb{E}_\mu[f(X)] + \mathbb{E}_\nu\left[f^*(Y)\right], \tag{3.8}$$

where $\tilde{\Phi}$ is the set of all convex functions in $L^1(d\mu)\times L^1(d\nu)$, $L^1(\mu) := \{f \text{ is measurable } \& \int f d\mu < \infty\}$, $f^*(y) = \sup_x\langle x,y\rangle - f(x)$ is f's convex

conjugate, and the optimal transport plan corresponds to gradient of the convex conjugate, $\gamma = \nabla f^\star$. Villani (2003, Theorem 2.9) then proves the existence of an optimal pair (f, f^*) of lower semi-continuous proper conjugate convex functions on $\mathbb{R}^n$ minimizing (3.7).

Part I

STATIC NEURAL OPTIMAL TRANSPORT

4

NEURAL OPTIMAL TRANSPORT

Tout va par degré dans la nature, et rien par saut, et cette regle à l'égard des changements est une partie de ma loi de la continuité.
— Gottfried Wilhelm Leibniz, *Nouveaux essais sur l'entendement humain* (1765)

4.1 NEURAL OPTIMAL TRANSPORT SOLVERS

4.1.1 *Relaxing the Optimal Transport Dual*

Villani (2003, Theorem 2.9) further simplifies the dual problem (3.7) over the pair of functions (f, g) to

$$W(\mu,\nu) = \underbrace{\frac{1}{2}\mathbb{E}\left[\|x\|_2^2 + \|y\|_2^2\right]}_{\mathcal{C}_{\mu,\nu}} - \inf_{f\in\tilde{\Phi}} \mathbb{E}_\mu[f(X)] + \mathbb{E}_\nu\left[f^*(Y)\right], \tag{4.1}$$

where $\tilde{\Phi}$ is the set of all convex functions in $L^1(d\mu) \times L^1(d\nu)$, $L^1(\mu) := \{f \text{ is measurable } \& \int f d\mu < \infty\}$, $f^*(y) = \sup_x \langle x, y\rangle - f(x)$ is f's convex conjugate, and the optimal transport plan corresponds to gradient of the convex conjugate, $\gamma = \nabla f^\star$. Villani (2003, Theorem 2.9) then proves the existence of an optimal pair (f, f^*) of lower semi-continuous proper conjugate convex functions on $\mathbb{R}^n$ minimizing (3.7).

4.1.2 *Convex Neural Architectures*

Input convex neural networks are neural networks $\psi_\theta(x)$ with specific constraints on the architecture and parameters θ, such that their output is a convex function of some (or all) elements of the input x (Amos et al., 2017). We consider in this work ICNNs, such that the output is a convex function of the entire input x. A typical ICNN is a L-layer, fully connected network such that, for $l = 0, \ldots, L-1$:

$$z_{l+1} = a_l(W_l^x x + W_l^z z_l + b_l) \text{ and } \psi_\theta(x) = z_L, \tag{4.2}$$

where by convention, z_0 and W_0^z are 0, a_l are convex non-decreasing (non-linear) activation functions, $\theta = \{b_l, W_l^z, W_l^x\}_{l=0}^{L-1}$ are the weights and biases of the neural network, with weight matrices W_l^z associated to latent representations z that have non-negative entries. Since Amos et al. (2017)'s work, convex neural architectures have been further extended and shown to capture relevant models despite these constraints (Amos et al., 2017; Makkuva et al., 2020; Huang et al., 2021a). In particular, Chen et al. (2019) provide a theoretical analysis that any convex function over a convex domain can be approximated in sup norm by an ICNN.

4.1.3 ...

To learn the optimal transport map, Makkuva et al. build upon celebrated results by Knott and Smith (1984) and Brenier (1991), which relate the optimal solutions for the primal (3.5) and the dual form (3.7), to derive a min-max formulation replacing the convex conjugate in (4.6) (Makkuva et al., 2020, Theorem 3.3)

$$W_2^2(\rho_c, \rho_k) = \sup_{\substack{f \in \tilde{\Phi} \\ f^* \in L^1(\rho_k)}} \inf_{g \in \tilde{\Phi}} \mathcal{C}_{\rho_c,\rho_k} - \underbrace{\mathbb{E}_{\rho_c}[f(x)] - \mathbb{E}_{\rho_k}[\langle y, \nabla g(y)\rangle - f(\nabla g(y))]}_{\mathcal{V}_{\rho_c,\rho_k}(f,g)}. \tag{4.3}$$

We can further relax the constraint $g \in \tilde{\Phi}$ to $L^1(\rho_k)$, as a function $g \in L^1(\rho_k)$ minimizing (4.7) is convex and equal to f^* for any convex function f. In order to learn the resulting optimal transport, i.e., the solution of the minimization problem in (4.7), Makkuva et al. (2020) parameterize both dual variables f and g using input convex neural networks (§ 4.1.2) (Amos et al., 2017) and yields a transport plan with the graident of g. The resulting approximate Wasserstein distance is thus defined as

$$\hat{W}_2^2(\rho_c, \rho_k) = \sup_{\phi} \inf_{\theta} \mathcal{C}_{\rho_c,\rho_k} - \mathcal{V}_{\rho_c,\rho_k}(f_\phi, g_\theta), \tag{4.4}$$

where θ and ϕ are the parameters of each ICNN.

4.1.3.1 *Alternative Approaches*

We focus in this work on neural approaches that parameterize the optimal maps T as neural networks. An early approach is the work on Wasserstein GANs (Arjovsky et al., 2017), albeit the transport map is not explicitly estimated. Several recent results have exploited a more explicit connection

between OT and NNs, derived from the celebrated Brenier theorem (1987), which states that Monge maps are necessarily gradients of convex functions. Such convex functions can be represented using ICNNs (Amos et al., 2017), to parameterize either the Monge map (Korotin et al., 2021a; Yang and Uhler, 2019; Bunne et al., 2023b, 2022b) or a dual potential (Makkuva et al., 2020; Korotin et al., 2021a) as, respectively, the gradient of an ICNN or an ICNN itself. In this paper, we build on this line of work, but substantially generalize it, to learn a *parametric* family of context-aware transport maps, using a collection of labeled pairs of measures.

Several approaches have been proposed on inferring transport map $\mathcal{T}_\theta$ from paired source and target populations, including the primal (3.4) or dual optimal transport problem (3.6).

A possible approach to learn our model could consist in minimizing a primal OT problem. In that case, we can learn $\mathcal{T}_\theta$ via the gradient of the Brenier potential parameterized via a PICNN, i.e., $\mathcal{T}_\theta = \nabla\psi^*_\theta = \nabla_1\text{PICNN}_\theta$. The PICNN is then trained using the entropy-regularized Wasserstein distance (3.2) between the predictions $\hat{\nu} = \nabla\psi^*_\sharp\mu = \nabla_1\text{PICNN}_\theta(\cdot, c)_\sharp\mu$ given source samples μ and condition c and the observed target population ν as a loss function, i.e.,

$$\ell_{\text{POT}}(\mu, \nu, c; \theta) = W_\varepsilon(\nabla_1\text{PICNN}_\theta(\cdot, c)_\sharp\mu, \nu). \tag{4.5}$$

Throughout this work, we choose a different route and propose instead to learn $\mathcal{T}_\theta$ via the dual optimal transport problem. We consider the strategy proposed by Makkuva et al. (2020) and utilized by Bunne et al. (2023b) in the context of single-cell perturbation analyses. $\mathcal{T}_\theta$ is then parameterized via the pair of dual potentials f and g, which themselves are defined by a pair of PICNNs $g : \text{PICNN}_{\theta_g}(\cdot, c)$ and $f : \text{PICNN}_{\theta_f}(\cdot, c)$ such that $\hat{\nu} = \nabla g_\sharp\mu = \nabla_1\text{PICNN}_{\theta_g}(\cdot, c)\sharp\mu$ is approximately ν, as well as $\hat{\mu} = \nabla f_\sharp\nu = \nabla_1\text{PICNN}_{\theta_f}(\cdot, c)_\sharp\nu$ is approximately μ on a labeled observation $((\mu, \nu), c)$ with parameters $\theta = (\theta_g, \theta_f)$. In order to optimize the pair of PICNNs, which parameterize the two dual functions, Makkuva et al. (2020) derive an approximate formulation of (3.6). First, Villani (2003, Theorem 2.9) rephrases (3.6) over the pair of dual potentials (f, g) to

$$W(\mu, \nu) = \underbrace{\frac{1}{2}\mathbb{E}\left[\|x\|_2^2 + \|y\|_2^2\right]}_{C_{\mu,\nu}} - \inf_{f \text{ convex}} \mathbb{E}_\mu[f(X)] + \mathbb{E}_\nu\left[f^*(Y)\right], \tag{4.6}$$

where $f^*(y) = \sup_x \langle x, y\rangle - f(x)$ is f's convex conjugate. In a second step, Makkuva et al. (2020) derive a min-max formulation by approximating the convex conjugate in (4.6) via

$$W(\mu,\nu) = \sup_{\substack{f \text{ convex} \\ f^* \in L^1(\nu)}} \inf_{g \text{ convex}} \mathcal{C}_{\mu,\nu} - \underbrace{\mathbb{E}_\mu[f(x)] - \mathbb{E}_\nu[\langle y, \nabla g(y)\rangle - f(\nabla g(y))]}_{\mathcal{V}_{\mu,\nu}(f,g)}, \tag{4.7}$$

and by relaxing the constraints on g. Thus, the dual potentials f and g can be learned via an alternate min-max optimization problem with loss functions

$$\ell^f_{\text{DOT}}(\mu,\nu,c;\theta_f) = \mathbb{E}_{x\sim\mu}[\text{PICNN}_{\theta_g}(x,c)] - \mathbb{E}_{y\sim\nu}[\text{PICNN}_{\theta_f}(\nabla\text{PICNN}_{\theta_g}(y,c),c)], \text{ and} \tag{4.8}$$

$$\ell^g_{\text{DOT}}(\mu,\nu,c;\theta_g) = -\mathbb{E}_{y\sim\nu}[\langle y, \nabla\text{PICNN}_{\theta_g}(y,c)\rangle - \text{PICNN}_{\theta_f}(\nabla\text{PICNN}_{\theta_g}(y,c),c)]. \tag{4.9}$$

For more details, see Makkuva et al. (2020); Korotin et al. (2021b).

Thus, dependent on the strategy chosen, $\mathcal{T}_\theta$ is parameterized via a single or a pair of PICNN. Each network takes as input the source distribution μ —in which it is input convex— as well as an embedded context variable $\hat{c}$, returned by combinator $\mathcal{C}_\Phi$ and embedding module $\mathcal{E}_\phi$. Parameters of all three modules are jointly trained based on the derived optimal transport loss $\ell = \{\ell_{\text{POT}}, \ell_{\text{DOT}}\}$, which measures how close predicted target cells $\hat{\nu}$ are from the observed target population ν, given source population μ and context c as inputs.

4.2 CELLOT: NEURAL OPTIMAL TRANSPORT FOR LEARNING CELLULAR PERTURBATION RESPONSES

Characterizing and modeling perturbation responses at the single-cell level from non-time-resolved data remains one of biology's grand challenges. It finds applications in predicting cellular reactions to environmental stress or a patient's response to drug treatments. Accurate inference of perturbation responses at the single cell level allows us, for instance, to understand how and why individual tumor cells evade cancer therapies (Frangieh et al., 2021). More generally, it deepens the mechanistic understanding of the molecular machinery that determines the respective responses to perturbations. Single-cell responses to genetic or chemical perturbations are highly heterogeneous (Liberali et al., 2014) due to multiple factors, including pre-existing variability in the abundance and subcellular organization of

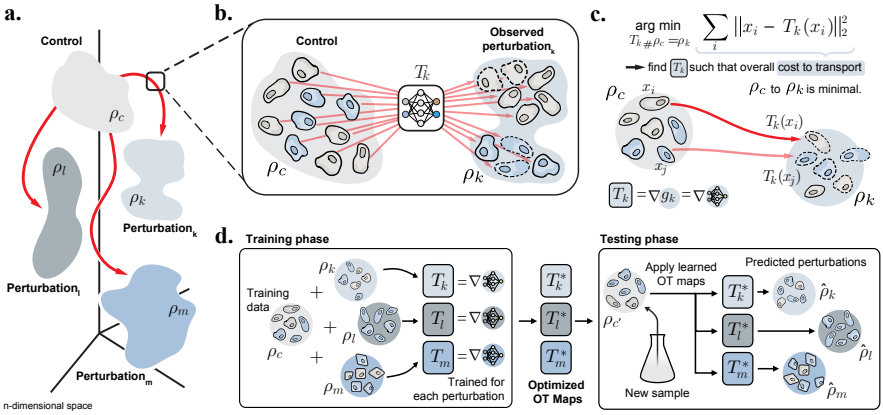


Figure 4.1: Overview of the CellOT Model. a. Distributions of single cells were measured in either an untreated control state (ρ_c) or in one of several perturbed states ($\rho_k, \rho_l, \rho_m, \ldots$). These distributions lie in a high-dimensional space of profiled features. **b.** For a perturbation k, we aim to model it with a function T_k that maps untreated cells in ρ_c to their treated counterparts in ρ_k. **c.** Lacking paired measurements, we assume that the perturbation transforms ρ_c into ρ_k under a principle of minimal effort. In particular, we learn T_k using optimal transport theory to directly estimate this distributional mapping as the gradient of the optimal transport dual potential ∇g_θ. **d.** OT maps are learned for all perturbations independently. Because these maps are fully parameterized, CellOT can be trained, for example, on a set of initially provided samples to then make predictions on untreated cells originating from new, previously unseen samples.

mRNA and proteins (Battich et al., 2013, 2015; Gut et al., 2018; Shaffer et al., 2017), cellular states (Kramer et al., 2022), and the cellular microenvironment (Snijder et al., 2009). To effectively predict the drug response of each cell in a population, whether derived from tissue culture or as primary cells from a patient biopsy, it is thus crucial to incorporate this heterogeneous multivariate subpopulation structure into the analysis.

A fundamental difficulty in learning perturbation responses is that cells are usually fixed and stained or chemically destroyed to obtain these measurements. Hence, it is only possible to measure the same cells before or after a perturbation is applied. Therefore, while we do not have access to a set of *paired* control/perturbed single-cell observations, we do have access to separate *sets* of single-cell observations from control and perturbed cells, respectively. To subsequently match single cells between conditions and, at the same time, account for cellular heterogeneity is a highly complex pairing problem.

Here, we seek to learn a perturbation model that robustly describes the cellular dynamics upon intervention while still accounting for underlying variability across samples. Learning the responses on an existing patient cohort enables inference of treatment responses for new, i.e., previously unseen patients, assuming that we captured the heterogeneous drug reactions of patients during training. It is crucial, however, to not simply model average perturbation responses of a patient cohort, but to capture the specificities of a single patient through personalized treatment effect predictions.

Previous methods to approximate single-cell perturbation responses fall short of solving this highly complex *pairing* problem while, at the same time, accounting for cellular heterogeneity and the strong subpopulation structure of cell samples (Wu et al., 2021; González-Silva et al., 2020; Li et al., 2022). Current state-of-the-art methods (Lopez et al., 2018; Lotfollahi et al., 2019; Yang et al., 2020) predict perturbation responses via *linear shifts* in a learned latent space. While this can capture nonlinear cell-type-specific responses, the use of linear interpolations reduces the alignment problem to the possibly more challenging task of learning representations that are invariant to the corresponding perturbation.

In this work, we introduce CellOT, a novel approach that predicts perturbation responses of single cells by *directly* learning and uncovering maps between control and perturbed cell states, thus explicitly accounting for heterogeneous subpopulation structures in multiplexed molecular readouts. Assuming perturbations incrementally alter molecular profiles of cells,

such as gene expression or signaling activities, we learn these changes and alignments using optimal transport theory (Villani, 2009). Optimal transport provides natural geometric and mathematical tools to manipulate probability distributions. It has found recent successes modeling cellular development processes (Lavenant et al., 2021; Schiebinger et al., 2019), albeit in a *non-parameterized* setting. Thus, current OT-based approaches are unable to make predictions on unseen cells, such as those from unseen samples, e.g., new patients.

Based on recent developments in neural optimal transport (Makkuva et al., 2020), CellOT learns an optimal transport map for each perturbation in a fully parameterized and highly scalable manner. Instead of directly learning a transport map (Korotin et al., 2021a; Yang and Uhler, 2019; Prasad et al., 2020), CellOT parameterizes a pair of dual potentials with convex neural networks (Amos et al., 2017). This choice induces an important theory-motivated inductive bias essential to model stability (Makkuva et al., 2020).

We demonstrate CellOT's effectiveness by (i) learning single-cell marker responses to different cancer drugs in melanoma cell lines, (ii) predicting single-cell transcriptome responses in biopsies of patients with systemic lupus erythematosus as well as Panobinostat treatment outcomes of glioblastoma patients, (iii) inferring LPS responses across different animal species, and (iv) modeling the transcriptome evolution of cell fates in hematopoiesis. Moreover, we benchmark CellOT against current state-of-the-art methods on multiple tasks (Lopez et al., 2018; Lotfollahi et al., 2019).

4.2.1 *Predicting Perturbation Responses via Neural Monge Maps*

Small molecule drugs can have profound effects on the cellular phenotype by, for instance, altering signaling cascades. Most of these effects depend on the context in which the perturbation occurs. Given the heterogeneity among single cells in cell populations and tissues, predicting cellular responses requires understanding the rules by which context shapes genome activity and its response to drugs. High-dimensional single-cell data measured via single-cell genomics or multiplexed imaging technologies can provide this contextual information but only return unpaired or unaligned observations of cell populations. Here, CellOT allows us to utilize such unpaired data and enables learning cell state transitions upon perturbation.

Recent high-throughput methods provide great insights on how cell populations respond to various perturbations on the level of individual

cells. The provided data, however, is non-time-resolved and unaligned. Hence, snapshots taken of biological samples before and after perturbations do not provide information on single-cell trajectories. Perturbations might include the application of drugs affecting molecular functions in cells, or changes in the cellular environment causing shifts in biological signaling, thus impacting cells and their states in various ways. In the following, we describe our approach, which uncovers single-cell perturbation responses by predicting couplings between control and perturbed cell states. Hereby, let $\mathcal{X}$ denote the biological data space spanned by the measured cell features. We then treat a cell's response to perturbation k as an evolution in a high-dimensional space of cell states $\mathcal{X} = \mathbb{R}^d$.

In formal terms, we denote the unperturbed control population by ρ_c consisting of n cells x_i for $i = 1, \dots, n$. Upon perturbation k, the multivariate state of each cell x_i of the unperturbed population changes, which we observe as the perturbed population ρ_k (Fig. 4.1a). To understand the mode of action and effect of perturbations, we seek to learn the transition and alignment between populations ρ_c and ρ_k via parameterizing a map T_k (see Fig. 4.1a-b), which explains the transition of each cell from the unperturbed cell population ρ_c into their perturbed state ρ_k upon treatment k. Despite originating from different observations, map T_k determines for each cell x_i the most likely corresponding cell $T_k(x_i)$ in the perturbed population (Fig. 4.1c). Finding this map then not only allows us to model single-cell trajectories upon perturbation but also to predict the perturbed state of previously unseen control cells. As a result, we can forecast the outcome of a perturbation k by applying the learned map T_k to a new unperturbed population ρ_c' (Fig. 4.1d).

The optimal map T_k aligning the control and perturbed population, which we seek to find, should best describe the incremental changes in the multivariate profile of each cell after applying a perturbation k. Using optimal transportation theory (Villani, 2003; Santambrogio, 2015) to recover these maps and unveil single-cell reprogramming trajectories has been proposed as a strong modeling hypothesis in the domain of single-cell biology (Schiebinger et al., 2019; Cang and Nie, 2020; Demetci et al., 2022; Huizing et al., 2022; Lavenant et al., 2021; Zhang et al., 2021). Optimal transport problems return the alignment between distributions ρ_c and ρ_k corresponding to the minimal overall cost between aligned molecular profiles, thus determining the most likely state of each cell upon perturbation (Fig. 4.1c). T_k is learned such that its image corresponds to ρ_k and mass is moved from ρ_c into ρ_k according to a principle of minimal effort. As directly parameterizing

the optimal transport map T_k (Korotin et al., 2021a; Yang and Uhler, 2019; Prasad et al., 2020) is unstable (Makkuva et al., 2020, Table 1), we parameterize the convex potentials of the dual optimal transport problem f and g by convex neural networks (Amos et al., 2017) and recover the optimal map T_k using the gradient of a convex function g_k, i.e., ∇g_k (Makkuva et al., 2020).

To put CellOT's performance in perspective, we benchmark it against current state-of-the-art methods based on autoencoders (Lotfollahi et al., 2019; Lopez et al., 2018), which attempt to add perturbation effects through the manipulation of a learned latent representation. To further test the hypothesis of the optimal transport modeling prior, we compare the learned OT map ∇g_k for each perturbation k with naive non-OT-based alignments.

4.2.2 ...

Given a dataset of n observations $\{x_1^c, \dots, x_n^c\}, x_i^c \in \mathcal{X}$ drawn from $\rho_c \in \mathcal{P}(\mathcal{X})$, the distribution of cells before applying a perturbation, we aim to learn the distribution of cells $\rho_k \in \mathcal{P}(\mathcal{X})$ upon some perturbation k, given a set of separate samples $\{x_1^k, \dots, x_m^k\}, x_i^k \in \mathcal{X}$.

Perturbation responses of cells are dynamic: After applying perturbation k, cell states evolve over time and thus can be modeled as a stochastic process in the cell data space. Despite this time-resolved nature of single-cell responses, we only have access to the distributions of cell states before, ρ_c, and after injecting perturbation k, ρ_k. We thus aim to understand the underlying stochastic process without access to time-resolved perturbation responses by uncovering the coupling γ between ρ_c and ρ_k. Given prior biological knowledge, we can assume that perturbations do not drastically or totally alter underlying cellular processes. We thus posit that the evolution of probability distributions of single cells upon perturbation can be modeled via the mathematical theory of optimal transport. The coupling γ then corresponds to an optimal transport plan (3.5) between ρ_c and ρ_k.

Following Makkuva et al. (2020), we learn the optimal map T (3.4) between ρ_c and ρ_k. Thus, instead of computing a coupling γ individually for each pair of cell samples using existing solvers (Cuturi, 2013), we learn a parameterized optimal transport map using neural networks. The parameterized OT map then serves as a robust predictor for cellular distribution shifts upon perturbations on unseen samples $\{x_i^c\}_{i=1}^{n'} \sim \rho_c$, i.e., of another patient.

… The framework described above allows us to recover maps between control $\{x_1^c, \dots, x_n^c\}$ and perturbed cells $\{x_1^k, \dots, x_m^k\}$, giving insights into cellular response trajectories upon application of a perturbation k. Given a set of perturbations K, and sample access to the control distribution ρ_c as well as distributions ρ_k for each perturbation $k \in K$, CellOT learns the optimal pair of dual potentials $(f^*_{\phi_k}, g_{\theta^*_k})$ by solving (4.4). Given parametrizations of the convex potentials for each k, CellOT then predicts the transformation of a control cell x_i^c upon perturbation k via $\hat{x}_i^k = \nabla g_{\theta_k^\star}(x_i^c)$, i.e., samples following the predicted perturbed distribution $\hat{\rho_k} = (\nabla g_{\theta_k^*})_{\#}\rho_c$. CellOT thus provides a general approach to predict state trajectories on a single-cell level, as well as understand how heterogeneous subpopulation structures evolve under the impact of external factors.

4.3 EMPIRICAL EVALUATION

4.3.1 *Predicting Treatment Outcomes of Cancer Drugs*

We apply CellOT to predict the responses of cell populations to cancer treatments using a proteomic dataset consisting of two melanoma cell lines (M130219 and M130429) (Raaijmakers et al., 2015), profiled by 4i (Gut et al., 2018), and a scRNAseq dataset (Srivatsan et al., 2020), which contain 34 and 9 different treatments, respectively. We benchmarked CellOT against two autoencoder-based tools, scGen (Lotfollahi et al., 2019) and cAE (Lopez et al., 2018), as well as PopAlign (Chen et al., 2020), a method based on aligning subpopulations of the control and treated space approximated through a mixture of Gaussian densities. Due to the high dimensional nature of scRNA-seq data, we apply CellOT on latent representations learned by an autoencoder. The marginal distributions for observed and predicted cell populations for two 4i treatments and two scRNAseq treatments are shown in Fig. 4.2a, d. Two features are selected for each perturbation. While the autoencoder baselines tend to capture the mean of the treated cell population, they are less successful in matching all heterogeneous states of the perturbed population, i.e., higher moments of the perturbed population. Thus, these models tend to learn over-simplified perturbation effects and are insufficient when aiming to understand heterogeneous rather than average cellular behaviors. CellOT, on the other hand, is able to capture these higher moments, yielding accurate and nuanced predictions.

This can be further quantified through distributional metrics such as the maximum-mean-discrepancy (MMD) (Gretton et al., 2012). Low values

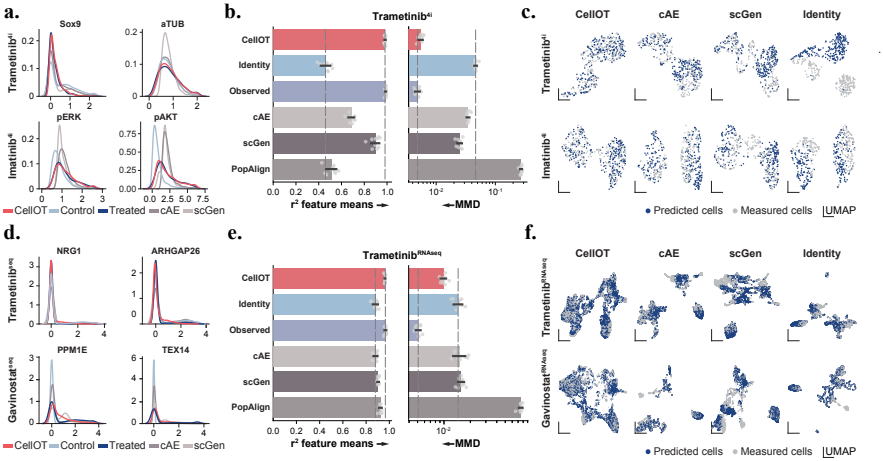


Figure 4.2: CellOT outperforms current state-of-the-art methods on different data modalities. Marginal distribution of marker gene expression (x-axis) of cells profiled by **a.** 4i and **d.** scRNA. Observed control and treated states are shown in light and dark blue. CellOT predictions are shown in red and baseline predictions (scGen, cAE, PopAlign) are shown in gray. We compare models based on the distributional distance MMD as well as average correlation coefficient r^2 between observed perturbed and predicted perturbed cells, for **b.** 4i and **e.** scRNA data. Error bars refer to the standard deviation over 10 bootstraps of the test set and the dashed lines correspond to the median of the identity and observed performances. Joint UMAPs of observed treated cells and cells predicted by each model for **c.** 4i and **f.** scRNA data. Projections are computed on a joint set of cells, down-sampled such that the number of observed perturbed (gray) and predicted perturbed cells (blue) are equal. An identity coupling compares treated cells to untreated cells. The analysis is conducted for drugs Trametinib, Imatinib, and Gavinostat. 4i data was generated using cell lines M130219 and M130429.

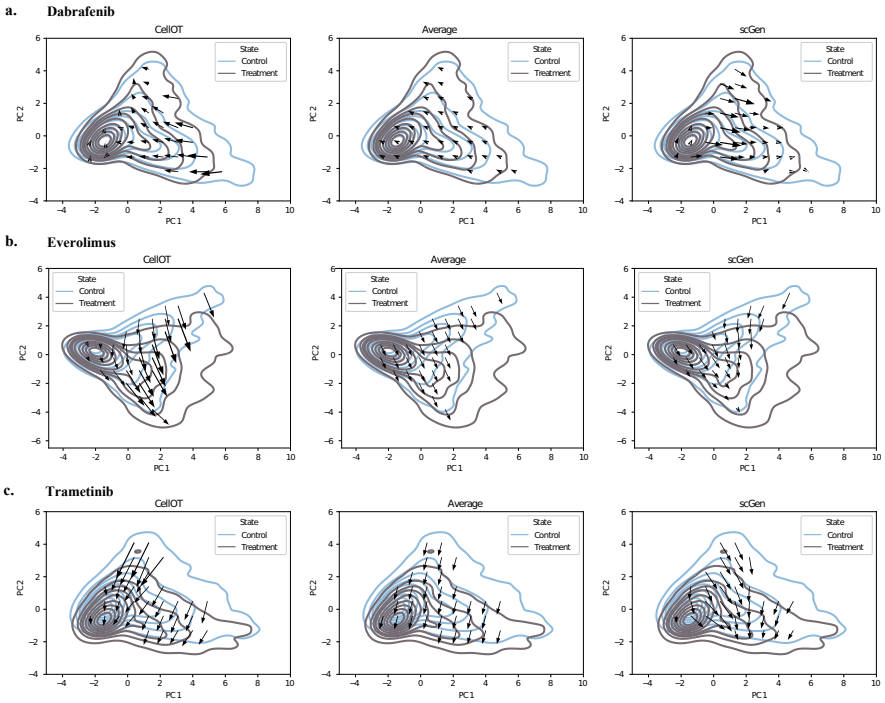


Figure 4.3: Visualization of the learned vector field describing the perturbation response on the single-cell level for **a.** Dabrafenib, **b.** Everolimus, and **c.** Trametinib of the 4i dataset for CellOT, the average effect, and scGen on the first two principal components. Cellular responses are computed as the predicted treated state minus the observed control state for each individual cell. Arrow tails are placed in a grid within PC space and arrow heads correspond to the average response of cells within each neighborhood, projected into PC space.

of MMD imply that all moments of two distributions are matched, and thus the entire distribution of perturbed cells is captured in fine detail, beyond the population average. The MMDs between the predicted and observed populations for the selected perturbations are shown in Fig. 4.2b, e. For scRNA-seq data, MMD evaluations are computed using the top 50 marker genes. In addition to the autoencoder baselines, we include the trivial *identity* baseline that predicts treatment effects simply by returning the untreated states, as well as a theoretical lower bound, *observed*, comprising a different set of observed perturbed cells, thus only varying from the true predictions up to experimental noise. We find that CellOT can approach the lower bound (*observed* setting), while the baseline methods often do not improve much over the *identity* setting. Fig. 4.3 visualizes the learned maps, further demonstrating CellOT's ability to model fine-grained responses.

Finally, we compute UMAP projections (McInnes et al., 2018) on a joint set of predicted and observed perturbed cells utilizing the full feature space, shown in Fig. 4.2c, f. We observe that the perturbed cell states inferred by CellOT are well integrated with the observed perturbed cells. Again, both baselines do not recover the perturbed distribution in its entirety and thus the perturbed state of different subpopulations is not captured consistently. CellOT outperforms the baselines in both metrics across all treatments, typically by one order of magnitude. We attribute the strong performance of CellOT to its ability to learn a transport function that considers explicitly the data geometries of cell populations through the theory of optimal transport.

4.3.2 *Capturing Cell-to-Cell Variability in Drug Responses*

Capturing distinct perturbation responses of different cell types within the same sample remains a challenging computational task. To reduce the task's complexity, prediction algorithms can be guided by predefined cell type labels both in the perturbed and unperturbed states (Chen et al., 2020) or set to approximate the mean drug response (Lotfollahi et al., 2019). These simplifications come at a cost: the reliance on a priori knowledge about present and relevant cell types, the assumption that cell types are characterized by the same features before and after a perturbation and that the drug response is uniform within a cell type. In the worst case, these limitations risk masking true and important drug response heterogeneity and thus hamper the discovery of novel cell type or cell state-specific perturbation responses.

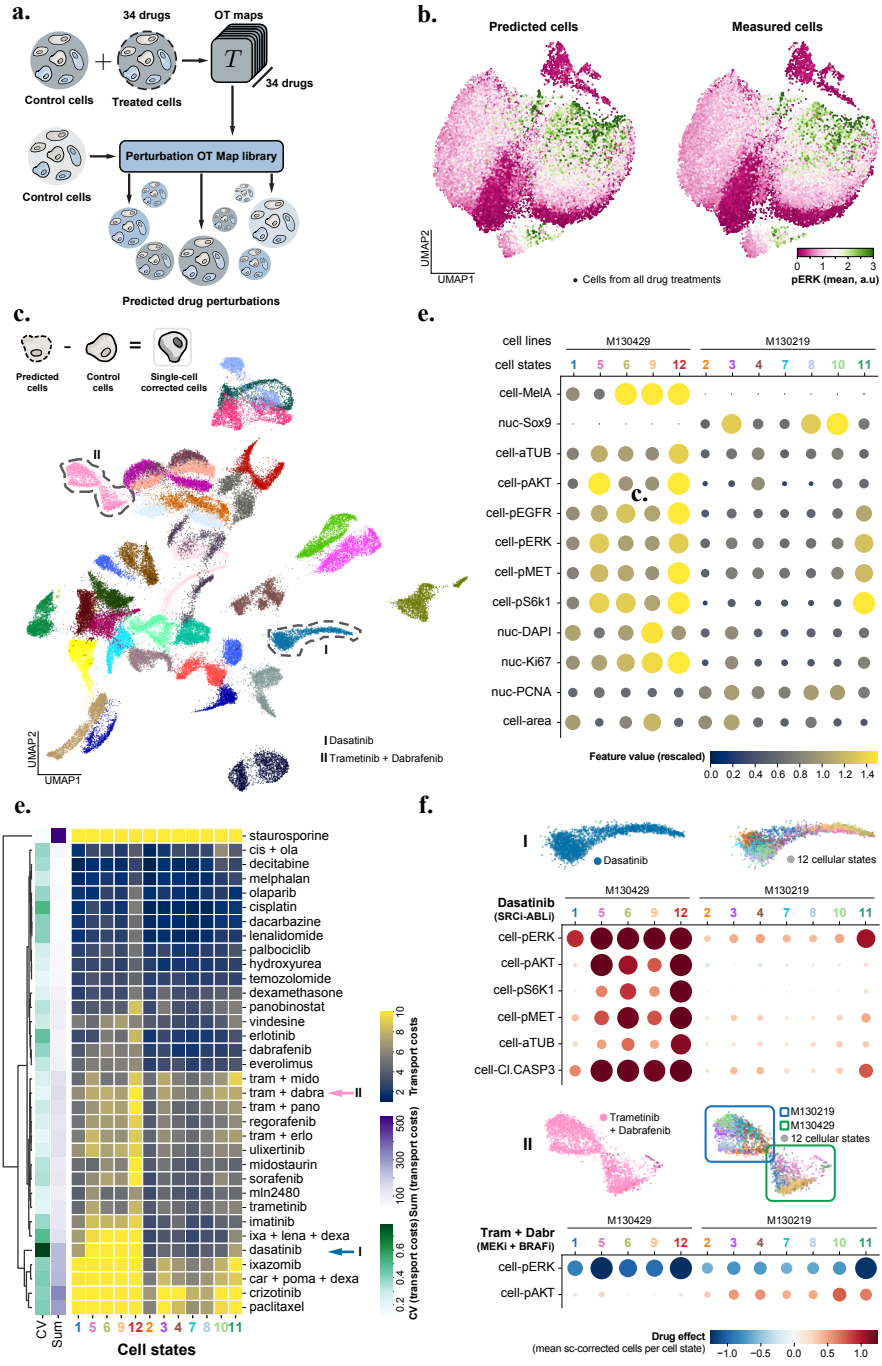

a.
34 drugs
OT maps
Control cells
Treated cells
34 drugs
Perturbation OT Map library
Control cells
Predicted drug perturbations
b.
Predicted cells
Measured cells
UMAP2
UMAP1
Cells from all drug treatments
pERK (mean, a.u)
c.
Predicted cells
Control cells
Single-cell corrected cells
II
I
UMAP2
UMAP1
I Dasatinib
II Trametinib + Dabrafenib
e.
cell lines
M130429
M130219
cell states
cell-MelA
nuc-Sox9
cell-aTUB
cell-pAKT
cell-pEGFR
cell-pERK
cell-pMET
cell-pS6k1
nuc-DAPI
nuc-Ki67
nuc-PCNA
cell-area
Feature value (rescaled)
e.
staurosporine
cis + ola
decitabine
melphalan
olaparib
cisplatin
dacarbazine
lenalidomide
palbociclib
hydroxyurea
temozolomide
dexamethasone
panobinostat
vindesine
erlotinib
dabrafenib
everolimus
tram + mido
tram + dabra
tram + pano
regorafenib
tram + erlo
ulixertinib
midostaurin
sorafenib
mln2480
trametinib
imatinib
ixa + lena + dexa
dasatinib
ixazomib
car + poma + dexa
crizotinib
paclitaxel
II
I
CV
Sum
Cell states
Transport costs
Sum (transport costs)
CV (transport costs)
f.
I
Dasatinib
12 cellular states
Dasatinib
(SRCI-ABLI)
M130429
M130219
cell-pERK
cell-pAKT
cell-pS6K1
cell-pMET
cell-aTUB
cell-Cl.CASP3
II
Trametinib
+ Dabrafenib
M130219
M130429
12 cellular states
Tram + Dabr
(MEKi + BRAFi)
cell-pERK
cell-pAKT
Drug effect
(mean sc-corrected cells per cell state)

Figure 4.4: CellOT facilitates the multiplexed single-cell characterization of cancer drugs. **a.** CellOT training and prediction setup. 34 CellOT models were trained, one for each drug perturbation. Subsequently, each model was used to predict perturbed cells from a common set of unseen control cells. **b.** UMAP projection constructed with equal numbers of predicted and measured cells from 34 perturbations. Dots correspond to a cells, color-coded for measured or predicted pERK intensity, **c.** UMAP projection of single-cell perturbation effects using predicted cells. Dots correspond to cells, color-coded for drug treatment (see Fig. A.1 for full legend for single-cell perturbation effect calculation). **d.** Cell states identified in control cells. Each column represents a cell state. Horizontal axis, cell states sorted based on their association to the cell lines M130219 and M130429. Vertical axis, cellular features (see Fig. A.1 for the full feature set). Size and hue of the circles are scaled on the feature values. **e.** Clustergram of transport cost (TC) of drug treatments for each cell state (main heatmap, blue-yellow color scheme), the sum of TCs (Sum) of all states per drug (first column left of the heatmap, purple), the coefficient of variation (CV) of TCs per drug (second column left of the heatmap, green) and the dendrogram based on the hierarchical clustering the drug's cell state TCs. Cell states are sorted as in **d**. **f.** Cell state-specific responses to drug treatments. Top panel (I) Dasatinib. Bottom panel (II) Trametinib + Dabrafenib. Panel organization: top-left, condition-focused enlargement of UMAP projection from **c**. Top-right, same as top-left but color-coded for cell state assignment. Bottom, columns represent a cell state, rows highlighted features. 'cell-' stands for mean cell intensity. Circles are scaled based on drug effect size, the stronger the effect the larger the circles. Negative values are encoded in hues of blue, positive values in red hues of the respective circles.

CellOT is free of these limitations and enables scientists to query the predicted single-cell responses at the granularity best suited to answer their biological questions. As a proof of concept, we co-cultured the aforementioned patient-derived melanoma cell lines at equal ratios and performed a boutique drug screen, during which we exposed cells 8h to a panel of 34 drugs and measured the single-cell drug responses with the 4i technology. Using CellOT, we predict the perturbed cell states of a shared set of control (DMSO-treated) cells (Fig. 4.4a) for each drug. Previous work (Kramer et al., 2022) shows that phosphorylation levels of signaling kinases upon drug treatments are tightly linked to the cellular state. To assess whether this relationship was retained in predicted compared to observed perturbed cells, we analyzed the phosphorylation levels of extracellular signal-regulated kinases (pERK) using the transport maps learned by CellOT on each drug. Using 750 predicted and 750 observed perturbed cells, we computed UMAP projections joint-wise from all features except pERK. Fig. 4.4b shows the predicted and observed population individually annotated with the respective pERK levels of each cell. We find the spatial organization of the two projections to look almost identical and that pERK levels had a highly comparable distribution across the cells of either class and all drug treatments (further analysis in Fig. A.1a, b).

4.3.3 *Disentangles Subpopulation-Specific Drug Effects*

CellOT allows us to isolate the mode of action of each drug by computing the difference between predicted perturbed cells and untreated control cells. A UMAP embedding of all cells color-coded by the treatment distinctly separates different treatments (Figs. 4.4c and A.1e), all of which CellOT is able to faithfully learn. Such distinct treatment embeddings are not present when accounting only for an average perturbation effect (Fig. A.1d), indicating the importance of capturing the cellular heterogeneity of drug responses.

Using Leiden clustering on the full feature set, we grouped unperturbed control cells in 12 cellular states (Fig. 4.4d, Fig. A.1g). Cellular states 1, 5, 6, 9, and 12 show high levels of MelA and no SOX9 and thus correspond to the melanocytic cell line M130429, whereas the $SOX9^{+}$ and $MelA^{-}$ states 2, 3, 4, 7, 8, 10, and 11 represent the mesenchymal cell line M130219. Overall, we find that M130429 cells have higher phosphorylation levels of the measured signaling kinases compared to M130219; a stereotypical spatial organization of cellular states is retained for the majority of the drugs, and cell states belonging to the same cell line cluster together (Fig. A.1f).

Computing the difference between the control and treated state of each drug, i.e., the optimal transport cost, allows us to further characterize a drug's severity. Apoptosis inducers (e.g., Staurosporine), proteasome inhibitors (e.g., Ixazomig and Carfilzomib or the combination treatment Carfilzomib + Pomalidomide + Dexamethasone), microtubule-stabilizing agents (e.g., Paclitaxel), c-Met inhibitors (e.g., Crizotinib), and ATP competitors for multiple tyrosine kinases such as c-KIT, and Bcr-Abl (i.e., Dasatinib) show high transport costs and thus substantial feature changes in all cellular states (Fig. 4.4e). Other drugs demonstrate less severe effects in the observed 8h incubation period. We find all perturbations to increase levels of cleaved Caspase 3, an apoptosis marker, in various cellular states and in both cell lines (Fig. A.1k), with the exception of Dasatinib, which specifically induced cell death in cellular states 5, 6, 9, and 19 associated to M130429 (Fig. 4.4f). Previous work by Smith et al. (2016) reports that M130429 cells reduce metabolic activity upon treatment with inhibitors of MEK (MEKi) and RAF (RAFi), while M130219 cells are resistant to these inhibitors. When comparing the responses of the two cell lines to Trametinib (MEKi) and MLN2480 (panRAFi) in the MEK and PI3K pathway using pERK and pAKT as the respective readouts, we find that MEKi-sensitive M130429 cells down-regulate pAKT and pERK, whereas the MEKi-resistant M130219 cells

only down-regulate pERK. Consistently, we also find that treatment with MLN2480 results in a similar differential drug response (Fig. A.1i). This suggests that *decoupling* of the MEK and PI3K pathways may confer resistance to MEK and Raf inhibitors and constitute an adaptation to the escape of cancer therapy (Kun et al., 2021). We find further supporting evidence of pathway crosstalk alteration when we analyze pAKT and pERK levels upon treatment with a cocktail of Trametinib (MEKi) and Dabrafenib (BRAFi). In response to two drugs impinging on the MEK pathway, we observe pERK to be reduced in both cell lines but increased pAKT levels in the MEKi-resistant cell line M130219 (which resistance was acquired during pre-exposing a patient to MEKi) (Fig. 4.4f). This finding points towards a compensatory feedback mechanism acquired by M130219 during MEKi treatment by which inhibition of the MEK pathway (quantified as a reduction of pERK) would stimulate signaling through the PI3K pathway, possibly through activation of an upstream receptor kinase (Caunt et al., 2015). Our results on two co-cultured primary melanoma cell lines treated with various anti-cancer drugs show that CellOT can accurately capture phenotypic heterogeneity in unperturbed cell populations and predict diverse drug responses by incorporating the underlying cell-to-cell variability without predefined cell line labels.

4.3.4 *Inferring Cellular Responses in Unseen Patients*

The maps between molecular states before and after treatments learned by CellOT contribute to a better understanding of the differences between cells that respond to certain drugs and cells that do not respond. This is crucial for inferring an incoming patient's response to drugs and settings with high cell-to-cell variability. To make predictions on unseen patients, however, we need to demonstrate that the learned maps T model perturbation responses across different patients coherently and robustly, while still predicting personalized treatment outcomes for each patient instead of mere population averages.

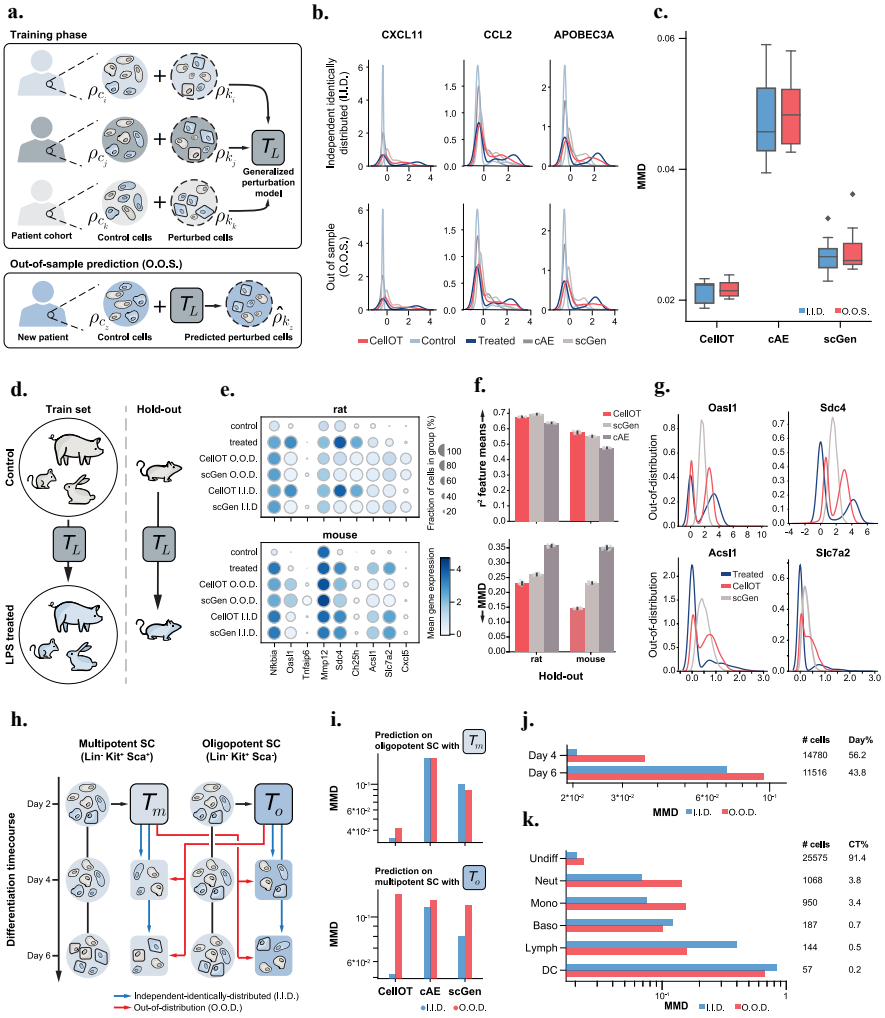

a.
Training phase
Generalized perturbation model
Patient cohort
Control cells
Perturbed cells
Out-of-sample prediction (O.O.S.)
New patient
Control cells
Predicted perturbed cells
b.
CXCL11
CCL2
APOBEC3A
Independent identically distributed (I.I.D.)
Out of sample (O.O.S.)
CellOT
Control
Treated
cAE
scGen
c.
MMD
I.I.D.
O.O.S.
CellOT
cAE
scGen
d.
Train set
Hold-out
Control
LPS treated
e.
rat
mouse
control
treated
CellOT O.O.D.
scGen O.O.D.
CellOT I.I.D.
scGen I.I.D.
Fraction of cells in group (%)
Mean gene expression
Nfkbia
Oasl1
Tnfaip6
Mmp12
Sdc4
Ch25h
Acsl1
Slc7a2
Cxcl5
f.
CellOT
scGen
cAE
r² feature means
MMD
rat
mouse
Hold-out
g.
Oasl1
Sdc4
Acsl1
Slc7a2
Out-of-distribution
Treated
CellOT
scGen
h.
Multipotent SC (Lin⁻ Kit⁺ Sca⁺)
Oligopotent SC (Lin⁻ Kit⁺ Sca⁻)
Differentiation timecourse
Day 2
Day 4
Day 6
Independent-identically-distributed (I.I.D.)
Out-of-distribution (O.O.D.)
i.
Prediction on oligopotent SC with
Prediction on multipotent SC with
MMD
CellOT
cAE
scGen
I.I.D.
O.O.D.
j.
Day 4
Day 6
cells
Day%
14780
56.2
11516
43.8
MMD
I.I.D.
O.O.D.
k.
cells
CT%
Undiff
25575
91.4
Neut
1068
3.8
Mono
950
3.4
Baso
187
0.7
Lymph
144
0.5
DC
57
0.2
MMD
I.I.D.
O.O.D.

Figure 4.5: CellOT generalizes to unseen patients and cell subpopulations. Out-of-sample (o.o.s., **a-c**), and out-of-distribution (o.o.d., **d-k**) setting. **a.** Cells from eight lupus patients are measured in an untreated and IFN-β treated state. **b.** Marginals of predicted cells from the holdout sample in the i.i.d. (top) and o.o.s. (bottom) setting. Predictions for both models are made on the same test set (not used for training the two models). **c.** MMD scores between the predicted distribution and the observed treated distribution across all holdout samples in the i.i.d. and o.o.s. settings. Box plots indicate the median and quartiles. **d.** As an o.o.d. task, we train CellOT and baselines to predict the response to LPS across different species, and test on rat (or mouse) as a holdout species. **e.** Mean gene expression for i.i.d. and o.o.d. predictions for CellOT and scGen for selected marker genes. **f.** Comparison of o.o.d. performance for r^2 correlation feature means and MMD of CellOT and baselines. Data is depicted as the mean +/- standard deviation across n=10 bootstraps of the test set. **g.** Marginals of the o.o.d. predictions for marker genes showing bimodal expression profiles when using rat as a holdout. **h.** We apply CellOT to predict how cells from day 2 develop into the combined set of day 4 and 6, when trained on only multipotent cells (T_m) or oligopotent cells (T_o). We then apply T_m to predict the o.o.d. oligopotent cells and T_o to predict the o.o.d. multipotent cells. **i.** MMD scores between the predicted and (observed) developed distributions for all models in both o.o.d. and i.i.d. prediction tasks (jointly for day 4 and 6). Performance of CellOT, when predicting **j.** day 4 states and day 6 states **k.** for different cell types in each setting using T_m.

To test the generalization capacity of CellOT in such an out-of-sample scenario, we use a peripheral blood mononuclear cells (PBMC) droplet scRNA-seq dataset. Kang et al. (2018) characterize the cell type specificity and inter-individual variability of the response of eight lupus patients to interferon beta (IFN-β), a potent cytokine that induces genome-scale changes in immune cell transcriptional profiles. In the following, we compare the performance of CellOT and other baselines in an independent-and-identically-distributed (i.i.d.) setting, where models see cells from all patients, as well as in the out-of-sample (o.o.s.) setting, where models do not see cells from a specific holdout patient (see Fig. 4.5a).

As in the previous analysis, we evaluate how accurately CellOT captures the change in the overall expression of different marker genes from control to IFN-β-treated cells and thus how well the predicted gene expression marginals are aligned with the treated population (Fig. 4.5b). Here, we consider the genes *CXCL11*, *CCL2*, and *APOBEC3A*, since they are connected with autoimmune diseases, including systemic lupus erythematosus (Hedrich and Tsokos, 2011; Perez-Bercoff et al., 2021) and thus potential therapeutic targets in the management of patients with lupus and, likely, other interferonopathies (Mathian et al., 2015; Rani et al., 1996; Hedrich and Tsokos, 2011; Mathian et al., 2015; Perez-Bercoff et al., 2021; Flier et al., 2001). These selected genes show a large change in expression from the control to the perturbed population, partially exhibiting a bimodal gene

expression profile upon perturbation. In contrast to CellOT, the baselines do not accurately predict these large transcriptomic shifts of these genes.
All models, including CellOT, show little performance drop when modeling the treatment outcome on a new patient using the generalized perturbation model T_L trained on the patient cohort and using the control cells ρ_{c_z} of the unseen patient as input. This becomes evident when comparing the predicted population $\hat{\rho}_{k_z}$ with observations ρ_{k_z} using the MMD metric. Fig. 4.5c displays summary results in which each individual patient was considered for the holdout set. CellOT outperforms previous baselines both in the i.i.d. and in the o.o.s. setting, while further showing a smaller performance drop when generalizing to the unseen patient. These results suggest that the learned optimal transport maps correctly model the shift in the structures of the cellular subpopulation present in all patients, thus robustly performing out-of-sample. We repeat the same evaluation for a glioblastoma cohort consisting of seven patients (Zhao et al., 2021). However, generalization within this setting proved to be difficult for CellOT and all baselines, due to the small size of the cohort and high degree of variance within the responses of each individual. For a complete analysis, see Fig. A.3.

4.3.5 *Reconstructing Innate Immune Responses across Different Species*

The innate immune response is a cell-intrinsic defense program showing high levels of heterogeneity among responding cells —and thus an ideal task for evaluating CellOT's capabilities. We rely our analysis on the dataset collected by Hagai et al. (2018), which studies the evolution of innate immunity programs of mononuclear phagocytes within different species, including pigs, rabbits, mice, and rats. For this, these primary bone marrow-derived cells are stimulated using lipopolysaccharide (LPS). In the following, we test how well CellOT and the baselines reconstruct innate immune responses within species that are not encountered during training. We refer to the generalization task as out-of-distribution (o.o.d.), since unlike the o.o.s. setting, we expect different species to have very distinct responses (see Fig. 4.5d). The holdout set thereby consists of cells derived from either rat or mouse. See Fig. A.2a,b for an analysis of cross-species similarity and the reasoning behind selecting the holdout set.
Indeed, CellOT accurately reconstructs the innate immune response in both mouse and rat in the i.i.d. and o.o.d. setting. This not only becomes evident through capturing more precisely the mean expression level of marker genes that show high differential expression levels upon addition of LPS,

e.g., *Nfkb1* (NF-κB), *Oasl1* (Oasl1), *Mmp12*, and *Cxcl5* (see Figs. 4.5e and A.2c-d), but also through the average correlation coefficient r^2 computed between o.o.d. predictions and holdout observations across all genes (see Fig. 4.5f). In particular, CellOT outperforms the baselines when analyzing how well each method captures the heterogeneity of innate immune responses in different species, as demonstrated by low levels of MMD (see Fig. 4.5f). Most impressively, our method shows a strong alignment or gene expression marginals of aforementioned marker genes that show complicated bimodal expression profiles upon perturbation (see Fig. 4.5g).

4.3.6 *Generalizing Developmental Fate Decisions from Multipotent to Oligopotent Cell Populations*

During developmental processes, stem and progenitor cells progress through a hierarchy of fate decisions, marked by a continuous differentiation of cells that refine their identity until reaching a functional end state. By tracking an initial cell population along the differentiation process, CellOT allows us to recover individual molecular cell fate decisions and developmental trajectories.

Weinreb et al. (2020) analyzed the fate potential of hematopoietic stem and progenitor cells (HSPCs), by tracking a broad class of oligopotent and multipotent progenitor cell subpopulations and observing samples on days 2, 4, and 6 (Fig. 4.5h). Here, we test how well CellOT and other baselines can learn the differentiation process of the cells observed on day 2 to the cells observed on days 4 and 6 (combined) and generalize from one subpopulation to another (o.o.d. setting). We learn two maps, where map T_o is trained exclusively on oligopotent cells, T_m on multipotent cells. I.i.d. versions of these maps are trained on both oligopotent and multipotent cells, such that each pair of i.i.d. and o.o.d. maps is evaluated on the same test set. Comparing the distributional distance between predicted and observed differentiated cell states using the MMD metric, CellOT outperforms current state-of-the-art methods in this i.i.d. setting for both the oligopotent and the multipotent subsets (see Fig. 4.5i). Furthermore, while baselines struggle to perform in either o.o.d. setting, CellOT is able to generalize its predictions in one direction, i.e., from multipotent cells to the oligopotent setting. In contrast to oligopotent cells, multipotent cells have a higher potency and thus can potentially differentiate into more cell types, and so we would expect T_m is more likely to generalize than T_o, trained on the less potent oligopotent cells. When predicting developmental

perturbations on multipotent cells using T_o, the differentiated cell fates cannot be recovered.
We further compare the performance at different time points and across cell types. Fig. 4.5j shows the accuracy of the modeled development of multipotent cells using map T_m individually for day 4 and day 6 cells, respectively. It is evident that CellOT achieves better results when predicting developmental dynamics short-range instead of states further away in time. This suggests a potential limitation for all of these methods, which might be unable to recover alignments over coarse time resolutions. Beyond, while the vast majority of cells on days 4 and 6 are still undifferentiated (undiff), some cells have evolved into neutrophils (neut), monocytes (mono), basophils (baso), lymphoid precursors (lymph), or dendritic cells (DC). As expected, the performance of CellOT drops in terms of the MMD metric for those cell types that are only sparsely represented in the dataset (see Fig. 4.5k).

4.4 DISCUSSION

Single-cell expression profiling provides a detailed look into the molecular states of individual cells, but it is destructive and does thus not allow continuous measurements of molecular properties over time. There have been numerous proposals for methods to uncover the dynamics of individual cells from population data, but all of them face the same challenge: sequentially observed distribution of cell states can be produced by multiple dynamics and mechanisms of gene regulation. The ill-defined nature of the problem makes it necessary to pose certain assumptions on the underlying cellular dynamics.
The mathematical foundation of this work builds on the biological intuition that perturbations incrementally alter the molecular profiles of cells. This principle aligns with the theory of optimal transport and, following previous work (Schiebinger et al., 2019), serves naturally as the model foundation of CellOT. If this principle is violated, however, and perturbations strongly disrupt the population to an unidentifiable level, the performance of CellOT as well as other methods drops (see Discussion). In these instances, more complicated mathematical machinery would be needed. Such tools, however, are currently unable to scale to settings with more than a few genes (Heydari et al., 2022). Thus, we rely on a fine granularity of the time course to recover large cell state changes between consecutive time points (Tritschler et al., 2019).

Furthermore, if a system exhibits rotations and oscillations within two consecutive snapshots not captured by measurements, models based on optimal transport and previous tools (Weinreb et al., 2018) will not be able to recover such complex dynamics. This is in part also due to the current choice of the cost function, which, due to theoretical constraints and practical performance, is set to the Euclidean distance (3.5). We leave it to future work, to investigate choices of alternative cost functions.
Beyond, the current system is not able to recover effects (other than cell flux) that change the distribution of cells between time points, for example, proliferation and death (Tritschler et al., 2019). Recent works propose extensions to the classical neural optimal transport scheme that account for cell death and birth (Lübeck et al., 2022).
Lastly, current developments in bioengineering aim at overcoming the technological limitation of destructive cell assays. Chen et al. (2022b) propose a transcriptome profiling approach that preserves cell viability. Weinreb et al. (2020) capture cell differentiation processes while clonally connecting cells and their progenitors through barcodes. These methods thus offer (lower-throughput) insights that provide individual trajectories of cells over time, i.e., an alignment between distinct measurement snapshots. Somnath et al. (2023) propose a novel algorithmic framework connected to optimal transport that are able to make use of such (partially) aligned datasets (Shi et al., 2023; Tong et al., 2023).

5

NEURAL OPTIMAL TRANSPORT WITH CONTEXT

...

— ..., ... (...)

A key challenge in the treatment of cancer is to predict the effect of drugs, or a combination thereof, on cells of a particular patient. To achieve that goal, single-cell sequencing can now provide measurements for individual cells, in treated and untreated conditions, but these are, however, not in correspondence. Given such examples of untreated and treated cells under different drugs, can we predict the effect of new drug combinations? We develop a general approach motivated by this and related problems, through the lens of *optimal transport (OT) theory*, and, in that process, develop tools that might be of interest for other application domains of OT. Given a collection of N pairs of measures (μ_i, ν_i) over $\mathbb{R}^d$ (cell measurements), tagged with a context c_i (encoding the treatment), we seek to learn a context-dependent, parameterized transport map $\mathcal{T}_\theta$ such that, on training data, that map $\mathcal{T}_\theta(c_i) : \mathbb{R}^d \to \mathbb{R}^d$ fits the dataset, in the sense that $\mathcal{T}_\theta(c_i)\sharp\mu_i \approx \nu_i$. Additionally, we expect that this parameterized map can generalize to unseen contexts and patients, to predict, given a patient's cells described in μ_{new}, the effect of applying context c_{new} on these cells as $\mathcal{T}_\theta(c_{\text{new}})\sharp\mu$.

Learning Mappings Between Measures From generative adversarial networks, to normalizing flows and diffusion models, the problem of learning maps that move points from a source to a target distribution is central to machine learning. OT theory (Santambrogio, 2015) has emerged as a principled approach to carry out that task: For a pair of measures μ, ν supported on $\mathbb{R}^d$, OT suggests that, among all maps T such that ν can be reconstructed by applying T to every point in the support of μ (abbreviated with the push-forward notation as $T\sharp\mu = \nu$), one should favor so-called Monge maps, which *minimize* the average squared-lengths of displacements $\|x - T(x)\|^2$. A rich literature, covered in Peyré and Cuturi (2019), addresses computational challenges of estimating such maps, with impactful applications to various areas of science (cf., Hashimoto et al., 2016; Schmitz et al., 2018; Schiebinger et al., 2019; Yang et al., 2020; Janati et al., 2020a; Bunne et al., 2023a).

Contributions We propose a framework that can leverage *labeled* pairs of measures $\{(c_i, (\mu_i, \nu_i))\}_i$ to infer a *global* parameterized map $\mathcal{T}_\theta$. Hereby, the context c_i belongs to an arbitrary set $\mathcal{C}$. We construct $\mathcal{T}_\theta$ so that it should be able, given a possibly unseen context label $c \in \mathcal{C}$, to output a map $\mathcal{T}_\theta(c) : \mathbb{R}^d \to \mathbb{R}^d$, that is itself the gradient of a convex function. To that end, we propose to learn these parameterized Monge maps $\mathcal{T}_\theta$ as the gradients of PICNNs, which we borrow from the foundational work of Amos et al. (2017). Our framework can be also interpreted as a hypernetwork (Ha et al., 2016): The PICNN architecture can be seen as an ICNN whose weights and biases are *modulated* by the context vector c, which parameterizes a *family* of convex potentials in $\mathbb{R}^d$. Because both ICNNs —and to a greater extent PICNNs— are notoriously difficult to train (Richter-Powell et al., 2021; Korotin et al., 2021a,b), we use closed-form solutions between Gaussian approximations to derive relevant parameter initializations for (P)ICNNs: These choices ensure that, *upon initialization*, the gradient of the (P)ICNNs mimics the affine Monge map obtained in closed form between Gaussian approximations of measures μ_i, ν_i (Gelbrich, 1990). Our framework is applied to three scenarios: Parameterization of transport through a real variable (time or drug dosage), through an auxiliary informative variable (cell covariates) and through action variables (genetic perturbations in combination) (see Fig. 5.1). Our results demonstrate the ability of our architectures to better capture on out-of-sample observations the effects of these variables in various settings, even when considering never-seen, composite context labels. These results suggest potential applications of conditional OT to model personalized medicine outcomes, or to guide novel experiments, where OT could serve as a predictor for never tested context labels.

5.1 ...

5.2 CONDOT: SUPERVISED TRAINING OF CONDITIONAL MONGE MAPS

We are given a dataset of N pairs of measures, each endowed with a label, $(c_i, (\mu_i, \nu_i)) \in \mathcal{C} \times \mathcal{P}(\mathbb{R}^d)^2$. Our framework builds upon two pillars: (i.) we formulate the hypothesis that an optimal transport $T_i^\star$ (or, equivalently, the gradient of a convex potential $f_i^\star$) explains how measure μ_i was mapped to ν_i, given context c_i; (ii.) we build on the multi-task hypothesis (Caruana, 1997) that all of the N maps $T_i^\star$ between μ_i and ν_i share a common set of parameters, that are *modulated* by context informations c_i. These ideas are summarized in an abstract regression model described below.

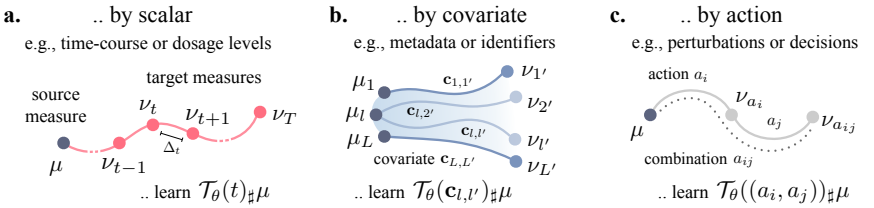


Figure 5.1: The evolution from a source μ to a target measure ν can depend on context variables c of various nature. This comprises **a.** scalars such as time or dosage t which determine the magnitude of an optimal transport, **b.** flow of measures into another one based on additional information (possibly different between μ and ν) stored in vectors $\mathbf{c}_{l,l'}$, or **c.** discrete and complex actions a_i, possibly in combination a_{ij}. We seek a unified framework to produce a map $\mathcal{T}_\theta(c)$ from any type of condition c.

5.2.1 *A Regression Formulation for Conditional OT Estimation*

$\theta \in \Theta \subset \mathbb{R}^r$, $\mathcal{T}_\theta$ describes a function that takes an input vector $c \in \mathcal{C}$, and outputs a *function* $\mathcal{T}_\theta(c) : \mathbb{R}^d \to \mathbb{R}^d$, as a hypernetwork would (Ha et al., 2016). Assume momentarily that we are given *ground truth* maps T_i, that describe the effect of context c_i on any measure, rather only pairs of measures (μ_i, ν_i). This is of course a major leap of faith, since even recovering an OT map $T^\star$ from two measures is in itself very challenging (Hütter and Rigollet, 2021; Rigollet and Stromme, 2022; Pooladian and Niles-Weed, 2021). If such maps were available, a direct supervised approach to learn a unique θ could hypothetically involve minimizing a fit function composed of losses between maps

$$\min_\theta \sum_{i=1}^N \int_{\mathbb{R}^d} \|\mathcal{T}_\theta(c_i)(x) - T_i(x)\|^2 \, \mathrm{d}\mu_i(x) \,. \tag{5.1}$$

Unfortunately, such maps T_i are not given, since we are only provided unpaired samples before μ_i and after ν_i that map's application. By Brenier's theorem, we know, however, that such an OT map $T_i^\star$ exists, and that it would be necessarily the gradient of a convex potential function that maximizes (3.6). As a result, we propose to modify (5.1) to (i.) parameterize, for any c, the map $\mathcal{T}_\theta(c)$ as the gradient w.r.t. x of a function $f_\theta(x, c) : \mathbb{R}^d \times \mathcal{C} \to \mathbb{R}$ that is convex w.r.t. x, namely $\mathcal{T}_\theta(c) := x \mapsto \nabla_1 f_\theta(x, c)$; (ii.) estimate θ by maximizing *jointly* the dual objectives (3.6) simultaneously for all N pairs of measures, in order to ensure that the maps are close to optimal, to form the aggregate problem

$$\max_\theta \textstyle\sum_{i=1}^N \mathcal{E}_{\mu_i,\nu_i}(f_\theta(\,\cdot\,, c_i)). \tag{5.2}$$

We detail in App. **??** how the Legendre transforms that appear in the energy terms $\mathcal{E}_{\mu_i,\nu_i}$ are handled with an auxiliary function.

5.2.2 *Integrating Context in Convex Architectures*

We propose to incorporate context variables, in order to modulate a family of convex functions $f_\theta(x, c)$ using partially input convex neural networks. PICNNs are neural networks that can be evaluated over a pair of inputs (x, c), but which are only required to be convex w.r.t. x. Given an input vector x and context vector c, a K-layer PICNN is defined as $\psi_\theta(x, c) = z_K$, where, recursively for $0 \leq k \leq K-1$ one has

$$\begin{aligned} u_{k+1} &= \tau_k \left(V_k u_k + v_k\right), \\ z_{k+1} &= \sigma_k \left(W_k^z \left(z_k \circ [W_k^{zu} u_k + b_k^z]_+\right) + W_k^x \left(x \circ (W_k^{xu} u_k + b_k^x)\right) + W_k^u u_k + b_k^u\right), \end{aligned} \tag{5.3}$$

where the PICNN is initialized as $u_0 = c, z_0 = \mathbf{0}$, $\circ$ denotes the Hadamard elementwise product, and τ_k is any activation function. The parameters of the PICNN are then given by

$$\theta = \{V_k, W_k^z, W_k^{zu}, W_k^x, W_k^{xu}, W_k^u, v_k, b_k^z, b_k^x, b_k^u\}.$$

Similar to ICNNs, the convexity w.r.t. input variable x is guaranteed as long as activation functions σ_i are convex and non-decreasing, and the weight matrices W_k^z have non-negative entries. We parameterize this by storing them as elementwise applications of softplus operations on precursor matrices of the same size, or, alternatively, by regularizing their negative part. Finally, much like ICNNs, all matrices at the $K-1$ layer are line vectors, and their biases scalars.

Such networks were proposed by Amos et al. (2017, Eq. 3) to address a problem that is somewhat symmetric to ours: Their inputs were labeled as (y, x), where y is a label vector, typically much smaller than that of vector x. Their PICNN is convex w.r.t. y, in order to easily recover, given a datapoint x (e.g., an image) the best label y that corresponds to x using gradient descent as a subroutine, i.e. $y^\star(x) = \arg\min_y \text{PICNN}_\theta(x, y)$. PICNN were therefore originally proposed to learn a parameterized, implicit classification layer, amortized over samples, whose motivation rests on the property that it is convex w.r.t. label variable y. By contrast, we use PICNNs that are convex w.r.t. data points x. In addition to that swap, we do not use the convexity of the PICNN to define an implicit layer (or to carry out gradient descent). Indeed, it does not make sense in our setting to minimize $\psi_\theta(x, c)$ as a

function of x, since x is an observation. Instead, our work rests on the property that $\nabla_1 \psi_\theta(x, c)$ describes a parameterized family of OT maps. We note that PICNNs were considered within the context of OT in (Fan et al., 2021, Appendix B). In that work, PICNN provide an elegant reformulation for neural Wasserstein barycenters. Fan et al. (2021) considered a context vector c that was restricted to be a small vector of probabilities.

5.2.3 *Conditional Monge Map Architecture*

Using PICNNs as a base module, the CONDOT architecture integrates operations on the contexts $\mathcal{C}$. As seen in Figure 5.1, context values c may take various forms:

1. A scalar t denoting a strength or a temporal effect. For instance, McCann's interpolation and its time parameterization, $\alpha_t = ((1-t)\mathrm{Id} + tT)_\sharp \alpha_0$ (McCann, 1997) can be interpreted as a trivial conditional OT model that creates, from an OT map ψ_θ, a set of maps parameterized by t, $\mathcal{T}_\theta(t) := x \mapsto \nabla_x \left((1-t)\|x\|^2/2 + t\psi_\theta(x)\right)$.
2. A covariate vector influencing the nature of the effect that led μ_i to ν_i, (capturing, e.g., patient feature vectors).
3. One or multiple actions, possibly discrete, representing decisions or perturbations applied onto μ_i.

To provide a flexible architecture capable of modeling different types of conditions as well as conditions appearing in combinations, the more general CONDOT architecture consists of the hypernetwork $\mathcal{T}_\theta$ that is fed a context vector through embedding and combinator modules. This generic architecture provides a one-size fits all approach to integrate all types of contexts c.

EMBEDDING MODULE To give greater flexibility when setting the context variable c, CONDOT contains an embedding module $\mathcal{E}$ that translates arbitrary contexts into real-valued vectors. Besides simple scalars t (Fig. 5.1a) for which no embedding is required, discrete contexts can be handled with an embedding module $\mathcal{E}_\phi$. When the set $\mathcal{C}$ is small, this can be done effectively using one-hot embeddings $\mathcal{E}_{\text{ohe}}$. For more complicated actions a such as treatments, there is no simple way to vectorize a context c. Similarly to action embeddings in reinforcement learning (Chandak et al., 2019; Tennenholtz and Mannor, 2019), we can learn embeddings for discrete actions into a learned continuous representation. This often requires domain-knowledge on the context values. For molecular drugs, for example, we can learn molec-

ular representations $\mathcal{E}_{\text{mol}}$ such as chemical, motif-based (Rogers and Hahn, 2010) or neural fingerprints (Rong et al., 2020; Schwaller et al., 2022). However, often this domain knowledge is not available. In this work, we thus construct so-called *mode-of-action* embeddings, by computing an embedding $\mathcal{E}_{\text{moa}}$ that encourages actions a with similar effect on target population ν to have a similar representation. In § 6.3, we analyze several embedding types for different use-cases.

COMBINATOR MODULE While we often have access to contexts c in isolation, it is crucial to infer the effect of contexts applied in combination. A prominent example are cancer combination therapies, in which multiple treatment modalities are administered in combination to enhance treatment efficacy (Kummar et al., 2010). In these settings, the mode of operation between individual contexts c is often not known, and can thus not be directly modeled via simple arithmetic operations such as `min, max, sum, mean`. While we test as a baseline the case, applicable to one-hot-embeddings, where simple additions are used to model these combinations, we propose to augment the CONDOT architecture with a parameterized combinator module $\mathcal{C}_\Phi$. If the order in which the actions are applied is irrelevant or unknown, the corresponding network $\mathcal{C}_\Phi$ needs to be permutation-invariant, which can be achieved by using a deep set architecture (Zaheer et al., 2017). Receiving a flexible number of inputs from the embedding module $\mathcal{E}_\phi$, CONDOT allows for a joint training of the PICNN parameters θ, embedding parameters ϕ, and combinator parameters Φ in a single, end-to-end differentiable architecture.

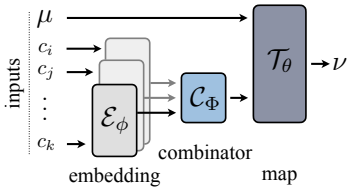


Figure 5.2: CondOT Architecture and Modules. The embedding module $\mathcal{E}_\phi$ embeds arbitrary conditions c, which are then combined via module $\mathcal{C}_\Phi$. Using the processed contexts c, the map $\mathcal{T}_\theta(c)$ acts on μ to predict the target measure ν.

TRAINING PROCEDURE. Given a dataset $\mathcal{D} = \{c_i, (\mu_i, \nu_i)\}_{i=0}^N$ of N pairs of populations before μ_i and after transport ν_i connected to a context c_i, we detail in Algorithm **??** provided in § **??**, a training loop that incorporates all of the architecture proposals described above. The training loss aims at making sure the map $\mathcal{T}_\theta(c_i)$ is an OT map from μ_i to ν_i, where c_i may either be the original label itself or its embedded/combined formulation in more advanced tasks. To handle the Legendre transform in (3.6), we

Method	Conditioned on Drug Dosage			
	In-Sample		Out-of-Sample	
	MMD	ℓ_2(PS)	MMD	ℓ_2(PS)
CPA (Lotfollahi et al., 2023)	0.1502 ± 0.0769	2.47 ± 2.89	0.1568 ± 0.0729	2.65 ± 2.75
ICNN OT (Makkuva et al., 2020)	0.0365 ± 0.0473	2.37 ± 2.15	0.0466 ± 0.0479	2.24 ± 2.39
CondOT (Identity initialization)	0.0111 ± 0.0055	0.63 ± 0.09	0.0374 ± 0.0052	2.02 ± 0.10
CondOT (Gaussian initialization)	0.0128 ± 0.0081	0.60 ± 0.11	0.0325 ± 0.0062	1.84 ± 0.14

Method	Conditioned on Cell Line	
	In-Sample	
	MMD	ℓ_2(PS)
CPA (Lotfollahi et al., 2023)	0.2551 ± 0.006	2.71 ± 1.51
ICNN OT (Makkuva et al., 2020)	0.0206 ± 0.0109	1.16 ± 0.75
CondOT (Identity initialization)	0.0148 ± 0.0078	0.39 ± 0.06
CondOT (Gaussian initialization)	0.0146 ± 0.0074	0.41 ± 0.07

Table 5.1: Evaluation of drug effect predictions from control cells to cells treated with drug Givinostat when conditioning on various covariates influencing cellular responses such as drug dosage and cell type. Results are reported based on MMD and the ℓ_2 distance between perturbation signatures of marker genes in the 1000 dimensional gene expression space.

use the proxy dual objective defined in (Makkuva et al., 2020, Eq. 6) (4.8)-(4.9) in place of (3.6) in our overall loss (5.2). This involves training the CondOT architecture using two PICNNs, i.e., PICNN_{θ_f} and PICNN_{θ_g}, that share the same embedding/combinator module, with a regularization (4.7) promoting that for any c, the $\text{PICNN}_{\theta_g}(\cdot, c)$ resembles the Legendre transform of the other, $\text{PICNN}^*_{\theta_f}(\cdot, c)$.

5.3 EMPIRICAL EVALUATION

Biological cells undergo changes in their molecular profiles upon chemical, genetic, or mechanical perturbations. These changes can be measured using recent technological advancements in high-resolution multivariate single-cell biology. Measuring single cells in their unperturbed or perturbed state requires, however, to destroy them, resulting in populations μ and ν that are unpaired. The relevance of OT to that comes from its ability to resolve such ambiguities through OT maps, holding promises of a better understanding of health and disease. We consider various high-dimensional problems arising from this scenario to evaluate the performance of CondOT (§ 5.2) versus other baselines.

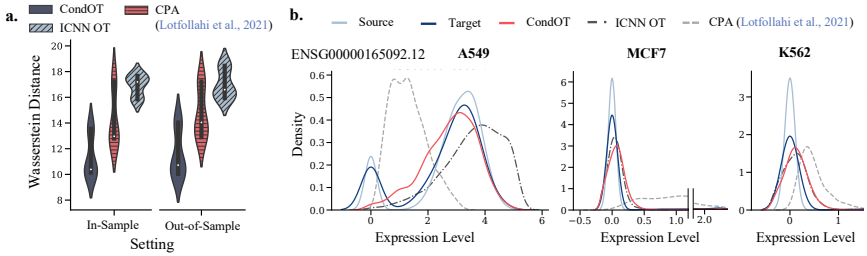


Figure 5.3: a. Predictive performance of CondOT and baselines w.r.t. the entropy-regularized Wasserstein distance on drug dosages *in-sample*, i.e., seen during training, and *out-of-sample*, i.e., unseen during training. **b.** Marginal distributions of observed source and target distributions, as well as predictions on perturbed distributions by CondOT and baselines of an exemplary gene across different cell lines. Predicted marginals of each method should match the marginal of the target population.

5.3.1 *Modeling Dosage-Sensitive Treatment Responses to Cancer Drugs*

Upon application of a molecular drug, the state of each cell x_i of the unperturbed population is altered, and observed in population ν. Molecular drugs are often applied at different dosage levels t, and the magnitude of changes in the gene expression profiles of single cells highly correlates with that dosage. We seek to learn a global, parameterized transport map $\mathcal{T}_\theta$ sensitive to that dosage.We evaluate our method on the task of inferring single-cell perturbation responses to the cancer drug Givinostat, a histone deacetylase inhibitor with potential anti-inflammatory, anti-angiogenic, and antineoplastic activities (Srivatsan et al., 2020), applied at different dosage levels, i.e., $t \in \{10\,\text{nM}, 100\,\text{nM}, 1,000\,\text{nM}, 10,000\,\text{nM}\}$. The dataset contains 3,541 cells described with the gene expression levels of 1,000 highly-variable genes. In a first experiment, we measure how well CondOT captures the drug effects at different dosage levels via distributional distances such as MMD (Gretton et al., 2012) and the ℓ_2-norm between the corresponding perturbation signatures (PS), as well as the entropy-regularized Wasserstein distance (Cuturi, 2013). We compute the metrics on 50 marker genes, i.e., genes mostly affected upon perturbation. To put CondOT's performance into perspective, we compare it to current state-of-the-art baselines (Lotfollahi et al., 2023) as well as parameterized Monge maps without context variables (Bunne et al., 2023b; Makkuva et al., 2020, ICNN OT). As visible in Table 5.1 and Fig. 5.3a, CondOT achieves consistently more accurate predictions of the target cell populations at

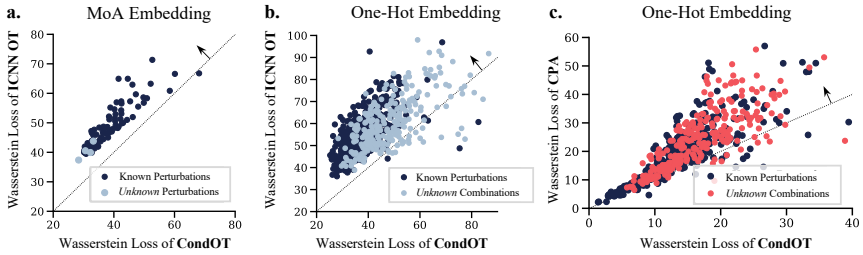


Figure 5.4: Comparison between **a.** CondOT and ICNN OT (Makkuva et al., 2020) based on embedding $\mathcal{E}_{\text{moa}}$ **b.** as well as $\mathcal{E}_{\text{ohe}}$, and **c.** CondOT and CPA (Lotfollahi et al., 2023) based on embedding $\mathcal{E}_{\text{ohe}}$ on *known* and *unknown* perturbations or combinations. Results above the diagonal suggest higher predictive performance of CondOT.

different dosage levels than OT approaches that cannot utilize context information, demonstrated through a lower average loss and a smaller variance. This becomes even more evident when moving to the setting where the population has been trained only on a subset of dosages and we test CondOT on *out-of-sample* dosages. Table 5.1 and Fig. 5.3a demonstrate that CondOT is able to generalize to previously *unknown* dosages, thus learning to interpolate the perturbation effects from dosages seen during training.

5.3.2 *Predicting Cell Type-Specific Treatment Responses to Cancer Drugs*

Molecular processes are often highly dependent on additional covariates that steer experimental conditions, and which are not present in the features measures in population μ or ν. This can be, for instance, factors such as different cell types clustered within the populations. When the model can only be conditioned w.r.t. a small and *fixed* set of metadata information, such as cell types, it is sufficient to encode these contexts using a one-hot embedding module $\mathcal{E}_{\text{ohe}}$. To illustrate this problem, we consider cell populations comprising three different cell lines (A549, MCF7, and K562). As visible in Table 5.1, CondOT outperforms current baselines which equally condition on covariate information such as CPA (Lotfollahi et al., 2023), assessed through various evaluation metrics. Figure 5.3b displays a gene showing highly various responses towards the drug Givinostat dependent on the cell line. CondOT captures the distribution shift from control to target populations consistently across different cell lines.

5.3.3 *Inferring Gene Knockout Responses*

To recommend personalized medical procedures for patients, or to improve our understanding of genetic circuits, it is key to be able to predict the outcomes of novel perturbations, arising from combinations of drugs or of genetic perturbations. Rather than learning individual maps T_θ^a predicting the effect of individual treatments, we aim at learning a global map $\mathcal{T}_\theta$ which, given as input the unperturbed population μ as well as the action a of interest, predicts the cell state perturbed by a. Thanks to its modularity, CondOT can not only learn a map T_θ for all actions *known* during training, but also to generalize to *unknown* actions, as well as potential *combinations* of actions. We will discuss all three scenarios below.

5.3.3.1 *Known Actions*

In the following, we analyze CondOT's ability to accurately predict phenotypes of genetic perturbations based on single-cell RNA-sequencing pooled CRISPR screens (Norman et al., 2019; Dixit et al., 2016), comprising 98,419 single-cell gene expression profiles with 92 different genetic perturbations, each cell measured via a 1,500 highly-variable gene expression vector. As, in a first step, we do not aim at generalizing beyond perturbations encountered during training, we utilize again a one-hot embedding $\mathcal{E}_{\text{ohe}}$ to condition $\mathcal{T}_\theta$ on each perturbation a. We compare our method to other baselines capable of modeling effects of a large set of perturbations such as CPA (Lotfollahi et al., 2023). Often, the effect of genetic perturbations are subtle in the high-dimensional gene expression profile of single cells. Using ICNN-parameterized OT maps without context information, we can thus assess the gain in accuracy of predicting the perturbed target population by incorporating context-awareness over simply predicting an average perturbation effect. Figure 5.4a and b demonstrate that compared to OT ablation studies, Fig. 5.4c and Fig. 5.5a for the current state-of-the-art method CPA (Lotfollahi et al., 2023). Compared to both, CondOT captures the perturbation responses more accurately w.r.t. the Wasserstein distance.

5.3.3.2 *Unknown Actions*

With the emergence of new perturbations or drugs, we aim at inferring cellular responses to settings not explored during training. One-hot embeddings, however, do not allow us to model *unknown* perturbations. This requires us to use an embedding $\mathcal{E}$, which can provide us with a representation of an unknown action a'. As genetic perturbations further

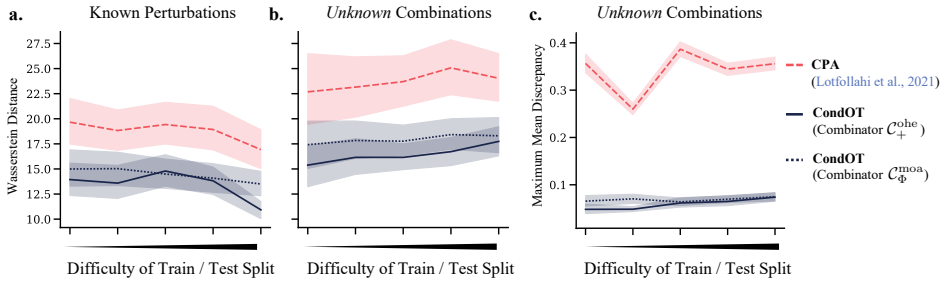


Figure 5.5: Predictive performance for **a.** known perturbations, **b.** unknown perturbations in combination w.r.t. regularized Wasserstein distance and **c.** MMD over different train / test splits of increasing difficulty for baseline CPA as well as CondOT with different combinators $\mathcal{C}_{+}^{\text{ohe}}$ and $\mathcal{C}_{\Phi}^{\text{moa}}$.

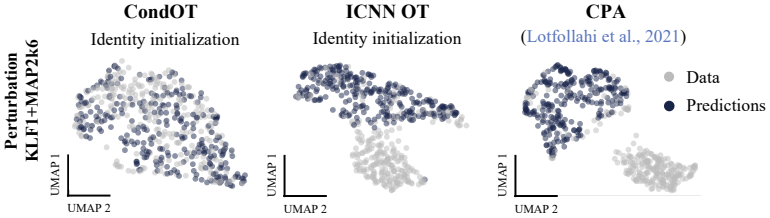


Figure 5.6: UMAP embeddings of cells perturbed by the combination KLF1+MAP2K6 (gray) and predictions of CondOT (ours), ICNN OT (Makkuva et al., 2020), and CPA (blue). While CondOT aligns well with observed perturbed cells, the baselines fail to capture subpopulations.

have no meaningful embeddings as, for example, molecular fingerprints for drugs, we resort to mode-of-action embeddings introduced in § 5.2.3. Assuming marginal sample access to all individual perturbations, we compute a multidimensional scaling (MDS)-based embedding from pairwise Wasserstein distances between individual target populations, such that perturbations with similar effects are closely represented. As current state-of-the-art methods are restricted to modeling perturbations via one-hot encodings, we compare our method to ICNN OT only. As displayed in Fig. 5.4a, CondOT accurately captures the response of *unknown* actions (BAK1, FOXF1, MAP2K6, MAP4K3), which were not seen during training, at a similar Wasserstein loss as perturbation effects seen during training.

5.3.3.3 *Actions in Combination*

While experimental studies can often measure perturbation effects in biological systems in isolation, the combinatorial space of perturbations in composition is too large to capture experimentally. Often, however, combination therapies are cornerstones of cancer therapy (Mokhtari et al., 2017). In the following, we test different combinator architectures to predict genetic perturbations in combination from single targets. Similarly to Lotfollahi et al. (2023), we can embed combinations by adding individual one-hot encodings of single perturbations (i.e., $\mathcal{C}_{+}^{\text{ohe}}$). In addition, we parameterize a combinator via a permutation-invariant deep set, as introduced in § 5.2.3, based on mode-of-action embeddings of individual perturbations (i.e., $\mathcal{C}_{\Phi}^{\text{moa}}$). We split the dataset into train / test splits of increasing difficulty: Initially containing all individual perturbations as well as some combinations, the number of perturbations seen in combination during training decreases over each split. We compare different combinators to ICNN OT (Fig. 5.4b) and CPA (Lotfollahi et al., 2023) (Fig. 5.4c, Fig. 5.5b, c). While the performance drops compared to inference on *known* perturbations (Fig. 5.5a) and decreases with increasing difficulty of the train / test split, CondOT outperforms all baselines. When embedding these high-dimensional populations in a low-dimensional UMAP space (McInnes et al., 2018), one can see that CondOT captures the entire perturbed population, while ICNN OT and CPA fail in capturing certain subpopulations in the perturbed state (see Fig. 5.6).

5.4 DISCUSSION

We have developed the CondOT framework that is able to infer OT maps from not only one pair of measures, but many pairs that come labeled with a context value. To ensure that CondOT encodes optimal transports, we parameterize it as a PICNN, an input-convex NN that modulates the values of its weights matrices according to a sequence of feature representations of that context vector. We showcased the generalization abilities of CondOT in the extremely challenging task of predicting outcomes for unseen combinations of treatments. These abilities and PICNN more generally hold several promises, both as an augmentation of the `OTT` toolbox (Cuturi et al., 2022), and for future applications of OT to single-cell genomics.

Part II

DYNAMIC NEURAL OPTIMAL TRANSPORT

6

LEARNING DYNAMICAL SYSTEMS VIA OPTIMAL TRANSPORT AND GRADIENT FLOWS

Ein solches mathematisch-definierbares System ist überhaupt nicht die Wirklichkeit selbst, sondern nur ein Schema, welches zur Beschreibung der Wirklichkeit dienen kann.

— Andrey Kolmogorov, *Über die analytischen Methoden in der Wahrscheinlichkeitsrechnung* (1931)

MODELING PARTICLE DYNAMICS AS A JKO SCHEME. In this paper, we draw inspiration from both approaches above—the intuition from the recent normalizing flows (NF) literature that flows should mimic an optimal transport (OT as prior), and be able, through training, to predict future configurations (OT as a loss)—to propose a causal model for population dynamics. Our approach relies on a powerful hammer: the Jordan-Kinderlehrer-Otto (JKO) flow (Jordan et al., 1998), widely regarded as one of the most influential mathematical breakthroughs in recent history. While the JKO flow was initially introduced as an alternative method to solve the Fokker-Planck partial differential equation (PDE), its flexibility can be showcased to handle more complex PDEs (Santambrogio, 2017, §4.7), or even describe the gradient flows of non-differentiable energies that have no PDE representation. On a purely mechanical level, a JKO step is to measures what the proximal step (Combettes and Pesquet, 2011) is to vectors: In a JKO step, particles move to decrease collectively an *energy* (a real-valued function defined on measures), yet remain close (in Wasserstein sense) to the previous configuration. Our goal in this paper is to treat JKO steps as parameterized modules, and fit their parameter (the energy function) so that its outputs agree repeatedly over time with observed data. This approach presents several challenges: While numerical approaches to solve JKO steps have been proposed in low dimensional settings (Burger et al., 2010; Carrillo et al., 2021; Peyré, 2015; Benamou et al., 2016a), scaling it to higher dimensions is an open problem. Moreover, minimizing a loss involving a JKO step w.r.t. energy requires not only solving the JKO problem, but also computing the (transpose) Jacobian of its output w.r.t. energy parameters.

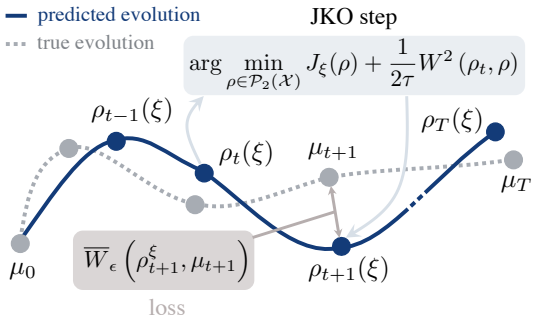


Figure 6.1: Given an observed trajectory $(\mu_0,\dots,\mu_T)$ of point clouds (gray), we seek parameters ξ for the energy J_ξ such that the predictions $\rho_1,\dots,\rho_T$ (blue) following a JKO flow from $\rho_0=\mu_0$ are close the observed trajectory (gray), by minimizing (as a function of ξ) the sum of Wasserstein distances between ρ_{t+1}, the JKO step from ρ_{t-1} using J_ξ, and data μ_{t+1}.

CONTRIBUTIONS. Our contributions are two-fold. First, we propose a method, given an input configuration and an energy function, to compute JKO steps using input convex neural networks input convex neural network (Amos et al., 2017; Makkuva et al., 2020) (see also concurrent works that have proposed similar approaches (Alvarez-Melis et al., 2022; Mokrov et al., 2021)). Second, we view the JKO step as an inner layer, a JKONET module parameterized by an energy function, which is tasked with moving the particles of an input configuration along an OT flow (the gradient of an optimal ICNN), trading off a lower energy with proximity to the previous configuration. We propose to estimate the parameters of the energy by minimizing a fitting loss computed between the outputs of the JKONET module (the prediction) and the ground truth displacements, as illustrated in Figure 6.1. We demonstrate JKONET's range of applications by applying in on synthetic potential- and trajectory-based population dynamics, as well as developmental trajectories of human embryonic stem cells based on single-cell genomics data.

6.1 ON THE CONNECTION BETWEEN OPTIMAL TRANSPORT AND FOKKER-PLANCK EQUATIONS

JKO FLOWS. In their seminal paper, Jordan et al. (1998) study diffusion processes under the lens of the OT metric (see also Ambrosio et al., 2006) and introduce a scheme that is now known as the JKO flow: Starting with

ρ_0, and given a real-valued energy function $J : \mathcal{P}(\mathbb{R}^d) \to \mathbb{R}$ driving the evolution of the system, they define iteratively for $t \geq 0$, :

$$\rho_{t+1} = \arg \min_{\rho \in \mathcal{P}_2(\mathbb{R}^d)} J(\rho) + \frac{1}{2\tau} W^2(\rho, \rho_t)\,, \tag{6.1}$$

where τ is a time step parameter. These successive minimization problems result in a sequence of probability measures in $\mathcal{P}(\mathbb{R}^d)$. The JKO flow can thus be seen as the analogy of the usual proximal descent scheme, tailored for probability measures (Santambrogio, 2015, p.285). Jordan et al. (1998) show that as step size $\tau \to 0$, and for a specific energy J that is the sum of a linear term and the negentropy, the measures describing the JKO flow recover solutions to a Fokker-Planck equation. In this work, following in the footsteps of more general applications of the JKO scheme (Santambrogio, 2017, §4.8), we model dynamics without necessarily having in mind PDE solutions in mind, to interpret instead the JKO step as a more general parametric type of dynamic for probability measures, exclusively parameterized by the energy J itself.

6.2 JKONET: A PROXIMAL OPTIMAL TRANSPORT MODEL

Given T discrete measures $\mu_0, \ldots, \mu_T$ describing the time evolution of a population, we posit that such an evolution follows a JKO flow for the free energy functional J, and assume that energy does not change throughout the dynamic. We parameterize the energy J as a neural network with parameters ξ, and fit ξ so that the JKO flow model matches the observed data.
Fitting parameter ξ with a reconstruction loss requires, using the chain rule, being able to differentiate the JKO step's output w.r.t. ξ (see Fig. 6.1), and more precisely provide a way to apply that transpose Jacobian to an arbitrary vector when using reverse-mode differentiation. To achieve this, we introduce a novel approach to numerically solve JKO flows using ICNNs (§ 6.2.1), resulting in a bilevel optimization problem targeting the energy J_ξ (§ 6.2.2).

6.2.1 *Reformulation of JKO Flows via ICNNs*

Given a starting condition ρ_t and energy functional J_ξ, the JKO step consists in producing a new measure ρ_{t+1} implicitly defined as the minimizer of (6.1). Solving directly (6.1) on the space of measures, involves substantial

computational costs. Different numerical schemes have been developed, e.g., based notably on Eulerian discretization of measures (Carrillo et al., 2021; Benamou et al., 2016b), and/or entropy-regularized optimal transport (Peyré, 2015). However, these methods are limited to small dimensions since the cost of discretizing such spaces grows exponentially. Except for the Eulerian approach proposed in (Peyré, 2015), obtained as the fixed point of a Sinkhorn type iteration, the differentiation would also prove extremely challenging as a function of the energy parameter ξ.

To reach scalability and differentiability, we build upon the approach outlined in Benamou et al. (2016b) to reformulate the JKO scheme as a problem solved over convex functions, rather than on measures ρ. Effectively, this is equivalent to making a change of variables in (6.1): Introduce a (variable) convex function ψ, and replace the variable ρ by the variable $\nabla\psi_{\#}\rho_t$. Writing

$$\mathcal{E}_J(\rho,\nu) := J(\rho) + \frac{1}{2\tau}W_2^2(\rho,\nu), \tag{6.2}$$

this identity states that, assuming μ and ν being absolutely continuous w.r.t. Lebesgue measure that

$$\min_{\rho}\mathcal{E}_J(\rho,\nu) = \min_{\psi\ \text{convex}}\mathcal{F}_J(\psi,\nu) := \mathcal{E}_J(\nabla\psi_{\#}\nu,\nu)\,,$$

simplifying the Wasserstein term in (6.2), using the assumption that ψ is convex and Brenier's theorem (**??**):

$$\mathcal{F}_J(\psi,\nu) = J(\nabla\psi_{\#}\nu) + \frac{1}{2\tau}\int\|x-\nabla\psi(x)\|^2 d\nu(x) \tag{6.3}$$

We pick an ICNN architecture to optimize over a restricted family of convex functions, $\{\psi_\theta\}$, and define, starting from $\rho_0(\xi) := \mu_0$, the recursive sequence for $t \geq 0$,

$$\rho_{t+1}(\xi) := \nabla\psi_{\theta^\star(\xi,\rho_t(\xi))\,\#}\,\rho_t(\xi)\,, \tag{6.4}$$

with $\theta^\star(\xi,\rho_t)$ defined implicitly using ξ and any ν as

$$\theta^\star(\xi,\nu) := \arg\min_{\theta}\mathcal{F}_J(\psi_\theta,\nu) \tag{6.5}$$

STRONG CONVEXITY OF ψ_θ. The strong convexity and smoothness of a potential ψ impacts the regularity of the corresponding OT map $\nabla\psi$ (Caffarelli, 2000; Figalli, 2010), since one can show that for a ℓ-strongly convex, L-smooth ψ one has (Paty et al., 2020) that

$$\ell\|x-y\| \leq \|\nabla\psi(x)-\nabla\psi(y)\| \leq L\|x-y\|.$$

While it is more difficult to enforce the L-smoothness of a neural network, and more generally its Lipschitz constants (Scaman and Virmaux, 2018) it is easy to enforce its strong convexity, by simply adding a term $\ell\|x\|^2/2$ to the corresponding potential, or a residual rescaled term ℓx to the output $\nabla\psi(x)$. This approach can be used to enforce that the push-forward of the gradient of an ICNN does not collapse to a single point, maintaining spatial diversity.

6.2.2 *Learning the Free Energy Functional*

The energy function $J_\xi : \mathcal{P}(\mathbb{R}^d) \to \mathbb{R}$ can be any parameterized function taking a measures as an input. Since our model assumes that the observed dynamic is parameterized entirely by that energy (and the initial observation ρ_0), the more complex this dynamic, the more complex one would expect the energy J_ξ to be. We focus in this first attempt on linear functions in the space of measures, that is expectations over ρ of a vector-input neural network E_ξ

$$J_\xi(\rho) := \int E_\xi(x) d\rho(x), \tag{6.6}$$

where $E_\xi : \mathbb{R}^d \to \mathbb{R}$ is a multi-layer perceptron (MLP).

Inferring nonlinear energies accounting for population growth and decline, as well as interactions between points, using the formalism of (De Bie et al., 2019), transformers (Vaswani et al., 2017) or set pooling methods (Edwards and Storkey, 2017; Zaheer et al., 2017), is an exciting direction for future work.

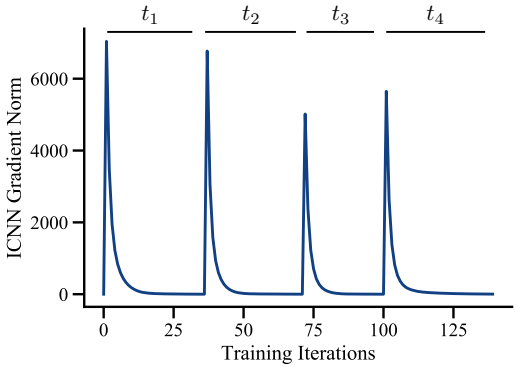


Figure 6.2: Optimization of the ICNN used in JKO steps. The bumps correspond to a change in the outer iteration, the smooth decrease in between correspond to a single minimization (6.5) of a time step t_i.

To address slow convergence and instabilities for dynamics with many snapshots, we use teacher forcing (Williams and Zipser, 1989) to learn J_ξ through time. In those settings, during training, J_ξ uses the

Algorithm 1 JKONET

Input: Dataset $\mathcal{D} = \{\{\mu_t^0\}_{t=0}^T, \ldots, \{\mu_t^N\}_{t=0}^T\}$ of N population trajectories, ξ^0 energy parameter initialization, θ^0 ICNN parameter initialization, learning rates lr_θ and lr_ξ, step τ, regularizer ε, tolerance α, `TeacherForcing` flag

Output: Free energy J_ξ explaining underlying population dynamics of snapshot data

$\xi \leftarrow \xi^0$
for $\{\mu_t\}_{t=0}^T \in \mathcal{D}$ **do**
  for $t \leftarrow 0$ **to** $T-1$ **do**
    $\theta \leftarrow \theta^0$
    if `TeacherForcing` **then**
      $\nu \leftarrow \mu_t$
    else
      $\nu \leftarrow \rho_t(\xi)$
    while $\frac{\sum_i \|\nabla_{\theta_i} \mathcal{F}_{J_\xi}(\theta)\|_2}{\sum_i \mathrm{count}(\theta_i)} \geq \alpha$ **do**
      $\theta \leftarrow \theta - \mathrm{lr}_\theta \times \nabla_\theta \mathcal{F}_{J_\xi, \nu}(\theta)$
    $\rho_{t+1}(\xi) \leftarrow \nabla\psi_{\theta\#}\nu$
    $\xi \leftarrow \xi - \mathrm{lr}_\xi \times \nabla_\xi \overline{W}_\varepsilon(\rho_{t+1}(\xi), \mu_{t+1})$

ground truth as input instead of predictions from the previous time step. At test time, we do not use teacher forcing.

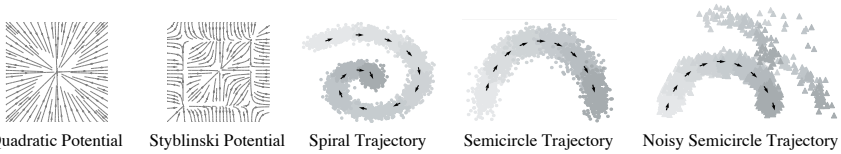
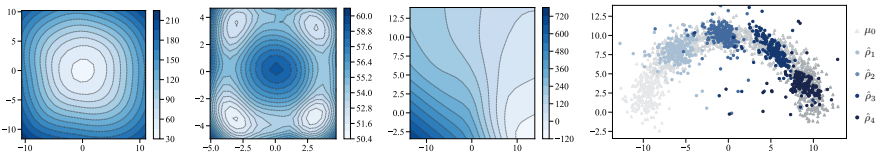


Figure 6.3: Overview on different tasks including trajectory- and potential-based dynamics.

6.2.3 *Bilevel Formulation of* JKONET

Learning the free energy functional J_ξ while solving each JKO step via an ICNN results in a challenging bilevel optimization problem. At each time



a. Quadratic Potential. **b.** Styblinski Potential. **c.** Semicircle Trajectory. **d.** Predicted Evolution on Semicircle Trajectory.

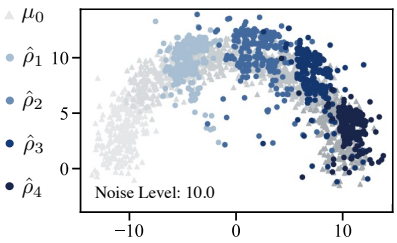
Figure 6.4: Results of JKOnet on potential- and trajectory-based dynamics. (a)-(c) Contour plots of the energy functionals J_ξ of JKONET on potential- and trajectory-based population dynamics, color gradients depict the magnitude of J_ξ. (d) Predicted population snapshots $(\hat{\rho}_1, \ldots, \hat{\rho}_4)$ (blue) and data trajectory $(\mu_0, \ldots, \mu_4)$ (gray).

step, the predicted dynamics are compared to the ground truth trajectory $(\mu_0, \mu_1, \ldots, \mu_T)$ with a Sinkhorn loss (3.3),

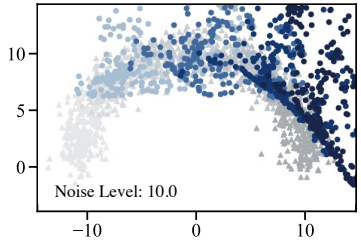
$$\begin{aligned} &\min_{\xi} \sum_{t=0}^{T-1} \overline{W}_\varepsilon(\rho_{t+1}(\xi), \mu_{t+1}), \\ &\text{s.t. } \rho_0(\xi) := \mu_0, \\ &\quad \rho_{t+1}(\xi) := \nabla\psi_{\theta^\star} \# \rho_t(\xi)\,, \\ &\quad \theta^\star := \arg\min_{\theta} \mathcal{F}_{J_\xi}(\psi_\theta, \rho_t(\xi)) \end{aligned} \tag{6.7}$$

The dependence of the Sinkhorn divergence losses in (6.7) on ξ only appears in the fact that the predictions $\rho_{t+1}(\xi)$ are themselves implicitly defined as solving a JKO step parameterized with the energy J_ξ. Learning J_ξ through the exclusive supervision of data observations requires therefore to differentiate the arg-minimum of a JKO problem, down therefore through to the lower-level optimization of the ICNN. We achieve this by implementing a differentiable double loop in `JAX`, differentiating first the Sinkhorn divergence using the `OTT` package (Cuturi et al., 2022), and then backpropagating through the ICNN optimization by unrolling Adam steps (Kingma and Ba, 2014; Metz et al., 2017; Lorraine et al., 2020).

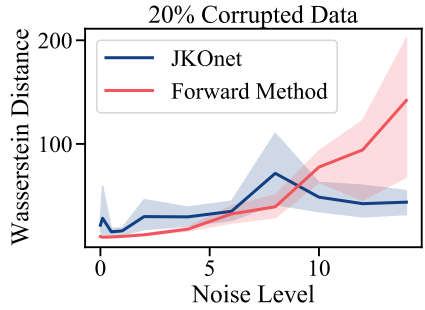
INNER LOOP TERMINATION. A question that arises when defining $\rho_{t+1}(\xi)$ lies in the budget of gradient steps needed or allowed to optimize the parameters θ of the ICNN, before taking a new gradient step on ξ in the outer loss. A straightforward approach in `JAX` (Bradbury et al., 2018) would be to use a preset number of iterations with a `for` loop (`jax.lax.scan`). We do observe, however, that the number of iterations needed to converge in relevant scenarios can vary significantly with the ICNN architecture



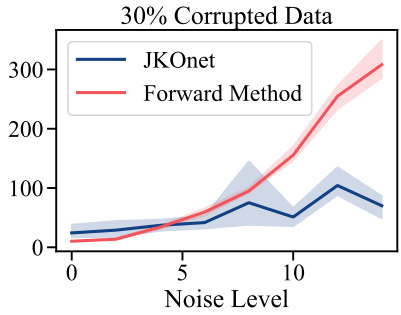
a. JKONET (30% corrupted data).



b. Forward method (30% corrupted data).



c. W_ϵ vs. noise level (20% corrupted data).

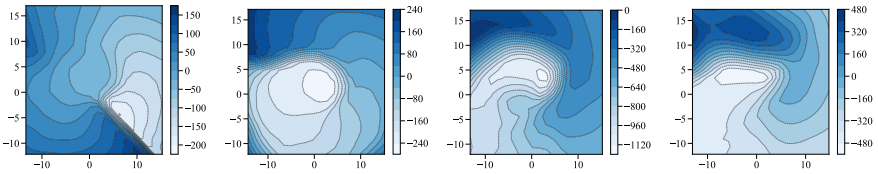


d. W_ϵ vs. noise level (30% corrupted data).

Figure 6.5: Comparison between JKONET and the forward method in settings of increasing noise on corrupted data on the semicircle trajectory task.

and/or the hardness of the underlying task. We propose to use instead a differentiable fixed-point loop to solve each JKO step up to a desired convergence threshold. We measure convergence of the optimization of the ICNN via the average norm of the gradient of the JKO objective w.r.t. the ICNN parameters θ, i.e., $\sum_i \|\nabla_{\theta_i} \mathcal{F}_{J_\xi}(\theta_i, \xi)\|_2 / \sum_i \text{count}(\theta_i)$. We observe that this approach is robust across datasets and architectures of the ICNN. An exemplary training curve for the ICNNs updated successively along a time sequence is shown in Figure 6.2.

REVERSE-MODE DIFFERENTIATION. The Jacobian $\partial \rho_{t+1} / \partial \xi$ arising when computing the gradient $\nabla_\xi \overline{W}_\varepsilon(\rho_{t+1}(\xi), \mu_{t+1})$ is obtained by unrolling the while loop above. The gradient term of the Sinkhorn divergence w.r.t the first argument is given by the Danskin envelope theorem (Danskin, 1967).



a. Forward method, teacher forcing. **b.** Forward method, *no* teacher forcing. **c.** JKONET, teacher forcing. **d.** JKONET, *no* teacher forcing.

Figure 6.6: Comparison between energy functionals J_ξ of the spiral trajectory task (see 6.3) between the forward method and JKONET, trained with or without teacher forcing § 6.2.2). When using teacher forcing, the forward method overfits a gap on the lower-right corner of the spiral, outputting a highly irregular energy. When taking into account the entire trajectory recursively, the Forward method does better overall, but is unable to recover an energy as precise as that returned by JKONET.

Method	**Prediction Loss (W_ε)**			
	Day 6 to 9	Day 12 to 15	Day 18 to 21	Day 24 to 27
One-Step Ahead				
Forward Method	0.187 ± 0.001	0.162 ± 0.010	0.185 ± 0.020	0.203 ± 0.004
JKONET	$\mathbf{0.133 \pm 0.020}$	$\mathbf{0.133 \pm 0.008}$	$\mathbf{0.172 \pm 0.0130}$	$\mathbf{0.169 \pm 0.004}$
All-Steps Ahead				
Forward Method	0.225 ± 0.023	0.160 ± 0.001	0.171 ± 0.016	0.183 ± 0.007
JKONET	$\mathbf{0.148 \pm 0.015}$	$\mathbf{0.144 \pm 0.013}$	$\mathbf{0.154 \pm 0.024}$	$\mathbf{0.138 \pm 0.034}$

Table 6.1: Evaluation of predictive performance w.r.t. the entropy-regularized Wasserstein distance W_ε (3.2) of JKONET and the forward method on the embryoid body scRNA-seq data per time step (using 3 runs).

SETTING τ IN (6.3). In usual JKO applications, τ needs to be tuned manually. In this work, the energy J_ξ is not fixed, but trained to fit data. Since we put no constraints on the scaling of J_ξ, τ can be set to 1 without loss of generality, as the parameter ξ will automatically adjust so that the scale of J_ξ induces steps of a relevant length to fit data. This only holds (as with a usual JKO step) if the trajectories are sampled regularly. For irregularly spaced time series, τ can be adapted at train and test time to the spacing of timestamps (shorter steps requiring larger τ).

6.3 EMPIRICAL EVALUATION

In the following, we evaluate our method empirically on a variety of tasks. This includes recovering synthetic potential- and trajectory-based popula-

tion dynamics (see Fig. 6.3), as well as the evolution of high-dimensional single-cell populations during a developmental process.

6.3.1 ...

ENERGY-DRIVEN TRAJECTORIES. The first task involves evolutions of partial differential equations with known potential. We hereby consider both convex (e.g., the quadratic function $J(x) = \|x\|_2^2$) and nonconvex potentials (e.g., Styblinski function) (see Fig. 6.3). These two-dimensional synthetic flows are generated using the Euler-Maruyama method (Kloeden and Platen, 1992). To recover the true potential via JKONET, we parameterize both energy J_ξ and ICNN ψ_θ with linear layers ($\epsilon = 1.0$, $\tau = 1.0$. Figure 6.4a-b demonstrate JKONET's ability to recover convex and nonconvex potentials via energy J_ξ.

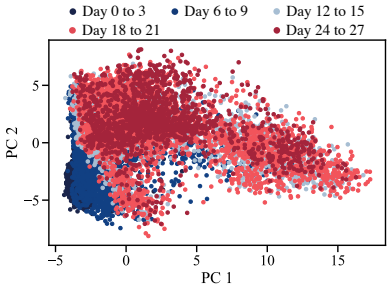
ARBITRARY TRAJECTORIES. As a sanity check, we evaluate if JKONET can recover an energy functional J_ξ from trajectories that are not necessarily arising from the gradient of an energy. Here, a 2-dimensional Gaussian moves along a predefined trajectory with nonconstant speed. We consider a line, a spiral, and movement along a semicircle (Fig. 6.3). As visible in Figure 6.4c (5 snapshots)], and Figure 6.6c-d (10 snapshots), JKONET learns energy functionals J_ξ that can then model the ground truth trajectories. These trajectory-based dynamics are learned using the strong convexity regularizer ($\ell = 0.8$, see § 6.2.1).

COMPARISON TO FORWARD METHODS. Instead of parameterizing the next iteration $\rho_{t+1}(\xi)$ as we do in the JKONET formulation (6.1), the *forward* scheme states that the prediction at time $t+1$, η_{t+1}, can be obtained as $(\nabla F_\xi)_\# \eta_t(\xi)$, where F_ξ is any arbitrary neural network, as considered in Hashimoto et al. (2016), namely $\eta_0 := \mu_0$ and subsequently $\eta_{t+1}(\xi) := (\nabla F_\xi)_\# \eta_t(\xi)$. Although OT still plays an important role in that paper, since the potential F is estimated by minimizing a Sinkhorn loss $\overline{W}_\varepsilon(\eta_{t+1}, \mu_{t+1})$, as we do in (6.7), the forward displacement operator $(\nabla F_\xi)_\#$ has no spatial regularity. Because of that, we observe that the forward method can get more easily trapped in local minima, and, in particular, overfits the training data as shown by a substantial decrease in performance in the presence of noise. We demonstrate this by comparing the robustness of both JKONET and the forward method to noise. For this, we corrupt 20% or 30% of the training data on the example of the semicircle trajectory with different levels

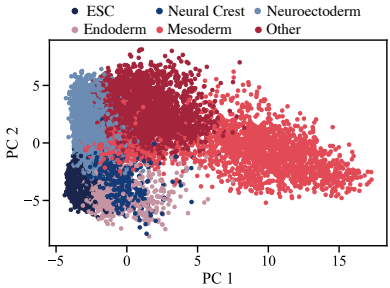
of noise (see Fig. 6.3). We insist that noise is only added at training time, as random shifts on both feature dimensions, while we test on the original semicircle trajectory. In low noise regimes, where train and test data are similar, the forward method overfits and performs marginally better than JKONET (see Fig. 6.5c,d). As noise increases, the performance of the forward method deteriorates (Fig. 6.5b), while JKONET, constrained to move points with OT maps, is robust (Fig. 6.5a).
Second, we compare the resulting energy functionals F_{ξ} and J_{ξ} of the forward method and JKONET, respectively, on the spiral trajectory (see Fig. 6.6). When learning long and complex population dynamics, teacher forcing improves training (see Fig. 6.4c-d). While facilitating training of the forward method in some settings, it likewise results in wrong energy functionals F_{ξ} (Fig. 6.6a). JKONET, on the other hand, is able to globally learn the energy functional J_{ξ}, despite being only exposed to a one-step history of snapshots during training with teacher forcing (see Fig. 6.6c).

6.3.2 ...

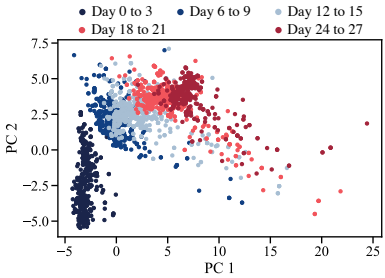
We investigate the ability of JKONET to predict the evolution of cellular and molecular processes through time. The advent of single cell profiling technologies has enabled the generation of high-resolution single-cell data, making it possible to profile individual cells at different states in the development. A key difficulty in learning the evolution of cell populations is that a cell is (usually) destroyed during a measurement. Thus, although one is able to collect features at the level of individual cells, the same cell cannot be measured twice. Instead, we collect independent samples at each snapshot, resulting in *unaligned* distributions across snapshots, without access to ground-truth single-cell trajectories. The goal of learning individual dynamics is to identify ancestor and descendant cells, and get a better understanding of biological differentiation or reprogramming mechanisms. We apply JKONET to embryoid body single-cell RNA sequencing (scRNA-seq) data (Moon et al., 2019), describing the differentiation of human embryonic stem cells grown as embryoid bodies into diverse cell lineages over a period of 27 days. During this time, cells are collected at 5 different snapshots (day 1 to 3, day 6 to 9, day 12 to 15, day 18 to 21, day 24 to 27) and measured via scRNA-seq (resulting in 15,150 cells). We run JKONET as well as the baseline on the first 20 components of a principal component analysis (PCA) of the 4000 highly differentiable genes. We split the dataset



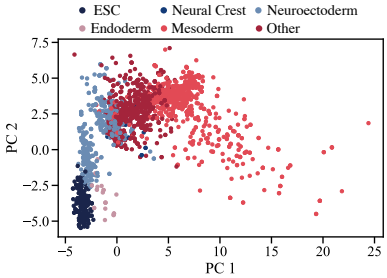
a. PCA embedding of the embryoid body scRNA-seq data colored by the snapshot time.



b. PCA embedding of the embryoid body scRNA-seq data colored by the lineage branch class.



c. PCA embedding of JKONET predictions colored by the snapshot time.

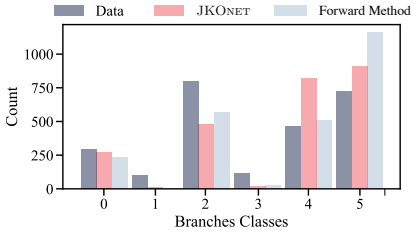


d. PCA embedding of JKONET predictions colored by the lineage branch class.

Figure 6.7: Analysis of population dynamics predictions of JKONET on the embryoid body scRNA-seq data.

into train and test data ($\sim 15\%$) and parameterize both energy J_ξ and ICNN ψ_θ with linear layers ($\epsilon = 1.0$, $\tau = 1.0$).

CAPTURING SPATIO-TEMPORAL DYNAMICS. Given the samples from the cell population at day 1 to 3 (μ_0), JKONET learns the underlying spatio-temporal dynamics giving rise to the developmental evolution of embryonic stem cells. As no ground truth trajectories are available in the data, we use distributional distances, i.e., the entropy-regularized Wasserstein distance W_ε (3.2) (Flamary et al., 2021), to measure the correctness of the predictions at each time step. We hereby measure the W_ε discrepancy between data and predictions for one-step ahead as well as inference of the entire evolution (all-steps ahead) for each time step t_i, see results in Table 6.1. JKONET outperforms the forward method in terms of W_ε (3.2) distance for both



Method	Cell Lineage Classification	
	ℓ_1	H^2
One-Step Ahead		
Forward Method	132.27 ± 5.00	0.026 ± 0.002
JKONET	$\mathbf{88.80 \pm 0.57}$	$\mathbf{0.016 \pm 0.001}$
All-Steps Ahead		
Forward Method	$\mathbf{185.47 \pm 12.18}$	0.033 ± 0.002
JKONET	215.60 ± 12.53	0.034 ± 0.004

Figure 6.8: Evaluation of cell lineage branch classification performance of JKONET and the forward method on the embryoid body scRNA-seq data based on the ℓ_1-distance of the histograms and the Hellinger distance H^2 (6.8) of the predicted branch class distributions (using 3 runs).

one-step ahead and all-steps ahead predictions for all time steps. The performance of both methods is relatively stable even until day 24 to 27, i.e., the W_ε distance does not significantly grow for future snapshots. We further visualize the first two principal components of the entire dataset (Fig. 6.7a) and of JKONET's predictions on the test dataset ($\sim$ 500 cells per snapshot, Fig. 6.7d).

CAPTURING BIOLOGICAL HETEROGENEITY. Besides measuring the ability of JKONET to model and predict the spatio-temporal dynamics of embryonic stem cells, we would like to guarantee, at a more macroscopic level, that JKONET is also able to learn the cell's differentiation into various cell lineages. Embryoid bodies differentiation covers key aspects of early embryogenesis and thus captures the development of embryonic stem cells into the mesoderm, endoderm, neuroectoderm, neural crest and others.
Following Moon et al. (2019, Fig. 6, Suppl. Note 4), we compute lineage branch classes (Fig. 6.7b) for all cells based on an initial k-means clustering ($k = 30$) in a 10-dimensional embedding space using PHATE, a non-linear dimensionality reduction method capturing a denoised representation of both local and global structure of a dataset. We then train a k-nearest neighbor (k-NN) classifier ($k = 5$) to infer the lineage branch class based on a 20-dimensional PCA embedding of a cell (classes: ESC: 0, neural crest: 1, neuroectoderm: 2, endoderm: 3, mesoderm: 4, other: 5).
We analyze the captured lineage branch heterogeneity of the population predicted by JKONET and the forward method by estimating the lineage branch class of each cell using the trained k-NN classifier. The predicted populations colored by the estimated lineage branch as well as the data with

the true lineage branch labels are visualized in Figure 6.7e and Figure 6.7b, respectively. The corresponding predicted and true distributions of lineage branch classes are shown in Figure 6.7c. To quantify how well JKONET and the forward method capture different cell lineage branches, we compute the ℓ_1 distance between the predicted and true histograms as well as the Hellinger distance

$$H^2(a,b) = \frac{1}{2}\sum_{i=1}^{k}\left(\sqrt{a_i/\|a\|_1} - \sqrt{b_i/\|b\|_1}\right)^2 \tag{6.8}$$

between both true and predicted class discrete distributions a and b. Figure 6.7c and Table 6.8 demonstrate that both, JKONET and the forward method, capture most lineage branches during the differentiation of embryonic stem cells. Both methods, however, have difficulties recovering cells of the neural crest (class 1) and the endoderm (class 3), lineage branches which are scarcely represented in the original data. The analysis further suggests that both methods reduce in performance w.r.t. biological heterogeneity when predicting the entire trajectory (all-steps ahead), instead of inferring the next snapshot only (one-step ahead).

6.4 DISCUSSION

We proposed JKONET, a model to infer and predict the evolution of population dynamics using a proximal optimal transport scheme, the JKO flow. JKONET solves local JKO steps using ICNNs and learns the energy that parameterizes these steps by fitting JKO flow predictions to observed trajectories using a fully differentiable bilevel optimization problem. We validate its effectiveness through experiments on synthetic potential- and trajectory-based population dynamics, and observe that it is far more robust to noise than a more direct Forward approach. We use JKONET to infer the developmental trajectories of human embryonic stem cells captured via high-dimensional and time-resolved single-cell RNAseq. Our analysis also shows that JKONET captures diverse cell fates during the incremental differentiation of embryonic cells into multiple lineage branches. Using proximal optimal transport to model real complex population dynamics thus makes for an exciting avenue of future work. Extensions could include modeling higher-order interactions among population particles in the energy function, e.g., cell-cell communication.

7

LEARNING DYNAMICAL SYSTEMS VIA OPTIMAL TRANSPORT AND STOCHASTIC CONTROL

Living matter evades the decay to equilibrium.
— Erwin Schrödinger, *What is Life?* (1944)

7.1 DIFFUSION SCHRÖDINGER BRIDGES

Recall the general SB problem (7.10). It turns out that the solution to (7.10) is itself given by two coupled SDEs of the form (Léonard, 2013)

$$\mathrm{d}X_t = (f_t + g_t Z_t)\,\mathrm{d}t + g_t\,\mathrm{d}\mathbb{W}_t,\ X_0 \sim \hat{\mathbb{P}}_0, \tag{7.1a}$$
$$\mathrm{d}X_t = (f_t - g_t \hat{Z}_t)\,\mathrm{d}t + g_t\,\mathrm{d}\mathbb{W}_t,\ X_1 \sim \hat{\mathbb{P}}_1, \tag{7.1b}$$

where $Z_t, \hat{Z}_t \colon \mathbb{R}^d \to \mathbb{R}^d$ are two time-indexed smooth *vector fields* called the optimal forward and backward drift, respectively, and (7.1b) runs backward in time (i.e., from $1 \to 0$). If we parametrize the forward drift by $Z_t^\theta(x)$ and the backward drift by $\hat{Z}_t^\phi(x)$ with some parameters θ, ϕ, then the negative likelihood function for θ and ϕ can be expressed as (Chen et al., 2022a)

$$\ell(x_0; \phi) = \int_0^1 \mathbb{E}_{(7.1a)}\Big[\frac{1}{2}\|\hat{Z}_t^\phi\|^2 + g\nabla_x \cdot \hat{Z}_t^\phi + \langle Z_t^\theta, \hat{Z}_t^\phi\rangle\,\mathrm{d}t \Big| X_0 = x_0\Big], \tag{7.2a}$$
$$\ell(x_1; \theta) = \int_0^1 \mathbb{E}_{(7.1b)}\Big[\frac{1}{2}\|Z_t^\theta\|^2 + g\nabla_x \cdot Z_t^\theta + \langle \hat{Z}_t^\phi, Z_t^\theta\rangle\,\mathrm{d}t \Big| X_1 = x_1\Big]. \tag{7.2b}$$

7.2 DATA-DRIVEN PRIORS FOR DIFFUSION SCHRÖDINGER BRIDGES

The Schrödinger bridge (SB) (Léonard, 2013; Chen et al., 2021b), alternatively known as the *dynamic* entropy-regularized optimal transport (OT), has recently received significant attention from the machine learning community. In contrast to the classical *static* OT where one seeks a coupling between measures that minimizes the average cost (Villani, 2009; Peyré and Cuturi, 2019), the goal of SBs is to find the optimal *stochastic processes* that evolve a given measure into another. As such, SBs are particularly suitable for learning complex continuous-time systems, and have been successfully applied to a wide range of applications such as sampling (Bernton et al.,

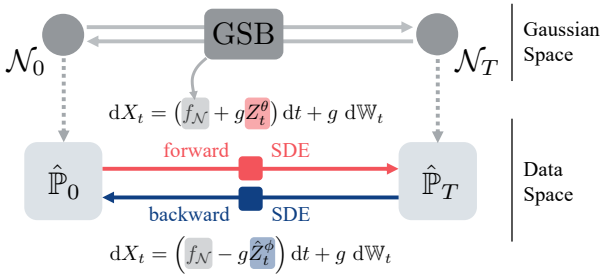


Figure 7.1: Solving the SB problem between $\hat{\mathbb{P}}_0$ and $\hat{\mathbb{P}}_1$ is notoriously difficult because it requires learning the time-dependent drifts of two SDEs that respect the desired marginals, and a random initialization for these drifts is usually extremely far from satisfying that constraint. We propose a data-dependent procedure that relies first on Gaussian approximations of the data measures, which provide a closed-form drift $f_{\mathcal{N}}$ in (7.31) (the GSB). We show that this facilitates the training of forward/backward drifts $\hat{Z}_t^\theta, \hat{Z}_t^\phi$.

2019; Huang et al., 2021c), generative modeling (Chen et al., 2022a; De Bortoli et al., 2021b; Wang et al., 2021), molecular biology (Holdijk et al., 2022), and mean-field games (Liu et al., 2022).
Despite of these impressive achievements, a common limitation of the existing works is that the SBs are typically solved in a purely numerical fashion. In sharp contrast, it is well-known that many important OT problems for *Gaussian* measures admit *closed-form* solutions, and the advantages of such solutions are numerous: they have inspired new learning methods (Rabin et al., 2011; Vayer et al., 2019; Bonneel et al., 2015), they can serve as the ground truth for evaluating numerical schemes (Janati et al., 2020b), and they have lead to the discovery of a new geometry that is both rich in theory and application (Takatsu, 2010).

CONTRIBUTIONS. The goal of our paper is to continue this pursuit of closed-form solutions and thereby extending these advantages to SB-based learning methods. For an overview of the method, see Fig. 7.1. To this end, we make the following contributions:

1. As our central result, we derive the closed-form expressions for Gaussian Schrödinger bridges (GSBs), i.e.,SBs between Gaussian measures. This is a challenging task for which all existing techniques fail, and thus we need to resort to a number of new ideas from entropic OT, Riemannian geometry, and generator theory; see Section 7.2.2.

2. We extend the deep connection between geometry and Gaussian OT to Gaussian Schrödinger bridges. In particular, our results can be seen as a vast generalization of the classical Bures-Wasserstein geodesics between Gaussian measures (Takatsu, 2010; Bhatia et al., 2019), which is the foundation of many computational methods (Chewi et al., 2020; Altschuler et al., 2021; Han et al., 2021).
3. Via a simple Gaussian approximation on real *single-cell genomics* data, we numerically demonstrate that many benefits of the closed-form expressions in static OT immediately carry over to SB-based learning methods: We report improved numerical stability and tuning insensitivity when trained on benchmark datasets, which ultimately lead to an overall better performance.

7.2.1 *Preliminaries on Gaussian Optimal Transport*

Throughout this paper, let $\xi \sim \mathcal{N}(\mu, \Sigma)$ and $\xi' \sim \mathcal{N}(\mu', \Sigma')$ denote two given Gaussian random variables. By abusing the notation, we will continue to denote the measures of these Gaussians by ξ and ξ', respectively. We will also denote by $\Pi(\xi, \xi')$ the set of all their couplings.

7.2.1.1 *Static Gaussian Optimal Transport*

The *static* entropy-regularized OT between Gaussians refers to the following minimization problem (Peyré and Cuturi, 2019):

$$\min_{\pi \in \Pi(\xi, \xi')} \int \|x - x'\|^2 \mathrm{d}\pi(x, x') + 2\sigma^2 D_{\mathrm{KL}} \pi \xi \otimes \xi', \tag{7.3}$$

where $\xi \otimes \xi'$ denotes the product measure of ξ and ξ', and $\sigma \geq 0$ is a regularization parameter. When $\sigma = 0$, (7.3) reduces to the classical 2-Wasserstein distance between ξ and ξ' (Villani, 2009), whose closed-form solution is classical (Dowson and Landau, 1982; Olkin and Pukelsheim, 1982). The case for general σ is more involved, and an analytical expression was only recently found (Bojilov and Galichon, 2016; del Barrio and Loubes, 2020; Janati et al., 2020b; Mallasto et al., 2021): Setting

$$D_\sigma := (4\Sigma^{\frac{1}{2}}\Sigma'\Sigma^{\frac{1}{2}} + \sigma^4 I)^{\frac{1}{2}}, \quad C_\sigma := \frac{1}{2}(\Sigma^{\frac{1}{2}} D_\sigma \Sigma^{-\frac{1}{2}} - \sigma^2 I), \tag{7.4}$$

then the solution $\pi^\star$ to (7.3) is itself a Gaussian:

$$\pi^\star \sim \mathcal{N}\left(\begin{bmatrix}\mu\\ \mu'\end{bmatrix}, \begin{bmatrix}\Sigma & C_\sigma\\ C_\sigma^\top & \Sigma'\end{bmatrix}\right). \tag{7.5}$$

7.2.1.2 *Dynamic Gaussian Optimal Transport*

In the literature, (7.3) is commonly referred to as the *static* OT formulation, since it merely asks *where* the mass should be transported to (i.e., $\pi(x, x')$ dictates how much mass at x should be transported to x'). In contrast, the more general problem of *dynamic* Gaussian OT seeks to answer *how* the mass the should be transported:

$$\min_{\rho_0=\xi,\ \rho_1=\xi'} \mathbb{E}\left[\int_0^1 \frac{1}{2}\|v_t\|^2 + \frac{\sigma^4}{8}\|\nabla \log \rho_t\|^2 \,\mathrm{d}t\right]. \tag{7.6}$$

Here, the minimization is taken over all pairs (ρ_t, v_t) where ρ_t is an absolutely continuous curve of measures (Ambrosio et al., 2006), and $v_t : \mathbb{R}^d \to \mathbb{R}^d$ is such that the continuity equation holds:

$$\partial_t \rho_t = -\nabla_x \cdot (\rho_t v_t), \tag{7.7}$$

where $(\nabla_x \cdot v_t)(x) := \sum_{i=1}^d \frac{\partial}{\partial x_i} v_t^i(x)$ denotes the divergence operator with respect to the x variable. It can be shown that, if $\rho_t^\star$ is the optimal curve for (7.6), then the joint distribution of the end marginals $(\rho_0^\star, \rho_1^\star)$ coincides with (7.5), hence the interpretation of $\rho_t^\star$ as the optimal *trajectory* in the space of measures (Chen et al., 2016; Gentil et al., 2017; Chen et al., 2021b; Gentil et al., 2020).

To our knowledge, the only work that has partially addressed the closed-form solution of (7.6) is Mallasto et al. (2021), whose results are nonetheless insufficient to cover important applications such as generative modeling. In Section 7.2.4, we will derive a vast generalization of the results in Mallasto et al. (2021) and provide a detailed comparison in Sections 7.2.2 to 7.2.3.

7.2.2 *The Gaussian Schrödinger Bridge Problem*

The purpose of this section is to introduce the core objectives in our paper, the Gaussian Schrödinger bridges, and establish their connection to the Gaussian OT problems in Section 7.2.1. To help the reader navigate our somewhat technical proofs in Sections 7.2.3 to 7.2.4, we illustrate in Section 7.2.2.2 the high-level challenges as well as our new techniques for solving Gaussian Schrödinger bridges.

7.2.2.1 *Schrödinger Bridges as Dynamic Entropy-Regularized Optimal Transport*

Let ν, ν' be two given measures and let $\mathbb{Q}_t$ be an arbitrary stochastic process. In its most generic form, the Schrödinger bridge refers to the following constrained KL-minimization problem over all stochastic processes $\mathbb{P}_t$ (Léonard, 2013; Chen et al., 2021b):

$$\min_{\mathbb{P}_0=\nu,\, \mathbb{P}_1=\nu'} D_{\mathrm{KL}}\mathbb{P}_t\mathbb{Q}_t. \tag{7.8}$$

In practice, ν and ν' typically arise as the (empirical) *marginal* distributions of a complicated continuous-time dynamics observed at the starting and end times, and $\mathbb{Q}_t$ is a "prior process" representing our belief of the dynamics before observing any data. The solution $\mathbb{P}_t^\star$ to (7.8) is thus interpreted as the best dynamics that conforms to the prior belief $\mathbb{Q}_t$ while respecting the data marginals ($\mathbb{P}_0^\star = \nu, \mathbb{P}_1^\star = \nu'$).
In this paper, we will consider a general class of $\mathbb{Q}_t$'s that includes most existing processes in the machine learning applications of SBs. Specifically, with some initial condition Y_0, we will take $\mathbb{Q}_t$ to be the measure of the linear stochastic differential equation (SDE):

$$\mathrm{d}Y_t = (c_t Y_t + \alpha_t)\,\mathrm{d}t + g_t\,\mathrm{d}\mathbb{W}_t := f_t\,\mathrm{d}t + g_t\,\mathrm{d}\mathbb{W}_t. \tag{7.9}$$

Here, $c_t : \mathbb{R}^+ \to \mathbb{R}$, $\alpha_t : \mathbb{R}^+ \to \mathbb{R}^d$, and $g_t : \mathbb{R}^+ \to \mathbb{R}^+$ are smooth functions. In this case, SBs can be seen as generalized dynamical OT between two (not necessarily Gaussian) measures:

Theorem 1. *Consider the Schrödinger bridge problem with Y_t as the reference process:*

$$\min_{\mathbb{P}_0=\nu,\, \mathbb{P}_1=\nu'} D_{\mathrm{KL}}\mathbb{P}_t Y_t. \tag{7.10}$$

Then (7.10) is equivalent to

$$\inf_{(\rho_t, v_t)} \mathbb{E}\left[\int_0^1 \frac{\|v_t\|^2}{2g_t^2} + \frac{g_t^2}{8}\|\nabla \log \rho_t\|^2 - \frac{1}{2}\langle f_t, \nabla \log \rho_t\rangle\,\mathrm{d}t\right] \tag{7.11}$$

where the infimum is taken all pairs (ρ_t, v_t) such that $\rho_0 = \nu, \rho_1 = \nu'$, ρ_t absolutely continuous, and

$$\partial_t \rho_t = -\nabla_x \cdot (\rho_t(f_t + v_t)). \tag{7.12}$$

The proof of Theorem 1, which we defer to Appendix A.2, is a straightforward extension of the argument in (Léonard, 2013; Chen et al., 2016;

Gentil et al., 2017) which establishes the equivalence when Y_t is a reversible Brownian motion, i.e., $f_t \equiv 0, g_t \equiv \sigma$, and Y_0 follows the Lebesgue measure.[1]

7.2.2.2 *The Gaussian Schrödinger Bridge Problem*

The central goal of our paper is to derive the closed-form solution of SBs when the marginal constraints are Gaussians $\xi \sim \mathcal{N}(\mu, \Sigma),\ \xi' \sim \mathcal{N}(\mu', \Sigma')$. Namely, we are interested in the following class of the SBs, termed Gaussian Schrödinger bridges:

$$\min_{\mathbb{P}_0=\xi,\ \mathbb{P}_1=\xi'} D_{\mathrm{KL}}\mathbb{P}_t Y_t. \tag{GSB}$$

To emphasize the dependence on the reference SDE, we will sometimes call (GSB) the Y_t-GSB.

TECHNICAL CHALLENGES; RELATED WORK. In order to analyze (GSB), we first notice that the objective in (7.11) becomes $\sigma^{-2}\,\mathbb{E}\left[\int_0^1 \frac{1}{2}\|v_t\|^2 + \frac{\sigma^4}{8}\|\nabla \log \rho_t\|^2\,\mathrm{d}t\right]$ for $\sigma\mathrm{W}_t$-GSBs. Up to a constant factor, this is simply (7.6), so Theorem 1 reduces to the well-known fact that $\sigma\mathrm{W}_t$-GSBs are a reformulation of the dynamic Gaussian OT (Léonard, 2013; Chen et al., 2016; Gentil et al., 2017).

At first sight, this might suggest that one can extend existing tools in Gaussian OT to analyze GSBs. Unfortunately, the major difficulty of tackling GSBs is that these existing tools are fundamentally insufficient for the generalized objective (7.11). To be more precise, there exist three prominent frameworks for studying Gaussian OT problems:

- **Convex analysis:** An extremely fruitful observation in the field is that many Gaussian OT instances can be reduced to a *convex* program, for which one can import various convex techniques such as KKT or fixed-point arguments. This is the case for static Gaussian OT (7.3), both when $\sigma = 0$ (Dowson and Landau, 1982; Olkin and Pukelsheim, 1982; Bhatia et al., 2019) and $\sigma > 0$ (Janati et al., 2020b). Furthermore, in the case of $\sigma = 0$, the solution to the dynamic formulation (7.6) can be recovered from the static one via a simple linear interpolation (McCann, 1997).
- **Ad hoc computations:** When $\sigma > 0$ in (7.6), the problem is no longer reducible to a convex program (Léonard, 2013; Chen et al., 2021b). In this case, the only technique we are aware of is the ad hoc approach of

1 The reversible Brownian motion is a technical construct to simplify the computations. For our purpose, one can think of $Y_0 \sim \xi$ instead of the Lebesgue measure, and our results still hold verbatim.

(Mallasto et al., 2021), which manages to find a closed form for (7.6) (and thus $\sigma\mathbb{W}_t$-GSBs) through a series of brute-force computations.

- **Control theory:** On a related note, in a series of papers, Chen et al. (2015, 2016, 2019) exploit the deep connection between $\sigma\mathbb{W}_t$-GSBs and control theory to study the *existence* and *uniqueness* of the solutions. Although a variety of new optimality conditions are derived in these works, they are all expressed in terms differential equations with coupled initial conditions, and it is unclear whether solving these differential equations is an easier task than (GSB) itself. In particular, no closed-form, even for $\sigma\mathbb{W}_t$-GSBs, can be found therein.

By Theorem 1, GSBs are more general than (7.6) and thus irreducible to convex programs, so there is no hope for the convex route. As for ad hoc computations, the time-dependent f_t and g_t terms in (7.11) present a serious obstruction for generalizing the approach of Mallasto et al. (2021) to Υ_t-GSBs when $f_t \neq 0$ or g_t is not constant; this is exemplified by the convoluted expressions in our Theorem 3, which hopefully will convince the reader that they are beyond any ad hoc guess. Finally, the control-theoretic view has so far fallen short of producing closed-form solutions even for $\sigma\mathbb{W}_t$-GSBs, so it is essentially irrelevant for our purpose.
To conclude, in order to find an analytic expression for general GSBs, we will need drastically different techniques.

OUR APPROACH To overcome the aforementioned challenges, in Section 7.2.3, we will first develop a principled framework for analyzing the closed-form expressions of $\sigma\mathbb{W}_t$-GSBs, i.e., (7.6). Unlike the ad hoc approach of Mallasto et al. (2021) which is very specific to Brownian motions, our analysis reveals the general role played by the *Lyapunov operator* (see (7.16)) on covariance matrices, thereby essentially reducing the solutions of GSBs to solving a matrix equation. This route is enabled via yet another equivalent formulation of (7.6), namely the action minimization problem on the *Bures-Wasserstein geometry*, which has recently emerged as a rich source for inspiring new computational methods (Chewi et al., 2020; Altschuler et al., 2021; Han et al., 2021). In Section 7.2.4, we show how the insight gained from our geometric framework in Section 7.2.3 can be easily adapted to GSBs with general reference processes, which ultimately leads to the full resolution of (GSB).

7.2.3 *The Bures-Wasserstein Geometry of $\sigma\mathbb{W}_t$-Gaussian Schrödinger Bridges*

This section illustrates the simple geometric intuition that underlies the somewhat technical proof of our main result (cf. Theorem 3). After briefly reviewing the action minimization problems on Euclidean spaces in Section 7.2.3.1, we present the main observation in Section 7.2.3.2: $\sigma\mathbb{W}_t$-GSBs are but action minimization problems on the Bures-Wasserstein manifolds, which can be tackled by following a standard routine in physics.

7.2.3.1 *A Brief Review on Action Minimization Problems*

Consider the following *action minimization* problem with fixed endpoints $x, x' \in \mathbb{R}^d$:

$$\min_{x(0)=x, x(1)=x'} \int_0^1 \frac{1}{2}\|\dot{x}(t)\|^2 - U(x(t))\,\mathrm{d}t, \tag{7.13}$$

where the minimum is taken over all piecewise smooth curves. A celebrated result in physics asserts that the optimal curve for (7.13) satisfies the *Euler-Lagrange* equation:

$$\ddot{x}(t) = -\nabla U(x(t)), \quad x(0) = x, \quad x(1) = x'. \tag{7.14}$$

In particular, when $U \equiv 0$, (7.14) reduces to $\ddot{x} \equiv 0$, i.e., $x(t)$ is a straight line connecting x and x'.
More generally, one can consider (7.13) on any *Riemannian manifold*, provided that the Euclidean norm $\|\cdot\|$ in (7.13) is replaced by the corresponding Riemannian norm. In this case, the Euler-Lagrange equation (7.14) still holds, with $\ddot{x}$ and ∇U replaced with their Riemannian counterparts (Villani, 2009).

7.2.3.2 *$\sigma\mathbb{W}_t$-GSBs as Action Minimization Problems*

We begin with the following simple observation. Based on the seminal work by Otto (2001), Gentil et al. (2020) show that SBs between two arbitrary measures can be formally understood as an action minimization problem of the form (7.13) on an *infinite*-dimensional manifold. Since we have restricted the measures in (GSB) to be Gaussian, and since Gaussian measures are uniquely determined by their means and covariances, Gentil et al. (2020) strongly suggests a *finite*-dimensional geometric interpretation of $\sigma\mathbb{W}_t$-GSBs. The main result in this section, Theorem 2 below, makes this link precise.

The proper geometry we need is the *Bures-Wasserstein manifold* (Takatsu, 2010; Bhatia et al., 2019) defined as follows. Consider the space of covariance matrices (i.e., symmetric positive definite matrices) of dimension d, which we denote by $\mathbb{S}^d_{++}$, and consider its natural tangent space as the space of symmetric matrices:

$$\mathcal{T}_\Sigma \mathbb{S}^d_{++} := \{U \in \mathbb{R}^{d\times d} : U^\top = U\}. \tag{7.15}$$

A notion that will play a pivotal role is the so-called *Lyapunov operator*: For any $\Sigma \in \mathbb{S}^d_{++}$ and $U \in \mathcal{T}_\Sigma \mathbb{S}^d_{++}$, we define $\mathcal{L}_\Sigma[U]$ to be the symmetric solution to the equation

$$A: \quad \Sigma A + A\Sigma = U. \tag{7.16}$$

It is shown in Takatsu (2010) that the Lyapunov operator defines a geometry on $\mathbb{S}^d_{++}$, known as the *Bures-Wasserstein geometry*: For any two tangent vectors $U, V \in \mathcal{T}_\Sigma \mathbb{S}^d_{++}$, the operation

$$\langle U, V\rangle_\Sigma := \frac{1}{2} \operatorname{tr} \mathcal{L}_\Sigma[U] V \tag{7.17}$$

satisfies all the axioms of the Riemannian metric; additional background on the Bures-Wasserstein geometry can be found in Appendix A.3.1.

We are now ready to state the main result of the section. Let $\|\cdot\|_\Sigma$ be the induced norm of $\langle\cdot,\cdot\rangle_\Sigma$. Fix $\sigma > 0$ and let $\mathbb{W}_t$ be a reversible Brownian motion. Consider the following special case of (GSB):

$$\min_{\mathbb{P}_0 = \mathcal{N}(0,\Sigma),\, \mathbb{P}_1 = \mathcal{N}(0,\Sigma')} D_{\mathrm{KL}} \mathbb{P}_t \sigma \mathbb{W}_t. \tag{7.18}$$

Then we have:

Theorem 2. *The minimizer of* (7.18) *(and hence* (7.6)*) coincides with the solution of the action minimization problem:*

$$\min_{\Sigma_0 = \Sigma, \Sigma_1 = \Sigma'} \int_0^1 \frac{1}{2} \|\dot\Sigma_t\|^2_{\Sigma_t} - \mathcal{U}_\sigma(\Sigma_t)\, \mathrm{d}t \tag{7.19}$$

where $\mathcal{U}_\sigma(\Sigma_t) := -\frac{\sigma^4}{8} \operatorname{tr} \Sigma_t^{-1}$ *and the minimum is taken over all piecewise smooth curves in* $\mathbb{S}^d_{++}$*. In particular, the minimizer of* (7.18) *solves the Euler-Lagrange equation in the Bures-Wasserstein geometry:*

$$\nabla_{\dot\Sigma_t} \dot\Sigma_t = -\operatorname{grad} \mathcal{U}_\sigma(\Sigma_t), \quad \Sigma_0 = \Sigma, \quad \Sigma_1 = \Sigma', \tag{7.20}$$

where $\nabla_{\dot\Sigma_t} \dot\Sigma_t$ *denotes the Riemannian acceleration and* grad *the Riemannian gradient in the Bures-Wasserstein sense.*

AN IMPORTANT IMPLICATION As alluded to in Section 7.2.2, the solution curve to (7.6) or (7.18) is not new; it is derived in Mallasto et al. (2021) via a strenuous and rather unenlightening calculation:

$$\Sigma_t := \bar{t}^2\Sigma + t^2\Sigma' + t\cdot\bar{t}\Big(C_\sigma + C_\sigma^\top + \sigma^2 I\Big). \tag{7.21}$$

Here, $\bar{t} := 1 - t$ and C_σ is defined in (7.4). However, the interpretation of (7.21) as the minimizer of (7.19) is new and suggests a principled avenue towards the closed-form solution of $\sigma\mathbb{W}_t$-GSBs: solve the Euler-Lagrange equation (7.20). Inspecting the formulas for $\nabla_{\dot{\Sigma}_t}\dot{\Sigma}_t$ and $\operatorname{grad}\mathcal{U}_\sigma(\Sigma_t)$ (see (A.3.5) and (A.3.6)), one can further reduce (7.20) to computing the Lyapunov operator $\mathcal{L}_{\Sigma_t}[\dot{\Sigma}_t]$, which presents the bottleneck in the proof of Theorem 2 as there is, in general, no closed form for the matrix equation (7.16). To this end, our main contribution is the following technical Lemma:

Lemma 1. *Define the matrix $\tilde{S}_t$ to be:*

$$\tilde{S}_t := t\Sigma' + \bar{t}C_\sigma - \bar{t}\Sigma - tC_\sigma^\top + \frac{\sigma^2}{2}(\bar{t} - t)I. \tag{7.22}$$

Then $\tilde{S}_t^\top\Sigma_t^{-1}$ is symmetric.

Armed with Lemma 1, it is straightforward to verify that $\mathcal{L}_{\Sigma_t}[\dot{\Sigma}_t] = \tilde{S}_t^\top\Sigma_t^{-1}$, i.e., $\tilde{S}_t^\top\Sigma_t^{-1}$ is symmetric and satisfies:

$$\tilde{S}_t^\top\Sigma_t^{-1}\cdot\Sigma_t^{-1} + \Sigma_t^{-1}\cdot\Sigma_t^{-1}\tilde{S}_t = \tilde{S}_t^\top + \tilde{S}_t = \dot{\Sigma}_t \tag{7.23}$$

which is more or less equivalent to the original Euler-Lagrange equation (7.20); we defer the details to Appendix A.3.2.
To conclude, in contrast to the purely technical approach of Mallasto et al. (2021), our Theorem 2 provides a geometric and conceptually clean solution for $\sigma\mathbb{W}_t$-GSBs: Compute the Lyapunov operator $\mathcal{L}_{\Sigma_t}[\dot{\Sigma}_t]$ via verifying the symmetry of the matrix in Lemma 1. *It turns out that this technique can be readily extended to general GSBs*, and therefore serves as the foundation for the proof of our main result; see Section 7.2.4.

REMARK. It is interesting to note that the matrix $\tilde{S}_t$ in (7.22) is itself *not* symmetric. Other consequences of Theorem 2 that might be of independent interest can be found in Appendix A.3.3. We also note that, when $\sigma = 0$, the solution to (7.19) is simply the Wasserstein geodesic between Gaussian measures, whose formula is well-known (Dowson and Landau, 1982; Takatsu,

2010). However, as explained in Section 7.2.2, the case of $\sigma > 0$ requires a completely different analysis since, unlike when $\sigma = 0$, it is not reducible to a convex program. This leads to the significantly more involved proofs of Theorem 2 and of (7.21) in Mallasto et al. (2021).

7.2.4 *Closed-Form Solutions of General Gaussian Schrödinger Bridges*

We now present the closed-form solutions of general GSBs.

7.2.4.1 *Linear Stochastic Differential Equations*

We need the following background knowledge on the linear SDE Y_t. Let $\tau_t := \exp\left(\int_0^t c_s \, \mathrm{d}s\right)$. Then the solution to (7.9) is (Platen and Bruti-Liberati, 2010):

$$Y_t = \tau_t \left(Y_0 + \int_0^t \tau_s^{-1} \alpha_s \, \mathrm{d}s + \int_0^t \tau_s^{-1} g_s \, \mathrm{d}W_s \right). \tag{7.24}$$

Another crucial fact in our analysis is that Y_t is a *Gaussian process given* Y_0, and is thus characterized by the first two moments. Using the independent increments of W_t and Itô's isometry (Protter, 2005), we compute:

$$\mathbb{E}[Y_t \mid Y_0] = \tau_t \left(Y_0 + \int_0^t \tau_s^{-1} \alpha_s \, \mathrm{d}s \right) =: \eta(t) \tag{7.25}$$

and, for any $t' \geq t$,

$$\begin{aligned} &\mathbb{E}\left[(Y_t - \eta(t)) \left(Y_{t'} - \eta\left(t'\right)\right)^\top \,\middle|\, Y_0 \right] \\ &\qquad = \left(\tau_t \tau_{t'} \int_0^t \tau_s^{-2} g_s^2 \, \mathrm{d}s \right) I =: \kappa\left(t, t'\right) I. \end{aligned} \tag{7.26}$$

7.2.4.2 *Main Result*

We now present the main result of our paper. With the important application of diffusion-based models in mind, we will not only derive solution curves as in (7.21) but also their SDE representations.
Let $\xi = \mathcal{N}(\mu_0, \Sigma_0)$ and $\xi' = \mathcal{N}(\mu_1, \Sigma_1)$ be two arbitrary Gaussian distributions in (GSB), and let D_σ, C_σ be as defined in (7.4).

SDE WITH $\alpha_t \equiv 0$	SETTING	$\kappa(t,t')$	$\sigma_\star^2$	r_t	$\bar{r}_t$	ρ_t	$\zeta(t)$
BM	$c_t{\equiv}0$ $g_t{\equiv}\omega{\in}\mathbb{R}^+$	$\omega^2 t$	ω^2	t	$1-t$	t	0
VESDE	$c_t{\equiv}0$ $g_t{=}\sqrt{q(t)}$	$q(t)$	$q(1)$	$\frac{q(t)}{q(1)}$	$1-\frac{q(t)}{q(1)}$	$\frac{q(t)}{q(1)}$	0
VPSDE	$-2c_t{=}g_t^2$	$\tau_{t'}\left(\tau_t^{-1}-\tau_t\right)$	$\tau_1^{-1}-\tau_1$	$\frac{\tau_t^{-1}-\tau_t}{\tau_1^{-1}-\tau_1}$	$\tau_1\left(\frac{\tau_t}{\tau_1}-\frac{\tau_t^{-1}-\tau_t}{\tau_1^{-1}-\tau_1}\right)$	$\frac{\tau_t^{-1}\left(\tau_t^{-1}-\tau_t\right)}{\tau_1^{-1}\left(\tau_1^{-1}-\tau_1\right)}$	0
SUB–VPSDE	$\frac{g_t^2}{-2c_t}{=}1{-}\tau_t^4$	$\tau_t\tau_{t'}\left(\tau_t^{-1}-\tau_t\right)^2$	$\tau_1\left(\tau_1^{-1}-\tau_1\right)^2$	$\frac{\tau_t}{\tau_1}\cdot\left(\frac{\tau_t^{-1}-\tau_t}{\tau_1^{-1}-\tau_1}\right)^2$	$\tau_t\left(1-\left(\frac{\tau_t^{-1}-\tau_t}{\tau_1^{-1}-\tau_1}\right)^2\right)$	$\left(\frac{\tau_t^{-1}-\tau_t}{\tau_1^{-1}-\tau_1}\right)^2$	0
SDE WITH $\alpha_t \not\equiv 0$	SETTING	$\kappa(t,t')$	$\sigma_\star^2$	r_t	$\bar{r}_t$	ρ_t	$\zeta(t)$
OU/VASICEK	$c_t{\equiv}{-}\lambda{\in}\mathbb{R}$ $\alpha_t{\equiv}\mathbf{v}{\in}\mathbb{R}^d$ $g_t{\equiv}\omega{\in}\mathbb{R}^+$	$\frac{\omega^2 e^{-\lambda t'}\sinh\lambda t}{\lambda}$	$\frac{\omega^2\sinh\lambda}{\lambda}$	$\frac{\sinh\lambda t}{\sinh\lambda}$	$\sinh\lambda t\coth\lambda t$ $-\sinh\lambda t\coth\lambda$	$e^{-\lambda(1-t)}$ $\cdot\frac{\sinh\lambda t}{\sinh\lambda}$	$\frac{\mathbf{v}}{\lambda}\left(1-e^{-\lambda t}\right)$
α_t-BDT	$c_t{\equiv}0$ $g_t{\equiv}\omega{\in}\mathbb{R}^+$	$\omega^2 t$	$\omega^2 1$	t	$1-t$	t	$\int_0^t \alpha_s\,\mathrm{d}s$

Table 7.1: Examples of reference SDEs and the corresponding solutions of GSBs. All relevant functions in the Table are either introduced in Section 7.2.4.1 or (7.27).

Theorem 3. *Denote by $\mathbb{P}_t$ the solution to Gaussian Schrödinger bridges (GSB). Set*

$$
\begin{aligned}
&r_t := \frac{\kappa(t,1)}{\kappa(1,1)}, \quad \bar{r}_t := \tau_t - r_t\tau_1, \quad \sigma_\star := \sqrt{\tau_1^{-1}\kappa(1,1)},\\
&\zeta(t) := \tau_t\int_0^t \tau_s^{-1}\alpha_s\,\mathrm{d}s, \rho_t := \frac{\int_0^t \tau_s^{-2}g_s^2\,\mathrm{d}s}{\int_0^1 \tau_s^{-2}g_s^2\,\mathrm{d}s},\\
&P_t := \dot{r}_t(r_t\Sigma_1 + \bar{r}_t C_{\sigma_\star}), \quad Q_t := -\dot{\bar{r}}_t(\bar{r}_t\Sigma_0 + r_t C_{\sigma_\star}),\\
&S_t := P_t - Q_t^\top + \Big[c_t\kappa(t,t)(1-\rho_t) - g_t^2\rho_t\Big]I. \qquad (7.27)
\end{aligned}
$$

Then the following holds:

1. *The solution $\mathbb{P}_t$ is a Markov Gaussian process whose marginal variable $X_t \sim \mathcal{N}(\mu_t,\Sigma_t)$, where*

$$\mu_t := \bar{r}_t\mu_0 + r_t\mu_1 + \zeta(t) - r_t\zeta(1), \qquad (7.28)$$

$$\Sigma_t := \bar{r}_t^2\Sigma_0 + r_t^2\Sigma_1 + r_t\bar{r}_t\left(C_{\sigma_\star} + C_{\sigma_\star}^\top\right) + \kappa(t,t)(1-\rho_t)I. \qquad (7.29)$$

2. *X_t admits a closed-form representation as the SDE:*

$$\mathrm{d}X_t = f_{\mathcal{N}}(t,X_t)\,\mathrm{d}t + g_t\,\mathrm{d}\mathbf{W}_t \qquad (7.30)$$

where

$$f_{\mathcal{N}}(t,x) := S_t^\top\Sigma_t^{-1}(x-\mu_t) + \dot{\mu}_t. \qquad (7.31)$$

Moreover, the matrix $S_t^\top\Sigma_t^{-1}$ is symmetric.

As in Theorem 2, the key step in the proof of Theorem 3 is to recognize the symmetry of the matrix $S_t^\top \Sigma_t^{-1}$ where S_t, defined in (7.27), simply becomes the $\tilde{S}_t$ in Lemma 1 (up to an additive factor of $\frac{\sigma^2 \bar{t}}{2} I$) for $\sigma \mathbb{W}_t$-GSBs. Although this can be directly verified via generalizing Lemma 1, the computation becomes quite tedious, so our proof of Theorem 3 will follow a slightly different route. In any case, given the symmetry of $S_t^\top \Sigma_t^{-1}$, the proof simply boils down to a series of straightforward calculations; see Appendix A.4.

CLOSED FORMS FOR CONDITIONAL DISTRIBUTIONS. In many practical applications such as generative modeling, a requirement to employ the SDE representation of GSBs in (7.30) is that its *conditional distributions* given the initial points can be computed efficiently. As an immediate corollary of Theorem 3, we obtain the following closed-form expressions for these conditional distributions.

Corollary 1. *Let $X_t \sim \mathbb{P}_t$ be the the solution to (GSB). Then the conditional distribution of X_t given end points has a simple solution: $X_t | X_0 = x_0 \sim \mathcal{N}\Big(\mu_{t|0}, \Sigma_{t|0}\Big)$, where*

$$\mu_{t|0} = \bar{r}_t x_0 + r_t \Big(\mu_1 + C_{\sigma_\star}^\top \Sigma_0^{-1} (x_0 - \mu_0)\Big) + \zeta(t) - r_t \zeta(1), \tag{7.32}$$

$$\Sigma_{t|0} = r_t^2 \Big(\Sigma_1 - C_\sigma^\top \Sigma_0^{-1} C_\sigma\Big) + \kappa(t,t)(1-\rho_t) I. \tag{7.33}$$

Similarly, $X_t | X_1 = x_1 \sim \mathcal{N}\Big(\mu_{t|1}, \Sigma_{t|1}\Big)$, where

$$\mu_{t|1} = r_t x_1 + \bar{r}_t \Big(\mu_0 + C_{\sigma_\star} \Sigma_1^{-1} (x_1 - \mu_1)\Big) + \zeta(t) - r_t \zeta(1), \tag{7.34}$$

$$\Sigma_{t|1} = \bar{r}_t^2 \Big(\Sigma_0 - C_\sigma \Sigma_1^{-1} C_\sigma^\top\Big) + \kappa(t,t)(1-\rho_t) I. \tag{7.35}$$

Examples of GSBs. Our framework captures most popular reference SDEs in the machine learning literature as well as other mathematical models in financial engineering; see Table 7.1. A non-exhaustive list includes:

- The basic Brownian motion (BM) and the Ornstein-Uhlenbeck (OU) processes, both widely adopted as the reference process for SB-based models (De Bortoli et al., 2021a,b; Lavenant et al., 2021; Vargas et al., 2021; Wang et al., 2021). We also remark that, even though (7.29) is known for BM (Mallasto et al., 2021), what is crucial in these applications is the SDE presentation (7.30), which is new even for BM.
- The variance exploding SDEs (VESDEs), which underlies the training of score matching with Langevin dynamics for diffusion-based generative modeling (Huang et al., 2021b; Song and Ermon, 2019; Song et al., 2021).

- The variance preserving SDEs (VPSDEs), which can be seen as the continuous limit of denoising diffusion probabilistic models (Ho et al., 2020; Sohl-Dickstein et al., 2015; Song et al., 2021), another important class of algorithms for diffusion-based generative modeling.
- The *sub-VPSDEs* proposed by (Song et al., 2021), which are motivated by reducing the variance of VPSDEs.
- Several important SDEs in financial engineering, such as the *Vasicek model* (which generalizes OU processes) and the *constant volatility* α_t*-Black-Derman-Toy (BDT) model* (Platen and Bruti-Liberati, 2010).

7.2.5 GSBflow: ...

Building on the closed-form solutions in **??**, we present an end-to-end learning paradigm that takes two marginal distributions $\hat{\mathbb{P}}_0, \hat{\mathbb{P}}_1$ to output the reconstruction of the underlying stochastic dynamics $\mathbb{P}_t$. Because our framework relies on GSBs, we call our algorithm the GSBflow.

Step 1: Moment estimates and GSB initialization. We first compute the means μ_0, μ_1 and covariances Σ_0, Σ_1 of the input distributions, and plug them into (7.31) and (7.32)-(7.35). Note that these computations are done only *once* for every dataset, and can be reused for all subsequent training.

Step 2: Forward and backward pretraining. Denoting by $\mathbb{Q}_t$ the measure of $f_{\mathcal{N}}\,\mathrm{d}t + g_t\,\mathrm{d}\mathbf{W}_t$ in (7.31), we propose to minimize the objective

$$\min_{\mathbb{P}_0=\hat{\mathbb{P}}_0,\ \mathbb{P}_1=\hat{\mathbb{P}}_1} D_{\mathrm{KL}}\mathbb{P}_t\mathbb{Q}_t. \tag{7.36}$$

Following the framework of Chen et al. (2022a), we see that the optimal solution to (7.36) is given by two SDEs of the form:

$$\mathrm{d}X_t = (f_{\mathcal{N}} + g_t Z_t)\,\mathrm{d}t + g_t\,\mathrm{d}\mathbf{W}_t, \quad X_0 \sim \hat{\mathbb{P}}_0, \tag{7.37a}$$
$$\mathrm{d}X_t = (f_{\mathcal{N}} - g_t \hat{Z}_t)\,\mathrm{d}t + g_t\,\mathrm{d}\mathbf{W}_t, \quad X_1 \sim \hat{\mathbb{P}}_1, \tag{7.37b}$$

where (7.37b) runs backward in time. After parameterizing Z_t and $\hat{Z}_t$ by two neural networks $Z_t^\theta(x), \hat{Z}_t^\phi(x)$ with parameters θ, ϕ, the corresponding negative likelihood in Section 7.1 becomes

$$\ell(x_0;\phi) = \int_0^1 \mathbb{E}_{(7.37a)}\Big[\frac{1}{2}\|\hat{Z}_t^\phi\|^2 + g\nabla_x\cdot\hat{Z}_t^\phi + \langle Z_t^\theta, \hat{Z}_t^\phi\rangle\,\mathrm{d}t\Big|X_0 = x_0\Big], \tag{7.38a}$$
$$\ell(x_1;\theta) = \int_0^1 \mathbb{E}_{(7.37b)}\Big[\frac{1}{2}\|Z_t^\theta\|^2 + g\nabla_x\cdot Z_t^\theta + \langle \hat{Z}_t^\phi, Z_t^\theta\rangle\,\mathrm{d}t\Big|X_1 = x_1\Big]. \tag{7.38b}$$

Following existing work on training SB-based objectives (Chen et al., 2022a; De Bortoli et al., 2021b; Vargas et al., 2021), we propose to initialize $\tilde{\theta}_0, \tilde{\phi}_0$ such that $Z_t^{\tilde{\theta}_0}(x), \hat{Z}_t^{\tilde{\phi}_0}(x) \equiv 0$, which can be easily achieved by zeroing out the last layer of the corresponding neural networks. In this case, estimating the conditional expectations in both (7.38a)-(7.38b) reduces to simulating (7.31) *conditioned* on the given start or end data points. Thanks to our closed-form expressions, this can be easily achieved by drawing Gaussian variables with mean and covariance prescribed in (7.37a)-(7.37b). The pretraining procedure is summarized in Algorithm 2.

Step 3: Alternating minimization. After the pretraining phase, we switch to minimizing (7.38a)-(7.38b) with general drifts in (7.37a)-(7.37b). We carry out this step in an alternating fashion: Since the bottleneck of our framework is to simulate the trajectories of SDEs, we perform several gradient updates for one parameter before drawing another batch of samples. See Algorithm 3 for a summary, and Fig. 5.1 for an illustration.

Algorithm 2 Forward and Backward Pretraining

Input: Marginal distributions $\hat{\mathbb{P}}_0, \hat{\mathbb{P}}_1$, initial parameters $\tilde{\theta}_0, \tilde{\phi}_0$ such that $Z_t^{\tilde{\theta}_0}(\cdot\,) = \hat{Z}_t^{\tilde{\phi}_0}(\cdot\,) \equiv 0$, iteration counts K_θ, K_ϕ, learning rates $\gamma_\theta, \gamma_\phi$
Output: Pretrained parameters θ_0, ϕ_0
Initialize $\theta_0 \leftarrow \tilde{\theta}_0$, $\phi_0 \leftarrow \tilde{\phi}_0$
for $k = 1$ **to** K_ϕ **do**
  Sample X_t from (7.32)-(7.33) with $x_0 \sim \hat{\mathbb{P}}_0$
  Compute $\ell(x_0; \phi)$ via (7.38a)
  Update $\phi_0 \leftarrow \phi_0 - \gamma_\phi \nabla \ell(x_0; \phi_0)$
for $k = 1$ **to** K_θ **do**
  Sample X_t from (7.34)-(7.35) with $x_1 \sim \hat{\mathbb{P}}_1$
  Compute $\ell(x_1; \theta)$ via (7.38b)
  Update $\theta_0 \leftarrow \theta_0 - \gamma_\theta \nabla \ell(x_1; \theta_0)$

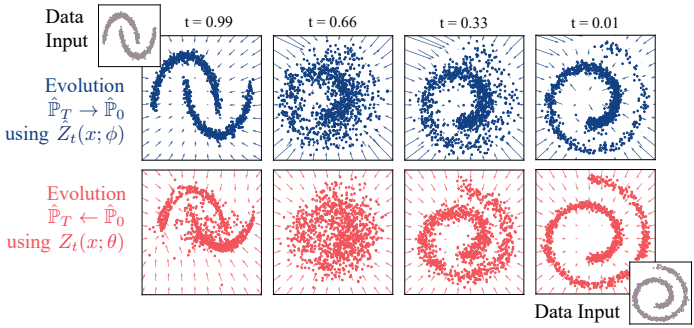


Figure 7.2: Illustration of the time-dependent drifts learned by GSBFLOW with VE SDE for two toy marginal distributions. *Top.* Evolution of $\hat{\mathbb{P}}_1$ (moons) $\to \hat{\mathbb{P}}_0$ (spiral) via backward policy $\hat{Z}_t^{\phi}(x)$. *Bottom.* Evolution of $\hat{\mathbb{P}}_0$ (spiral) $\to \hat{\mathbb{P}}_1$ (moons) via forward policy $Z_t^{\theta}(x)$.

Algorithm 3 GSBFLOW

Input: Marginal distributions $\hat{\mathbb{P}}_0, \hat{\mathbb{P}}_1$, pretrained parameters θ_0, ϕ_0, caching frequency M, iteration counts $K_{\text{in}}, K_{\text{out}}$, learning rates $\gamma_\theta, \gamma_\phi$

Output: Optimal forward and backward drifts $Z_t(\cdot), \hat{Z}_t(\cdot)$ for (7.36)

Initialize $\theta \leftarrow \theta_0$, $\phi \leftarrow \phi_0$.
for $k = 1$ **to** K_{out} **do**
  for $j = 1$ **to** K_{in} **do**
    if $j \mod M = 0$ **then**
      Simulate (7.37a) with $x_0 \sim \hat{\mathbb{P}}_0$
    Compute $\ell(x_0;\phi)$ via (7.38a)
    Update $\phi \leftarrow \phi - \gamma_\phi \nabla \ell(x_0;\phi)$
  for $j = 1$ **to** K_{in} **do**
    if $j \mod M = 0$ **then**
      Simulate (7.37b) with $x_1 \sim \hat{\mathbb{P}}_1$
    Compute $\ell(x_1;\theta)$ via (7.38b)
    Update $\theta \leftarrow \theta - \gamma_\theta \nabla \ell(x_1;\theta)$

7.2.6 *Empirical Evaluation*

The purpose of our experiments is to demonstrate that, by leveraging moment information, GSBFLOW is significantly more stable compared to other SB-based objectives, especially when moving beyond the *generative* setting

Method	Tasks	
	Wasserstein Loss $W_\varepsilon \downarrow$	
	Moon et al. (2019)	Schiebinger et al. (2019)
Song et al. (2021)		
VESDE	20.83 ± 0.18	40.81 ± 0.42
sub-VPSDE	$\mathbf{19.96 \pm 0.58}$	48.15 ± 3.38
GSBFLOW (*ours*)		
VESDE	25.18 ± 0.10	$\mathbf{27.85 \pm 0.68}$

Table 7.2: Evaluation of predictive performance w.r.t. the entropy-regularized Wasserstein distance W_ε (Cuturi, 2013) of GSBFLOW and baselines on generating different single-cell datasets (using 3 runs).

where $\hat{\mathbb{P}}_1$ is a simple Gaussian. Indeed, while performing competitively in the generative setting ($\mathcal{N}_0 \to \hat{\mathbb{P}}_1$), our method *outperforms* when modeling the evolution of two complex distributions ($\hat{\mathbb{P}}_0 \to \hat{\mathbb{P}}_1$), the most general and ambitious setting to estimate a bridge. This is demonstrated on synthetic data as well as a task from molecular biology concerned with modeling the dynamics of cellular systems, i.e., single-cell genomics (Macosko et al., 2015; Frangieh et al., 2021; Kulkarni et al., 2019).

7.2.6.1 *Synthetic Dynamics*

Before conducting the single-cell genomics experiments, we first test GSBFLOW on a synthetic setting. Our first task involves recovering the stochastic evolution of two-dimensional synthetic data containing two interleaving half circles ($\hat{\mathbb{P}}_1$) into a spiral ($\hat{\mathbb{P}}_0$). Fig. 7.2 shows the trajectories learned by GSBFLOW based on the VESDE (see Table 7.1 and **??**).
While it is sufficient to parameterize only a single policy ($\hat{Z}_t^\phi(x)$) in generative modeling, the task of learning to evolve $\hat{\mathbb{P}}_0$ into $\hat{\mathbb{P}}_1$ requires one to recover *both* vector fields $\hat{Z}_t^\phi(x)$ and $Z_t^\theta(x)$. As demonstrated in Fig. 7.2, GSBFLOW is able to successfully learn both policies $Z_t^\theta(x)$ and $\hat{Z}_t^\phi(x)$ and reliably recovers the corresponding targets of the forward and backward evolution. While initializing the reference process through the closed-form SB between the Gaussian approximations of both synthetic datasets provides good results, the power of GSBFLOW becomes evident in more complex applications which we tackle next.

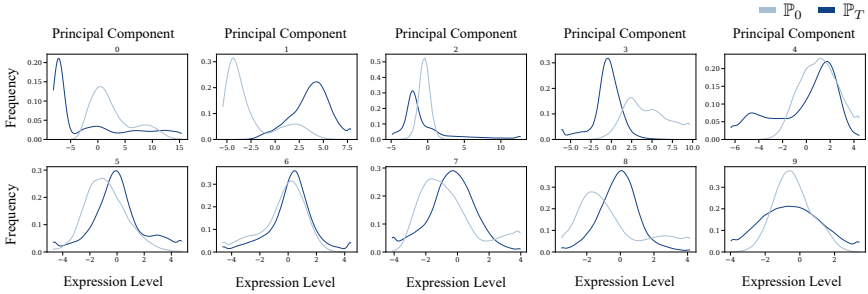


Figure 7.3: The expression levels of the first 10 principal components from the dataset by Schiebinger et al. (2019).

7.2.6.2 *Single-Cell Dynamics*

Modern single-cell profiling technologies are able to provide rich feature representations (e.g., gene expression) of *individual* cells at any development state. A crucial issue that arises with such profiling methods is their destructive nature: Measuring a cell requires destroying it and thus a cell cannot be measured twice. As a result, independent samples are collected at each snapshot, with no access to ground-truth single-cell trajectories throughout time, resulting in challenging, *unaligned*, datasets. Recovering cellular dynamics from such unaligned snapshots, i.e., $\hat{\mathbb{P}}_0$ to $\hat{\mathbb{P}}_1$, has, however, extremely important scientific and biomedical relevance (Kulkarni et al., 2019). For example, it determines our understanding on how and why tumor cells evade cancer therapies (Frangieh et al., 2021) or unveils mechanisms of cell differentiation and development (Schiebinger et al., 2019). Following related work, in particular previous methods based on optimal transport (Schiebinger et al., 2019; Bunne et al., 2023b, 2022a; Tong et al., 2020), the task is thus to learn the stochastic process that described the evolution of single cells from $\hat{\mathbb{P}}_0$ to $\hat{\mathbb{P}}_1$.

EXPERIMENTAL SETUP.

SINGLE-CELL GENOMICS VIA SBS. Let us consider the evolution of a gene, for which we can collect the empirical distributions $\hat{\mathbb{P}}_0, \hat{\mathbb{P}}_1$ of its expression levels at the times $t = 0, 1$ (Schiebinger et al., 2019; Moon et al., 2019). Our goal is to two-fold:

1. To solve the **generative modeling** problem, i.e., to generate $\hat{\mathbb{P}}_0$ or $\hat{\mathbb{P}}_1$ from a standard Gaussian noise, and

2. to **evolve** $\mathbb{P}_0 \to \mathbb{P}_1$ or $\mathbb{P}_1 \to \mathbb{P}_0$, i.e., to recover a stochastic process $\mathbb{P}_t$ satisfying $\mathbb{P}_0 = \hat{\mathbb{P}}_0, \mathbb{P}_1 = \hat{\mathbb{P}}_1$.

Although there are numerous algorithms for generative modeling, to our knowledge, the only framework that can simultaneously solve both tasks is the SB-based scheme recently proposed in (Chen et al., 2022a). In order to apply this framework, one has to choose a prior process Y_t, which is taken by the authors to be the high-performing VESDE and sub-VPSDE. These SB-based methods, as well as several standard generative modeling algorithms (Ho et al., 2020; Sohl-Dickstein et al., 2015; Song et al., 2021; Huang et al., 2021b; Song and Ermon, 2019; Song et al., 2021) for the first task, constitute strong baselines for our experiments.

OUR CHOICE OF Y_t; THE GSBFLOW. Instead of directly diving into the numerical solution of SBs as in Chen et al. (2022a), we first empirically verify that the distributions $\hat{\mathbb{P}}_0, \hat{\mathbb{P}}_1$ in single-cell genomics are typically close to *non-standard* Gaussian distributions: See Fig. 7.3 for the canonical dataset (Schiebinger et al., 2019).

Since the solutions of SBs are Lipschitz in terms of $\hat{\mathbb{P}}_0, \hat{\mathbb{P}}_1$ (Carlier et al., 2022), a reasonable approximation to the original SB objective is to replace $\hat{\mathbb{P}}_0, \hat{\mathbb{P}}_1$ by Gaussians with matching moments. This results in a GSB problem which can be solved in closed form by our Theorem 3. Intuitively, if we denote an existing prior process by Y_t and the solution of its corresponding GSB by X_t, then X_t presents a more appealing prior process than Y_t since it carries the moment information of $\hat{\mathbb{P}}_0$ and $\hat{\mathbb{P}}_1$, whereas Y_t is completely data-oblivious.

Motivated by these observations, we propose a simple modification of the framework in Chen et al. (2022a): Replace the prior process Y_t by its GSB approximation and keep everything else the same. The resulting scheme, which we term the GSBFLOW, learns a pair of forward $Z_t^\theta(x)$ and backward parametrized drifts $\hat{Z}_t^\phi(x)$ that progressively transport samples from $\hat{\mathbb{P}}_0 \to \hat{\mathbb{P}}_1$ and $\hat{\mathbb{P}}_1 \to \hat{\mathbb{P}}_0$, respectively. The full algorithm is presented in Algorithm 3 for completeness.

RESULTS. We investigate the ability of GSBFLOW to generate cell populations $\hat{\mathbb{P}}_1$ from noise $\mathcal{N}_0$ ($\mathcal{N}_0 \to \hat{\mathbb{P}}_1$, Fig. 7.4a, b) on the the canonical datasets (Moon et al., 2019; Schiebinger et al., 2019); as well as to predict the dynamics of single-cell genomics ($\hat{\mathbb{P}}_0 \to \hat{\mathbb{P}}_1$, Fig. 7.4c) (Moon et al., 2019), i.e., the inference of cell populations $\hat{\mathbb{P}}_1$ resulting from the developmental process of an initial cell population $\hat{\mathbb{P}}_0$, with the goal of learning individual

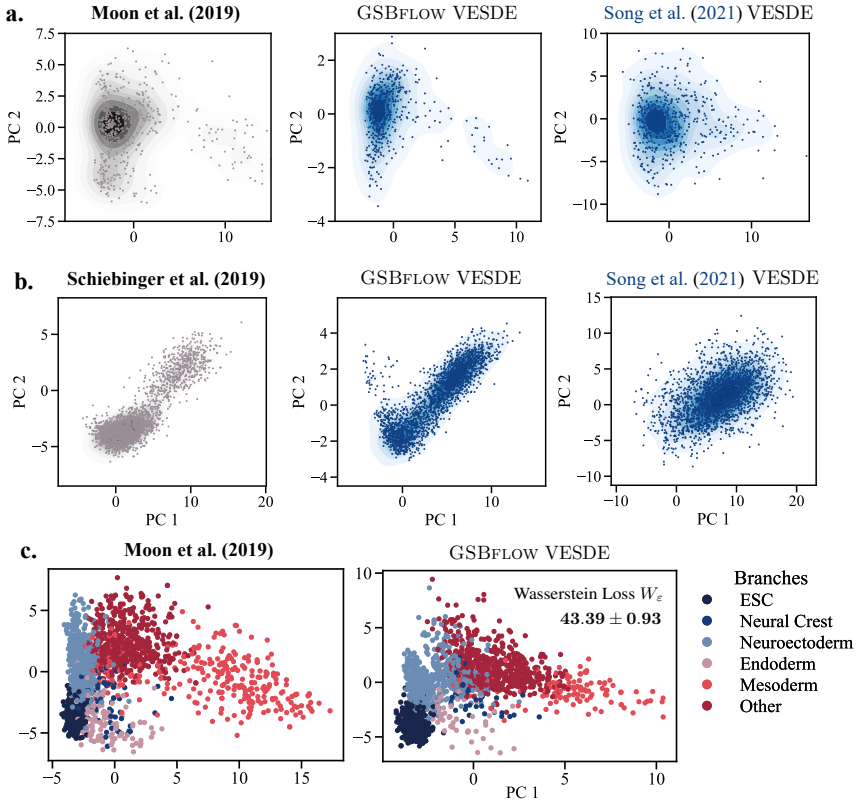


Figure 7.4: a.-b. Visual evaluation of the ability of our method to model the **generation** of data from **a.** Moon et al. (2019) and **b.** Schiebinger et al. (2019). Density plots are visualized in 2D PCA space and show generated data points using either GSBFLOW (our method) or the procedure in Song et al. (2021). **c.** Evaluation of GSBFLOW's ability to model the entire **evolution** of a developmental process of Moon et al. (2019), visualized by the data and GSBFLOW predictions colored by the lineage branch class.

dynamics, identify ancestor and descendant cells. The evaluation is conducted on the first 20 or 30 components of the PCA space of the > 1500 highly differentiable genes.
We evaluate the quality of the generated cellular states through the entropy-regularized Wasserstein distance W_ε (see Table 7.2) and by visualizing the first two principal components (PC), see Fig. 7.4a, b. GSBFLOW performs competitively on reconstructing embryoid body differentiation landscapes (Moon et al., 2019), and outperforms score-based generative models baselines on the iPSC reprogramming task (Schiebinger et al., 2019) as quantified by W_ε between data and predictions. Further, we analyze GSBFLOW's ability to predict the temporal evolution of embryoid body differentiation (Moon et al., 2019), where cells measured at day 1 to 3 serve as samples of $\hat{\mathbb{P}}_0$, while $\hat{\mathbb{P}}_1$ is constructed from samples between day 12 to 27. As no ground truth trajectories are available in the data, we compare the predicted evolution to the data and compare how well the heterogeneity of lineage (Fig. 7.4c, upper panel) is captured. Fig. 7.4c (lower panel) thereby closely resembles the data (see W_ε in Fig. 7.4c) and thus demonstrate GSBFLOW's ability to learn cell differentiation into various lineages and to capture biological heterogeneity on a more macroscopic level.

7.2.7 *Discussion*

We derive closed-form solutions of GSBs, an important class of dynamic OT problems. Our technique originates from a deep connection between Gaussian OT and the Bures-Wasserstein geometry, which we generalize to the case of general SB problems. Numerically, we demonstrate that our new closed forms inspire a simple modification of existing SB-based numerical schemes, which can however lead to significantly improved performance.
Limitation of our framework. In a broader context, we hope our results can serve as the inspiration for more learning algorithms, much like how existing closed-form solutions of Gaussian OT problems have contributed to the machine learning community. We thus acknowledge a severe limitation of our closed-form solutions: These formulas require matrix inversions, which might face scalability issues for high-dimensional data. In addition, existing matrix inversion algorithms are typically extremely sensitive to the condition number, and thus our formulas are not as useful for ill-conditioned data. Lifting these constraints to facilitate further applications, such as to image datasets, is an important future work.

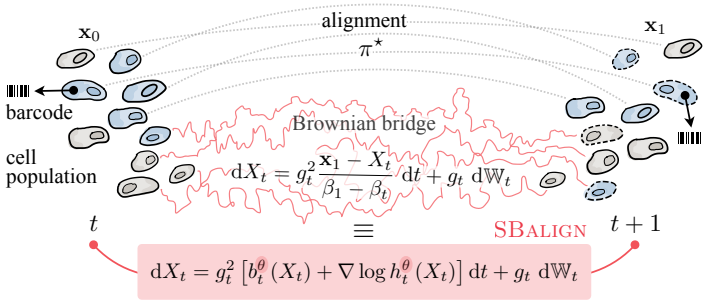


Figure 7.5: Overview of SBALIGN: In biological tasks such as protein docking, one is naturally provided with *aligned* data in the form of unbound and bound structures of participating proteins. Our goal is to therefore recover a stochastic trajectory from $\mathbf{x}_0$ to $\mathbf{x}_1$. To achieve this, we connect the characterization of an SDE conditioned on $\mathbf{x}_0$ and $\mathbf{x}_1$ (utilizing the Doob's *h-transform*) with that of a Brownian bridge between $\mathbf{x}_0$ and $\mathbf{x}_1$ (classical Schrödinger bridge theory). We show that this leads to a simpler training procedure with lower variance and strong empirical results.

7.3 LEARNING DIFFUSION SCHRÖDINGER BRIDGES FROM SPARSE TRAJECTORIES

Interpolation, the task of transforming one given distribution into another, lies at the heart of many modern machine learning applications such as single-cell genomics (Tong et al., 2020; Schiebinger et al., 2019; Bunne et al., 2022a), meteorology (Fisher et al., 2009), and robotics (Chen et al., 2021a). To this end, diffusion Schrödinger bridges (De Bortoli et al., 2021b; Chen et al., 2022a; Vargas et al., 2021; Liu et al., 2022) have recently emerged as a powerful paradigm due to their ability to generalize prior deep diffusion-based models, notably score matching with Langevin dynamics (Song and Ermon, 2019; Song et al., 2021) and denoising diffusion probabilistic models (Ho et al., 2020), which have achieved the state-of-the-art on many generative modeling problems.

Despite the wide success, a significant limitation of existing frameworks for solving DSBs is that they fail to capture the *alignment* of data: If $\hat{\mathbb{P}}_0, \hat{\mathbb{P}}_1$ are two (empirical) distributions between which we wish to interpolate, then a tacit assumption in the literature is that the dependence of $\hat{\mathbb{P}}_0$ and $\hat{\mathbb{P}}_1$ is unknown and somehow has to be recovered. Such an assumption, however, ignores important scenarios where the data is *aligned*, meaning that the samples from $\hat{\mathbb{P}}_0$ and $\hat{\mathbb{P}}_1$ naturally come in pairs $(\mathbf{x}_0^i, \mathbf{x}_1^i)_i^N$, which is common in many biological phenomena. Proteins, for instance, undergo

conformational changes upon interactions with other biomolecules (protein docking, see Fig. 7.5). The goal is to model conformational changes by recovering a (stochastic) trajectory $\mathbf{x}_t$ based on the positions observed at two-time points $(\mathbf{x}_0, \mathbf{x}_1)$. Failing to incorporate this alignment would mean that we completely ignore information on the correspondence between the initial and final points of the molecules, resulting in a much harder problem than necessary. Beyond, the recent use of SBs has been motivated by an important task in molecular biology: Cells change their molecular profile throughout developmental processes (Schiebinger et al., 2019; Bunne et al., 2022b) or in response to perturbations such as cancer drugs (Lotfollahi et al., 2019; Bunne et al., 2023b). As most measurement technologies are destructive assays, i.e., the same cell cannot be observed twice nor fully profiled over time, these methods aim at reconstructing cell dynamics from *unpaired* snapshots. Recent developments in molecular biology, however, aim at overcoming this technological limitation. For example, Chen et al. (2022b) propose a transcriptome profiling approach that preserves cell viability. Weinreb et al. (2020) capture cell differentiation processes by clonally connecting cells and their progenitors through barcodes (see illustrative Figure in Supplement). Motivated by these observations, the goal of this paper is to propose a novel algorithmic framework for solving DSBs with (partially) *aligned* data. Our approach is in stark contrast to existing works which, due to the lack of data alignment, all rely on some variants of IPF (Fortet, 1940; Kullback, 1968) and are thus prone to numerical instability. On the other hand, via a combination of the original theory of Schrödinger bridges (Schrödinger, 1931; Léonard, 2013) and the key notion of Doob's *h-transform* (Doob, 1984; Rogers and Williams, 2000), we design a novel loss function that completely bypasses the IPF procedure and can be trained with much lower variance. To summarize, we make the following contributions:

- To our best knowledge, we consider, for the first time, the problem of interpolation with *aligned* data. We rigorously formulate the problem in the DSB framework.
- Based on the theory of Schrödinger bridges and h-transform, we derive a new loss function that, unlike prior work on DSBs, does not require an IPF-like procedure to train. We also propose principled regularization schemes to further stabilize training.
- We describe how interpolating aligned data can provide better reference processes for use in classical DSBs, paving the way to hybrid aligned/non-aligned SBs.

- We evaluate our proposed framework on both synthetic and real data. For experiments utilizing real data, we consider two tasks where such aligned data is naturally available. The first is the task of developmental processes in single-cell biology, and the second is (*rigid*) protein docking, where the goal is to predict the 3D structure of the bound complex formed by two proteins, given their unbound 3D structures. Our method demonstrates a considerable improvement over prior methods across various metrics, thereby substantiating the importance of taking the data alignment into account.

RELATED WORK. Solving DSBs is a subject of significant interest in recent years and has flourished in a number of different algorithms (De Bortoli et al., 2021b; Chen et al., 2022a; Vargas et al., 2021; Bunne et al., 2023a; Liu et al., 2022). However, all these previous approaches focus on *unaligned* data, and therefore the methodologies all rely on IPF and are hence drastically different from ours. In the experiments, we will demonstrate the importance of taking the alignment of data into consideration by comparing our method to these baselines.

An important ingredient in our theory is Doob's h-transform, which has recently also been utilized by Liu et al. (2023) to solve the problem of constrained diffusion. However, their fundamental motivation is different from ours. Liu et al. (2023) focus on learning the drift of the diffusion model and the h-transform *together*, whereas ours is to read off the drift *from* the h-transform with the help of *aligned data*. Consequently, there is no overlap between the two algorithms and their intended applications.

To the best of our knowledge, the concurrent work of Tong et al. (2023) is the only existing framework that can tackle aligned data, which, however, is not their original motivation. In the context of solving DSBs, their algorithm can be seen as learning a vector field that generates the correct *marginal* probability (cf. Tong et al., 2023, Proposition 4.3). Importantly, this is different from our aim of finding the *pathwise* optimal solution of DSBs: If $(\mathbf{x}^i_{0,\text{test}})_{i=1}^m$ is a test data set for which we wish to predict their destinations, then the framework of Tong et al. (2023) can only ensure that the marginal distribution $(\mathbf{x}^i_{1,\text{test}})_{i=1}^m$ is correct, whereas ours is capable of predicting that $\mathbf{x}^i_{1,\text{test}}$ is precisely the destination of $\mathbf{x}^i_{0,\text{test}}$ for each i. This latter property is highly desirable in tasks like ML-accelerated protein docking.

PROBLEM FORMULATION. Suppose that we are given access to i.i.d. *aligned* data $(\mathbf{x}^i_0, \mathbf{x}^i_1)_{i=1}^N$, where the marginal distribution of $\mathbf{x}^i_0$'s is $\hat{\mathbb{P}}_0$ and of

$\mathbf{x}_1^i$'s is $\hat{\mathbb{P}}_1$. Typically, we view $\hat{\mathbb{P}}_0$ as the empirical marginal distribution of a stochastic process observed at time $t = 0$, and likewise $\hat{\mathbb{P}}_1$ the empirical marginal observed at $t = 1$. The goal is to reconstruct the stochastic process $\mathbb{P}_t$ based on $(\mathbf{x}_0^i, \mathbf{x}_1^i)_{i=1}^N$, i.e., to *interpolate* between $\hat{\mathbb{P}}_0$ and $\hat{\mathbb{P}}_1$.
Such a task is ubiquitous in biological applications. For instance, understanding how proteins dock to other biomolecules is of significant interest in biology and has become a topic of intense study in recent years (Ganea et al., 2022; Tsaban et al., 2022; Corso et al., 2023). In the protein docking task, $\mathbf{x}_0^i$ represents the 3D structures of the unbound proteins, while $\mathbf{x}_1^i$ represents the 3D structure of the bound complex. Reconstructing a stochastic process that diffuses $\mathbf{x}_0^i$'s to $\mathbf{x}_1^i$'s is tantamount to recovering the energy landscape governing the docking process. Similarly, in molecular dynamics simulations, we have access to trajectories $\left(\mathbf{x}_t^i\right)_{t\in[0,1]}$, where $\mathbf{x}_0^i$ and $\mathbf{x}_1^i$ represent the initial and final positions of the i-th molecule respectively. Any learning algorithm using these simulations should be able to respect the provided alignment.

7.3.1 SBALIGN: *Aligned Diffusion Schrödinger Bridges*

In this section, we derive a novel loss function for DSBs with aligned data by combining two classical notions: The theory of Schrödinger bridges (Schrödinger, 1931; Léonard, 2013; Chen et al., 2021b) and Doob's h-transform (Doob, 1984; Rogers and Williams, 2000). We then describe how solutions to DSBs with aligned data can be leveraged in the context of classical DSBs.

STATIC SB AND ALIGNED DATA Our starting point is the simple and classical observation that (7.8) is the continuous-time analogue of the *entropic optimal transport*, also known as the *static* Schrödinger bridge problem (Léonard, 2013; Chen et al., 2021b; Peyré and Cuturi, 2019):

$$\pi^\star := \underset{\mathbb{P}_0=\hat{\mathbb{P}}_0,\, \mathbb{P}_1=\hat{\mathbb{P}}_1}{\arg\min} \; D_{\mathrm{KL}}\mathbb{P}_{0,1}\mathbb{Q}_{0,1} \tag{7.39}$$

where the minimization is over all *couplings* of $\hat{\mathbb{P}}_0$ and $\hat{\mathbb{P}}_1$, and $\mathbb{Q}_{0,1}$ is simply the joint distribution of $\mathbb{Q}_t$ at $t = 0, 1$. In other words, if we denote by $\mathbb{P}_t$ the stochastic process that minimizes (7.8), then the joint distribution $\mathbb{P}_{0,1}$ necessarily coincides with the $\pi^\star$ in (7.39). Moreover, since in DSBs, the data is always assumed to arise from $\mathbb{P}_t$, we see that:

> The *aligned* data $(\mathbf{x}_0^i, \mathbf{x}_1^i)_{i=1}^N$ constitutes samples of $\pi^\star$.

This simple but crucial observation lies at the heart of all derivations to come.

Our central idea is to represent $\mathbb{P}_t$ via two different, but equivalent, characterizations, both of which involve $\pi^\star$: That of a *mixture* of reference processes with pinned end points, and that of conditional stochastic differential equations.

$\mathbb{P}_t$ FROM $\pi^\star$: $\mathbb{Q}_t$ WITH PINNED END POINTS For illustration purposes, from now on, we will assume that the reference process $\mathbb{Q}_t$ is a Brownian motion with diffusion coefficient g_t:

$$\mathrm{d}\mathbb{Q}_t = g_t \, \mathrm{d}\mathbb{W}_s. \tag{7.40}$$

In this case, it is well-known that $\mathbb{Q}_t$ *conditioned* to start at $\mathbf{x}_0$ and end at $\mathbf{x}_1$ can be written in another SDE (Mansuy and Yor, 2008; Liu et al., 2023):

$$\mathrm{d}X_t = g_t^2 \frac{\mathbf{x}_1 - X_t}{\beta_1 - \beta_t} \, \mathrm{d}t + g_t \, \mathrm{d}\mathbb{W}_s \tag{7.41}$$

where $X_0 = \mathbf{x}_0$ and

$$\beta_t := \int_0^t g_s^2 \, \mathrm{d}s. \tag{7.42}$$

We call the processes in (7.41) the *scaled Brownian bridges* as they generalize the classical Brownian bridge, which corresponds to the case of $g_t \equiv 1$.

The first characterization of $\mathbb{P}_t$ is then an immediate consequence the following classical result in Schrödinger bridge theory: Draw a sample $(\mathbf{x}_0, \mathbf{x}_1) \sim \pi^\star$ and connect them via (7.41). The resulting path is a sample from $\mathbb{P}_t$ (Léonard, 2013; Chen et al., 2021b). In other words, $\mathbb{P}_t$ is a *mixture* of scaled Brownian bridges, with the mixing weight given by $\pi^\star$.

$\mathbb{P}_t$ FROM $\pi^\star$: SDE REPRESENTATION Another characterization of $\mathbb{P}_t$ is that it is itself given by an SDE of the form (Léonard, 2013; Chen et al., 2021b)

$$\mathrm{d}X_t = g_t^2 b_t(X_t) \, \mathrm{d}t + g_t \, \mathrm{d}\mathbb{W}_s. \tag{7.43}$$

Here, $b_t : \mathbb{R}^d \to \mathbb{R}^d$ is a time-dependent drift function that we wish to learn. Now, by Doob's h-transform, we know that the SDE (7.43) *conditioned* to start at $\mathbf{x}_0$ and end at $\mathbf{x}_1$ is given by another SDE (Doob, 1984; Rogers and Williams, 2000):

$$\mathrm{d}X_t = g_t^2 [b_t(X_t) + \nabla \log h_t(X_t)] \, \mathrm{d}t + g_t \, \mathrm{d}\mathbb{W}_s \tag{7.44}$$

where $h_t(\mathbf{x}) := \mathbb{P}(X_1 = \mathbf{x}_1 | X_t = \mathbf{x})$ is the *Doob's h function*. Notice that we have suppressed the dependence of h_t on $\mathbf{x}_0$ and $\mathbf{x}_1$ for notational simplicity.

LOSS FUNCTION Since both (7.41) and (7.44) represent $\mathbb{P}_t$, the solution of the DSBs, the two SDEs must coincide. In other words, suppose we parametrize b_t as b_t^θ, then, by matching terms in (7.41) and (7.44), we can learn the optimal parameter $\theta^\star$ via optimization of the loss function

$$L(\theta) := \mathbb{E}\left[\int_0^1 \left\| \frac{\mathbf{x}_1 - X_t}{\beta_1 - \beta_t} - \nabla \log h_t^\theta(X_t) \right\|^2 \mathrm{d}t\right] \tag{7.45}$$

where h_t^θ is determined by b_t^θ as well as the drawn samples $(\mathbf{x}_0, \mathbf{x}_1)$. In short, assuming that, for each θ, we can compute h_t^θ *based only on* b_t^θ, we can then backprop through (7.45) and optimize it using any off-the-shelf algorithm.

A SLIGHTLY MODIFIED (7.45) Even with infinite data and a neural network with sufficient capacity, the loss function defined in (7.45) does converge to 0. For the purpose of numerical stability, we instead propose to modify (7.45) to:

$$L(\theta) := \mathbb{E}\left[\int_0^1 \left\| \frac{\mathbf{x}_1 - X_t}{\beta_1 - \beta_t} - \left(b_t^\theta + \nabla \log h_t^\theta(X_t)\right) \right\|^2 \mathrm{d}t\right] \tag{7.46}$$

which is clearly equivalent to (7.45) at the true solution of b_t. Notice that (7.46) bears a similar form as the popular score-matching objective employed in previous works (Song and Ermon, 2019; Song et al., 2021):

$$L(\theta) := \mathbb{E}\left[\int_0^1 \left\| \nabla \log p(\mathbf{x}_t|\mathbf{x}_0) - s^\theta(X_t, t) \right\|^2 \mathrm{d}t\right], \tag{7.47}$$

where the term $\frac{\mathbf{x}_1 - X_t}{\beta_1 - \beta_t}$ is akin to $\nabla \log p(\mathbf{x}_t|\mathbf{x}_0)$, while $\left(b_t^\theta + \nabla \log h_t^\theta(X_t)\right)$ corresponds to $s^\theta(X_t, t)$.

COMPUTING h_t^θ. Inspecting h_t in (7.44), we see that, given $(\mathbf{x}_0, \mathbf{x}_1)$, it can be written as the conditional expectation of an indicator function:

$$h_t(\mathbf{x}) = \mathbb{P}(X_1 = \mathbf{x}_1 | X_t = \mathbf{x}) = \mathbb{E}\left[\mathbb{1}_{\{\mathbf{x}_1\}} | X_t = \mathbf{x}\right] \tag{7.48}$$

where the expectation is over (7.43). Functions of the form (7.48) lend itself well to computation since it solves simulating the *unconditioned* paths. Furthermore, in order to avoid overfitting on the given samples, it is customary to replace the "hard" constraint $\mathbb{1}_{\{\mathbf{x}_1\}}$ by its *smoothed* version (Zhang and Chen, 2022; Holdijk et al., 2022):

$$h_{t,\tau}(\mathbf{x}) := \mathbb{E}\left[\exp\left(-\frac{1}{2\tau}\|X_1 - \mathbf{x}_1\|^2\right) | X_t = \mathbf{x}\right]. \tag{7.49}$$

Algorithm 4 SBALIGN

Input: Aligned data $(\mathbf{x}_0^i, \mathbf{x}_1^i)_{i=1}^N$, learning rates $\gamma_\theta, \gamma_\phi$, training iterations K.
Output: Optimal drift b_t^θ and parameterization m^ϕ of the "softened" Doob's h-transform $h_{t,\tau}$

Initialize $\theta \leftarrow \theta_0$, $\phi \leftarrow \phi_0$
for $k = 1$ **to** K **do**
  Draw a mini-batch of samples from $(\mathbf{x}_0^i, \mathbf{x}_1^i)_{i=1}^N$
  Compute empirical average of loss L (7.50) with mini-batch
  Update $\phi \leftarrow \phi - \gamma_\phi \nabla L(\theta, \phi)$
  Update $\theta \leftarrow \theta - \gamma_\theta \nabla L(\theta, \phi)$

Here, τ is a regularization parameter that controls how much we "soften" the constraint, and we have $\lim_{\tau \to 0} h_{t,\tau} = h_t$.
Although the computation of (7.49) can be done via a standard application of the Feynman-Kac formula (Rogers and Williams, 2000), an altogether easier approach is to parametrize $h_{t,\tau}$ by a second neural network m^ϕ and perform alternating minimization steps on b_t^θ and m^ϕ. This way, we can also avoid simulating even the unconditional paths of (7.43), and thereby further reducing the variance in training.

REGULARIZATION Since it is well-known that $\nabla \log h_t$ typically explodes when $t \to 1$ (Liu et al., 2023), it is important to regularize the behavior of m^ϕ for numerical stability, especially when $t \to 1$. Moreover, in practice, it is desirable to learn a drift b_t^θ that respects the data alignment *in expectation*: If $(\mathbf{x}_0, \mathbf{x}_1)$ is an input pair, then multiple runs of the SDE (7.43) starting from $\mathbf{x}_0$ should, on average, produce samples that are in the proximity of $\mathbf{x}_1$. This observation implies that we should search for drifts whose corresponding h-transforms are diminishing.
A simple way to simultaneously achieve the above two requirements is to add an ℓ^2-regularization term, resulting in the loss function:

$$L(\theta, \phi) := \mathbb{E}\left[\int_0^1 \left\| \frac{\mathbf{x}_1 - X_t}{\beta_1 - \beta_t} - \left(b_t^\theta + m^\phi(X_t) \right) \right\|^2 + \lambda_t \| m^\phi(\mathbf{x}_t) \|^2 \, \mathrm{d}t \right] \tag{7.50}$$

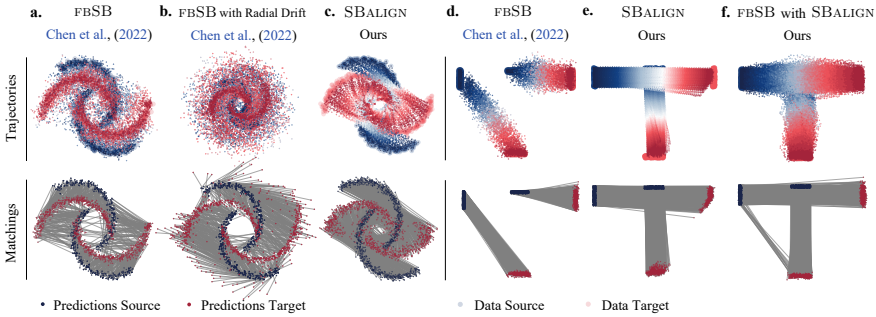


Figure 7.6: Experimental results on the Moon dataset (**a-c**) and T-dataset (**d-f**). The top row shows the trajectory sampled using the learned drift, and the bottom row shows the matching based on the learnt drift. Compared to other baselines, SBALIGN is able to learn an appropriate drift respecting the true alignment. (**f**) further showcases the utility of SBALIGN's learnt drift as a suitable reference process to improve other training methods.

where λ_t can either be constant or vary with time. The overall algorithm is depicted in Algorithm 4.

7.3.2 *Aligned Schrödinger Bridges as Prior Processes*

Classical SBs are unsuitable in cases where the alignments are known, because they only consider samples from $\hat{\mathbb{P}}_0$ and $\hat{\mathbb{P}}_1$ and disregard those drawn from the (optimal) coupling $\pi^\star$. However, the reliance of our method on this crucial knowledge is critical to avoid the necessity of IPF-like iterates but may become a limitation when insufficient information on alignments is available.

In such a situation, while it is unrealistic to hope for an accurate solution to the aligned SB problem, the interpolation between $\hat{\mathbb{P}}_0$ and $\hat{\mathbb{P}}_1$ learned by SBALIGN (7.43) can potentially still be leveraged to obtain a better reference process, when solving a classical SB on the same marginals —i.e. the term $b_t(X_t)$ learned via SBALIGN can, in fact, be used *as is* to construct a data-informed alternative $\tilde{\mathbb{Q}}_t$ to the standard Brownian motion (7.40).

Improved reference processes, either using pre-trained or data-informed ones, have been previously considered in the literature. For instance, both De Bortoli et al. (2021b) and Chen et al. (2022a) use a pre-trained reference process for challenging image interpolation tasks. This approach, however, relies on DSBs trained using the classical score-based generative modeling objective between a Gaussian and the data distribution. It therefore pre-

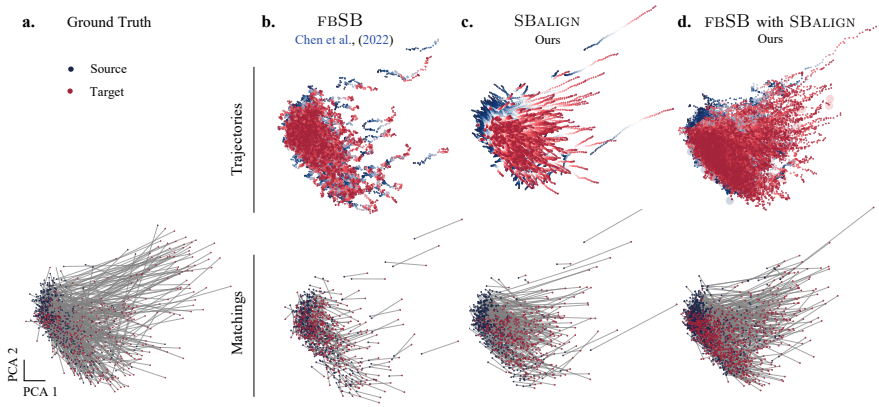


Figure 7.7: Cell differentiation trajectories based on (**a**) the ground truth and (**b-d**) learned drifts. SBALIGN is able to learn an appropriate drift underlying the true differentiation process while respecting the alignment. (**d**) Using the learned drift from SBALIGN as a reference process helps improve the drift learned by other training methods.

trains the reference process on a related —but different— process, i.e., the one mapping Gaussian noise to data rather than $\hat{\mathbb{P}}_0$ to $\hat{\mathbb{P}}_1$. An alternative, proposed by Bunne et al. (2023a), draws on the closed-form solution of SBs between two Gaussian distributions, which are chosen to approximate $\hat{\mathbb{P}}_0$ and $\hat{\mathbb{P}}_1$, respectively. Unlike our method, these alternatives construct better prior drifts by falling back to simpler and related tasks, or approximations of the original problem. We instead propose to shape a coarse-grained description of the drift based on alignments sampled directly from $\mathbb{P}_{0,1}$.

7.3.3 *Empirical Evaluation*

In this section, we evaluate SBALIGN in different settings involving 2-dimensional synthetic datasets, the task of reconstructing cellular differentiation processes, as well as predicting the conformation of a protein structure and its ligand formalized as rigid protein docking problem.

7.3.3.1 *Synthetic Experiments*

We run our algorithm on two synthetic datasets and compare the results with classic Schrödinger bridge models, i.e., the forward-backward SB formulation proposed by Chen et al. (2022a), herein referred to as FBSB. We

equip the baseline with prior knowledge, as elaborated below, to further challenge SBALIGN.

MOON DATASET. The first synthetic dataset (Fig. 7.6a-c) consists of two distributions, each supported on two semi-circles ($\hat{\mathbb{P}}_0$ drawn in *blue* and $\hat{\mathbb{P}}_1$ in *red*). $\hat{\mathbb{P}}_1$ was obtained from $\hat{\mathbb{P}}_0$ by applying a clockwise rotation around the center, i.e., by making points in the upper blue arm correspond to those in the right red one. This transformation is clearly not the most likely one under the assumption of Brownian motion of particles and should therefore not be found as the solution of a classical SB problem. This is confirmed by FBSB trajectories (Fig. 7.6a), which tend to map points to their closest neighbor in $\hat{\mathbb{P}}_1$ (e.g., some points in the upper arm of $\hat{\mathbb{P}}_0$ are brought towards the left rather than towards the right). While being a minimizer of (7.8), such a solution completely disregards our prior knowledge on the alignment of particles, which is instead reliably reproduced by the dynamics learned by SBALIGN (Fig. 7.6b).
One way of encoding this additional information on the nature of the process is to modify $\mathbb{Q}_t$ by introducing a clockwise radial drift, which describes the prior tangential velocity of particles moving circularly around the center. Solving the classical SB with this updated reference process indeed generates trajectories that respect most alignments (Fig. 7.6b), but requires a hand-crafted expression of the drift that is only possible in very simple cases.

T DATASET. In most real-world applications, it is very difficult to define an appropriate reference process $\mathbb{Q}_t$, which respects the known alignment without excessively distorting the trajectories from a solution to (7.8). This is already visible in simple examples like (Fig. 7.6d-f), in which the value of good candidate prior drifts at a specific location needs to vary wildly in time. In this dataset, $\hat{\mathbb{P}}_0$ and $\hat{\mathbb{P}}_1$ are both bi-modal distributions, each supported on two of the four extremes of an imaginary T-shaped area. We target alignments that connect the two arms of the T as well as the top cloud with the bottom one. We succeed in learning them with SBALIGN (Fig. 7.6e) but unsurprisingly fail when using the baseline FBSB (Fig. 7.6d) with a Brownian motion prior.
In this case, however, attempts at designing a better reference drift for FBSB must take into account the additional constraint that the horizontal and vertical particle trajectories intersect (see Fig. 7.6e), i.e., they cross the same area at times t_h and t_v (with $t_h > t_v$). This implies that the drift b_t, which ini-

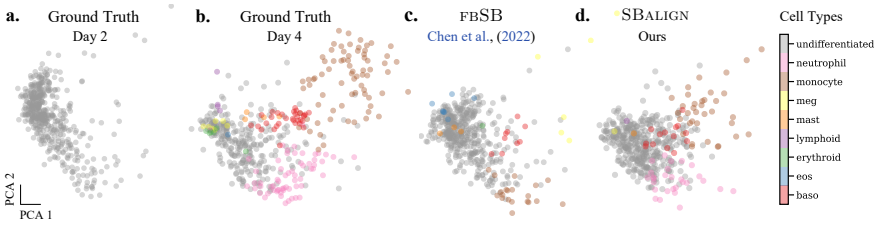


Figure 7.8: Cell type prediction on the differentiation dataset. All distributions are plotted on the first two principal components. **a-b:** Ground truth cell types on day 2 and day 4 respectively. **c-d:** FBSB and SBALIGN cell type predictions on day 4. SBALIGN is able to better model the underlying differentiation processes and capture the diversity in cell types.

tially points downwards (when $t < t_v$), should swiftly turn rightwards (for $t > t_h$). Setting imprecise values for one of t_h and t_v when defining custom reference drifts for classical SBs would hence not lead to the desired result and, worse, would actively disturb the flow of the other particle group.
As described in § 7.3.2, in presence of hard-to-capture requirements on the reference drift, the use of SBALIGN offers a remarkably easy and efficient way of learning a parameterization of it. For instance, when using the drift obtained by SBALIGN as reference drift for the computation of the SB baseline (FBSB), we find the desired alignments (Fig. 7.6f).

7.3.3.2 *Cell Differentiation*

Biological processes are determined through heterogeneous responses of single cells to external stimuli, i.e., developmental factors or drugs. Understanding and predicting the dynamics of single cells subject to a stimulus is thus crucial to enhance our understanding of health and disease and the focus of this task. Most single-cell high-throughput technologies are destructive assays —i.e., they destroy cells upon measurement— allowing us to only measure *unaligned* snapshots of the evolving cell population. Recent methods address this limitation by proposing (lower-throughput) technologies that keep cells alive after transcriptome profiling (Chen et al., 2022b) or that genetically tag cells to obtain a clonal trace upon cell division (Weinreb et al., 2020).

DATASET To showcase SBALIGN's ability to make use of such (partial) alignments when inferring cell differentiation processes, we take advantage of the genetic barcoding system developed by Weinreb et al. (2020). With a focus on fate determination in hematopoiesis, Weinreb et al. (2020) use expressed DNA barcodes to clonally trace single-cell transcriptomes over

	Cell Differentiation				
Methods	MMD ↓	W_ε ↓	ℓ_2(PS) ↓	RMSD ↓	Class. Acc. ↑
FBSB	1.58e-2	12.6	4.07	9.63e-1	58.0%
FBSB with SBALIGN	5.15e-3	10.6	0.95	9.88e-1	49.0%
SBalign	9.77e-3	11.2	1.24	9.28e-1	56.0%

Table 7.3: Cell differentiation prediction results. Shown are distributional metrics (MMD, W_ϵ), alignment-based metrics (ℓ_2, RMSD), and cell type classification accuracy for different methods on the cell differentiation dataset.

time. The dataset consists of two snapshots: the first, recorded on day 2, when most cells are still undifferentiated (see Fig. 7.8a), and a second, on day 4, comprising many different mature cell types (see Fig. 7.8b). Using SBALIGN as well as the baseline FSSB, we attempt to reconstruct cell evolution between day 2 and day 4, all while capturing the heterogeneity of emerging cell types.

BASELINES We benchmark SBALIGN against previous DSBs such as (Chen et al., 2022a, FBSB). Beyond, we compare SBALIGN in the setting of learning a prior reference process. Naturally, cell division processes and subsequently the propagation of the barcodes are very noisy. While this genetic annotation provides some form of assignment, it does not capture the full developmental process. We thus test SBALIGN in a setting where it learns a prior from such partial alignments and, plugged into FBSB, is fine-tuned on the full dataset.

EVALUATION METRICS To assess the performance of SBALIGN and the baselines, we monitor several metrics, which include distributional distances, i.e., MMD (Gretton et al., 2012) and W_ϵ (Cuturi, 2013), as well as average scores, i.e., ℓ_2(PS) (Bunne et al., 2023b) and RMSD. Moreover, we also train a simple neural network-based classifier to annotate the cell type on day 4 and we report the accuracy of the predicted vs. true cell type for all the models.

RESULTS SBALIGN accurately predicts cellular differentiation processes in hematopoiesis from day 2 to day 4, as visible from the (2D projections of the) learned trajectories and alignments (Fig. 7.7c) and the quantitative evaluation in Table 7.3. SBALIGN outperforms FBSB in all but the cell-type accuracy metric: Remarkably, our method exceeds the performances of the

baseline also on distributional metrics and not uniquely on alignment-based ones. Further, we evaluate how well SBALIGN recovers the heterogeneity of emerging cell types throughout the developmental process on day 4. The results are displayed in Fig. 7.8d and show that, while capturing the overall differentiation trend, SBALIGN (as well as FBSB) struggles to isolate rare cell types. Lastly, we employ SBALIGN to learn a prior process from noisy alignments based on genetic barcode annotations. When using this reference process within FBSB, we learn an SB which compensates for inaccuracies stemming from the stochastic nature of cell division and barcode redistribution and which achieves better scores on distributional metrics (see Tab. 7.3).

7.3.4 *Discussion*

In this paper, we propose a new framework to tackle the interpolation task with aligned data via diffusion Schrödinger bridges. Our central contribution is a novel algorithmic framework derived from the Schrödinger bridge theory and Doob's h-transform. Via a combination of the two notions, we derive novel loss functions which, unlike all prior methods for solving diffusion Schrödinger bridges, do not rely on the iterative proportional fitting procedure and are hence numerically stable. We verify our proposed algorithm on various synthetic and real-world tasks and demonstrate noticeable improvement over the previous state-of-the-art, thereby substantiating the claim that data alignment is a highly relevant feature that warrants further research.

8

CONCLUSION AND FUTURE DIRECTIONS

It's odd the way life works, the way it mutates and wanders, the way one thing becomes another.
— Siri Hustvedt, *What I Loved* (2003)

In this work we propose CellOT, a framework to model single-cell perturbation responses from unpaired treated and untreated cell states using neural optimal transport. By adequately modeling the nature of the problem through the lens of optimal transport, CellOT determines how perturbations affect cellular properties, reconstructs the most likely trajectory single cells take upon perturbation, and subsequently assists in a better understanding of driving factors of cell fate decision and cellular evasion mechanisms. CellOT builds on the recent successes of optimal transport applications in single-cell biology (Schiebinger et al., 2019; Lavenant et al., 2021), by introducing a fully parameterized transport map that can be applied to incoming unseen samples. Previous methods (Korotin et al., 2021a; Yang and Uhler, 2019; Prasad et al., 2020) rely on an unconstrained parameterization of the *primal* optimal transport map. However, the unconstrained nature of these models makes robust optimization challenging and results in reduced performance (Makkuva et al., 2020, Table 1). Instead, we learn the transformation of unperturbed to perturbed cell states through the *dual* optimal transport problem, parameterized via a pair of neural networks constrained to be convex (Makkuva et al., 2020). These constraints are important inductive biases that facilitate learning and result in a reliable and easy-to-train framework, as evidenced by the consistently strong performance of CellOT on several problems without the need for extensive hyperparameter tuning (see Online Methods).

CellOT infers the highly complex and nonlinear evolution of cell populations in response to perturbations without making strong simplifying assumptions on the nature of these dynamics. Unlike current approaches comprising autoencoder-based baselines (Lopez et al., 2018; Lotfollahi et al., 2019; Yang et al., 2020), CellOT does not necessarily rely on learning meaningful low-dimensional embeddings in which perturbations are modeled as linear shifts. We confirm this advantage through experiments on single-cell responses to different drugs in cancer cell lines obtained

with RNA-seq and spatially resolved 4i measurements, where CellOT consistently outperforms (Fig. 4.2). Our evaluations went beyond the often-used average treatment effect and correlation analysis across all cells; we analyzed marginals and computed MMD scores, a strong measure of how well predicted and observed distributions match.

Using CellOT to perform cell-state-aware drug profiling enables us to quantify perturbation effects as a function of the underlying heterogeneity of the studied system, in our cases a co-culture of two melanoma cell lines with different sensitivities to drug treatments. In doing so, we *sharpen* the response profiles of the measured drugs and reveal cell-state-specific responses of multiple signaling pathway in relation to treatment history of the cell line donor. We find the signaling activity associated to the MEK and PI3k pathways to decouple in cells pre-exposed to MEK inhibitors, a known adaptation mechanism for therapy evasion in melanoma cells (Kun et al., 2021). This *pathway rewiring* is associated to alteration in the molecular feedback structure of cells from effectors to receptors (Kun et al., 2021; Turke et al., 2012). Thus, combining CellOT with a larger set of combination treatments, multiplexed imaging, and cellular systems reflective of disease adaptations may help us to elucidate the molecular mechanisms of signaling pathway evolution in the context of cancer therapy.

We further analyze how well the learned maps generalize beyond samples used for training (o.o.s. setting) and to different sample compositions (o.o.d. setting). In Fig. 4.5, we therefore test CellOT's ability to to predict treatment responses in unseen lupus patients, infer developmental trajectories on stem cells of lower potency, and translate innate immune responses across patients. In all cases, CellOT's accuracy and precision are superior to current state-of-the-art methods (Fig. 4.5). Moreover, the predicted cell states after perturbation are still very close to the actually observed cell states. We consider these results as particularly promising, as it illustrates that accurate o.o.s. and o.o.d. predictions are indeed possible.

The ability to make predictions out-of-distribution, such as on unseen patients, is, however, only feasible if a) similar samples have been observed in the unperturbed setting, and b) the training set contains cases that are similar not only in their unperturbed state but also their perturbation response. An analysis of glioblastoma patients treated with Panobinostat (Zhao et al., 2021) (see Fig. A.3a-c) indeed confirms this restriction: CellOT and the baselines are able to predict treatment outcomes for those patients that are similar to other patients in both unperturbed state as well as perturbation effect (see Fig. A.3f), but fail to capture perturbation effects for patients

that exhibit unique responses (see Fig. A.3g). This limitation is important to consider when applying CellOT in o.o.d. settings. To overcome such problems, larger cohorts, additional meta-information, and methodological extensions are required. Bunne et al. (2022a) partially address this issue by deriving a neural optimal transport scheme that can be conditioned on a context, e.g., patient meta-data, when predicting perturbation responses.

We also observe that the predictive performance for CellOT drops when perturbations are too strong, i.e., the cell distributions before and after perturbations are very different (see Fig. 4.5j); a similar drop is observed for the other methods. The principle underlying the optimal transport theory is ideally suited for acute cellular perturbations during which single cells do not redistribute entirely and randomly in multidimensional measurement space, but typically only in a few dimensions, such that the overall correlation structure is preserved. While this modeling hypothesis is satisfied when perturbation responses are observed via regularly and frequently sampled snapshots, molecular transitions cannot be reconstructed when perturbation responses have progressed too far. For particularly strong or complicated perturbations, cellular multiplex profiles might change too drastically, violating OT assumptions and making it challenging to reconstruct the alignments between unperturbed and perturbed populations based on the *minimal effort* principle. In such settings, additional information is likely needed, for instance, a model of the underlying biology or models that integrate observations of multiple smaller time steps.
Despite the stochastic nature of cell fate decisions and the fact that cellular dynamics are intrinsically noisy (Wilkinson, 2009), CellOT models cell responses as deterministic trajectories. Approaches treating cell fate decisions as probabilistic events have previously allowed estimation of the full dynamical model to a greater extent than their deterministic counterparts (Bergen et al., 2020). By connecting OT and stochastic difference equations, recent work (Bunne et al., 2023a; Somnath et al., 2023) can build up on CellOT to account for biological heteroscedasticity, at the cost of added model complexity and other simplifying assumptions.

Despite having provided a proof-of-concept of the capacity of CellOT to model various chemical perturbations for different data modalities through an in-depth analysis of the nature of the learned mapping as well as a demonstration of its versatility in a broad class of applications, CellOT's generalization capacity has been evaluated on relatively small datasets. Crucially, large cohorts comprised of patients with different molecular profiles, such as cancer patients with various underlying genetics, could

result in strongly heterogeneous treatment responses. It is evident that approaches addressing these challenges could readily exploit the upcoming availability of large-scale patient cohort studies. The use of neural optimal transport to learn single-cell drug responses makes thus for an exciting avenue for future work, including its use to improve our understanding of cell therapies, study drug responses from patient samples, and better account for cell-to-cell variability in large-scale drug design efforts.

BIBLIOGRAPHY

Jason Altschuler, Francis Bach, Alessandro Rudi, and Jonathan Niles-Weed. Massively scalable Sinkhorn distances via the Nyström method. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019.

Jason Altschuler, Sinho Chewi, Patrik R Gerber, and Austin Stromme. Averaging on the Bures-Wasserstein manifold: dimension-free convergence of gradient descent. *Advances in Neural Information Processing Systems (NeurIPS)*, 34, 2021.

David Alvarez-Melis, Yair Schiff, and Youssef Mroueh. Optimizing Functionals on the Space of Probabilities with Input Convex Neural Networks. *Transactions on Machine Learning Research (TMLR)*, 2022.

L. Ambrosio, N. Gigli, and G. Savaré. *Gradient Flows in Metric Spaces and in the Space of Probability Measures*. Springer, 2006.

Brandon Amos. On amortizing convex conjugates for optimal transport. In *International Conference on Learning Representations (ICLR)*, 2023.

Brandon Amos, Lei Xu, and J Zico Kolter. Input Convex Neural Networks. In *International Conference on Machine Learning (ICML)*, volume 34, 2017.

Brandon Amos, Samuel Cohen, Giulia Luise, and Ievgen Redko. Meta Optimal Transport. *arXiv preprint arXiv:2206.05262*, 2022.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein Generative Adversarial Networks. In *International Conference on Machine Learning (ICML)*, 2017.

Nico Battich, Thomas Stoeger, and Lucas Pelkmans. Image-based transcriptomics in thousands of single human cells at single-molecule resolution. *Nature Methods*, 10(11), 2013.

Nico Battich, Thomas Stoeger, and Lucas Pelkmans. Control of transcript variability in single mammalian cells. *Cell*, 163(7), 2015.

Jean-David Benamou and Yann Brenier. A computational fluid mechanics solution to the Monge-Kantorovich mass transfer problem. *Numerische Mathematik*, 84(3), 2000.

Jean-David Benamou, Guillaume Carlier, and Maxime Laborde. An augmented lagrangian approach to wasserstein gradient flows and applications. *ESAIM: Proceedings and Surveys*, 54, 2016a.

Jean-David Benamou, Guillaume Carlier, Quentin Mérigot, and Edouard Oudet. Discretization of functionals involving the Monge–Ampére operator. *Numerische Mathematik*, 134(3), 2016b.

Volker Bergen, Marius Lange, Stefan Peidli, F Alexander Wolf, and Fabian J Theis. Generalizing RNA velocity to transient cell states through dynamical modeling. *Nature Biotechnology*, 38(12), 2020.

Espen Bernton, Jeremy Heng, Arnaud Doucet, and Pierre E Jacob. Schrödinger Bridge Samplers. In *arXiv preprint arXiv:1912.13170*, 2019.

Rajendra Bhatia, Tanvi Jain, and Yongdo Lim. On the Bures–Wasserstein distance between positive definite matrices. *Expositiones Mathematicae*, 37 (2), 2019.

Raicho Bojilov and Alfred Galichon. Matching in Closed-Form: Equilibrium, Identification, and Comparative Statics. *Economic Theory*, 61(4), 2016.

Nicolas Bonneel, Julien Rabin, Gabriel Peyré, and Hanspeter Pfister. Sliced and Radon Wasserstein Barycenters of Measures. *Journal of Mathematical Imaging and Vision*, 51(1), 2015.

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

Yann Brenier. Décomposition polaire et réarrangement monotone des champs de vecteurs. *CR Acad. Sci. Paris Sér. I Math.*, 305, 1987.

Yann Brenier. Polar Factorization and Monotone Rearrangement of Vector-Valued Functions. *Communications on Pure and Applied Mathematics*, 44(4), 1991.

Charlotte Bunne, Andreas Krause, and Marco Cuturi. Supervised Training of Conditional Monge Maps. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022a.

Charlotte Bunne, Laetitia Meng-Papaxanthos, Andreas Krause, and Marco Cuturi. Proximal Optimal Transport Modeling of Population Dynamics. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 25, 2022b.

Charlotte Bunne, Ya-Ping Hsieh, Marci Cuturi, and Andreas Krause. The Schrödinger Bridge between Gaussian Measures has a Closed Form. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2023a.

Charlotte Bunne, Stefan G Stark, Gabriele Gut, Jacobo Sarabia del Castillo, Kjong-Van Lehmann, Lucas Pelkmans, Andreas Krause, and Gunnar Ratsch. Learning Single-Cell Perturbation Responses using Neural Optimal Transport. *Nature Methods*, 2023b.

Martin Burger, Josè A. Carrillo, and Marie-Therese Wolfram. A mixed finite element method for nonlinear diffusion equations. *Kinetic & Related Models*, 3(1), 2010.

Luis A Caffarelli. Monotonicity Properties of Optimal Transportation and the FKG and Related Inequalities. *Communications in Mathematical Physics*, 214(3), 2000.

Zixuan Cang and Qing Nie. Inferring spatial and signaling relationships between cells from single cell transcriptomic data. *Nature Communications*, 11(1), 2020.

Guillaume Carlier, Lénaïc Chizat, and Maxime Laborde. Lipschitz Continuity of the Schrödinger Map in Entropic Optimal Transport. *arXiv preprint arXiv:2210.00225*, 2022.

Jose A Carrillo, Katy Craig, Li Wang, and Chaozhen Wei. Primal Dual Methods for Wasserstein Gradient Flows. *Foundations of Computational Mathematics*, 2021.

Rich Caruana. Multitask Learning. *Machine Learning*, 28(1), 1997.

Christopher J Caunt, Matthew J Sale, Paul D Smith, and Simon J Cook. MEK1 and MEK2 inhibitors and cancer therapy: the long and winding road. *Nature Reviews Cancer*, 15(10), 2015.

Yash Chandak, Georgios Theocharous, James Kostas, Scott Jordan, and Philip Thomas. Learning Action Representations for Reinforcement Learning. In *International Conference on Machine Learning (ICML)*, 2019.

Sisi Chen, Paul Rivaud, Jong H Park, Tiffany Tsou, Emeric Charles, John R Haliburton, Flavia Pichiorri, and Matt Thomson. Dissecting heterogeneous cell populations across drug and disease conditions with PopAlign. *Proceedings of the National Academy of Sciences (PNAS)*, 117(46), 2020.

Tianrong Chen, Guan-Horng Liu, and Evangelos A Theodorou. Likelihood Training of Schrödinger Bridge using Forward-Backward SDEs Theory. In *International Conference on Learning Representations (ICLR)*, 2022a.

Wanze Chen, Orane Guillaume-Gentil, Pernille Yde Rainer, Christoph G Gäbelein, Wouter Saelens, Vincent Gardeux, Amanda Klaeger, Riccardo Dainese, Magda Zachara, Tomaso Zambelli, et al. Live-seq enables temporal transcriptomic recording of single cells. *Nature*, 608, 2022b.

Yize Chen, Yuanyuan Shi, and Baosen Zhang. Optimal Control Via Neural Networks: A Convex Approach. In *International Conference on Learning Representations (ICLR)*, 2019.

Yongxin Chen, Tryphon T Georgiou, and Michele Pavon. Optimal Steering of a Linear Stochastic System to a Final Probability Distribution, Part I-III. *IEEE Transactions on Automatic Control*, 61(5), 2015.

Yongxin Chen, Tryphon T Georgiou, and Michele Pavon. On the relation between optimal transport and Schrödinger bridges: A stochastic control viewpoint. *Journal of Optimization Theory and Applications*, 169(2), 2016.

Yongxin Chen, Tryphon T Georgiou, and Michele Pavon. Optimal Transport in Systems and Control. *Annual Review of Control, Robotics, and Autonomous Systems*, 4, 2021a.

Yongxin Chen, Tryphon T Georgiou, and Michele Pavon. Stochastic control liaisons: Richard Sinkhorn meets Gaspard Monge on a Schrödinger bridge. *SIAM Review*, 63(2), 2021b.

Sinho Chewi, Tyler Maunu, Philippe Rigollet, and Austin J Stromme. Gradient descent algorithms for Bures-Wasserstein barycenters. In *Conference on Learning Theory (COLT)*, 2020.

Lénaïc Chizat, Gabriel Peyré, Bernhard Schmitzer, and Franccois-Xavier Vialard. Scaling algorithms for unbalanced optimal transport problems. *Mathematics of Computation*, 87(314), 2018.

Lénaïc Chizat, Stephen Zhang, Matthieu Heitz, and Geoffrey Schiebinger. Trajectory Inference via Mean-field Langevin in Path Space. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Patrick L Combettes and Jean-Christophe Pesquet. Proximal Splitting Methods in Signal Processing. In *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*. Springer, 2011.

Gabriele Corso, Hannes Stärk, Bowen Jing, Regina Barzilay, and Tommi Jaakkola. Diffusion Steps, Twists, and Turns for Molecular Docking. In *International Conference on Learning Representations (ICLR)*, 2023.

Marco Cuturi. Sinkhorn Distances: Lightspeed Computation of Optimal Transport. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 26, 2013.

Marco Cuturi, Laetitia Meng-Papaxanthos, Yingtao Tian, Charlotte Bunne, Geoff Davis, and Olivier Teboul. Optimal Transport Tools (OTT): A JAX Toolbox for all things Wasserstein. *arXiv Preprint arXiv:2201.12324*, 2022.

Marco Cuturi, Michal Klein, and Pierre Ablin. Monge, Bregman and Occam: Interpretable Optimal Transport in High-Dimensions with Feature-Sparse Maps. In *International Conference on Machine Learning (ICML)*, 2023.

Arnak S Dalalyan. Theoretical guarantees for approximate sampling from smooth and log-concave densities. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 2017.

Max Daniels, Tyler Maunu, and Paul Hand. Score-based Generative Neural Networks for Large-Scale Optimal Transport. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, 2021.

John M Danskin. *The Theory of Max-Min and its Applications to Weapons Allocation Problems*, volume 5. Springer, 1967.

Gwendoline De Bie, Gabriel Peyré, and Marco Cuturi. Stochastic Deep Networks. In *International Conference on Machine Learning (ICML)*, volume 36, 2019.

Valentin De Bortoli, Arnaud Doucet, Jeremy Heng, and James Thornton. Simulating Diffusion Bridges with Score Matching. In *arXiv preprint arXiv:2111.07243*, 2021a.

Valentin De Bortoli, James Thornton, Jeremy Heng, and Arnaud Doucet. Diffusion Schrödinger Bridge with Applications to Score-Based Generative Modeling. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, 2021b.

Valentin De Bortoli, Emile Mathieu, Michael Hutchinson, James Thornton, Yee Whye Teh, and Arnaud Doucet. Riemannian Score-Based Generative Modelling. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Eustasio del Barrio and Jean-Michel Loubes. The statistical effect of entropic regularization in optimal transportation. *arXiv preprint arXiv:2006.05199*, 2020.

Pinar Demetci, Rebecca Santorella, Björn Sandstede, William Stafford Noble, and Ritambhara Singh. SCOT: Single-Cell Multi-Omics Alignment with Optimal Transport. *Journal of Computational Biology*, 29(1), 2022.

Vincent Divol, Jonathan Niles-Weed, and Aram-Alexandre Pooladian. Optimal transport map estimation in general function spaces. *arXiv preprint arXiv:2212.03722*, 2022.

Atray Dixit, Oren Parnas, Biyu Li, Jenny Chen, Charles P Fulco, Livnat Jerby-Arnon, Nemanja D Marjanovic, Danielle Dionne, Tyler Burks, Raktima Raychowdhury, et al. Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens. *Cell*, 167 (7), 2016.

Joseph Doob. *Classical Potential Theory and Its Probabilistic Counterpart*, volume 549. Springer, 1984.

D. C. Dowson and B. V. Landau. The Fréchet distance between multivariate normal distributions. *Journal of Multivariate Analysis*, 12(3), 1982.

Harrison Edwards and Amos Storkey. Towards a Neural Statistician. In *International Conference on Learning Representations (ICLR)*, volume 5, 2017.

Jiaojiao Fan, Amirhossein Taghvaei, and Yongxin Chen. Scalable Computations of Wasserstein Barycenter via Input Convex Neural Networks. In *International Conference on Machine Learning (ICML)*, 2021.

Jean Feydy, Thibault Séjourné, Franccois-Xavier Vialard, Shun-Ichi Amari, Alain Trouvé, and Gabriel Peyré. Interpolating between Optimal Transport and MMD using Sinkhorn Divergences. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 22, 2019.

Alessio Figalli. The Optimal Partial Transport Problem. *Archive for Rational Mechanics and Analysis*, 195(2), 2010.

Alessio Figalli. *The Monge–Ampère equation and Its Applications*. Zurich Lectures in Advanced Mathematics, 2017.

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

R Flamary, N Courty, D Tuia, and A Rakotomamonjy. Optimal Transport for Domain Adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1, 2016.

Rémi Flamary, Nicolas Courty, Alexandre Gramfort, Mokhtar Z. Alaya, Aurélie Boisbunon, Stanislas Chambon, Laetitia Chapel, Adrien Corenflos, Kilian Fatras, Nemo Fournier, Léo Gautheron, Nathalie T.H. Gayraud, Hicham Janati, Alain Rakotomamonjy, Ievgen Redko, Antoine Rolet, Antony Schutz, Vivien Seguy, Danica J. Sutherland, Romain Tavenard, Alexander Tong, and Titouan Vayer. POT: Python Optimal Transport. *Journal of Machine Learning Research*, 22, 2021.

Jacoba Flier, Dick M Boorsma, Peter J van Beek, Cees Nieboer, Tom J Stoof, Rein Willemze, and Cornelis P Tensen. Differential expression of CXCR3 targeting chemokines CXCL10, CXCL9, and CXCL11 in different types of skin inflammation. *The Journal of Pathology: A Journal of the Pathological Society of Great Britain and Ireland*, 194(4), 2001.

Aden Forrow and Geoffrey Schiebinger. LineageOT is a unified framework for lineage tracing and trajectory inference. *Nature Communications*, 12(1), 2021.

Robert Fortet. Résolution d'un système déquations de M. Schrödinger. *J. Math. Pure Appl. IX*, 1, 1940.

Chris J Frangieh, Johannes C Melms, Pratiksha I Thakore, Kathryn R Geiger-Schuller, Patricia Ho, Adrienne M Luoma, Brian Cleary, Livnat Jerby-Arnon, Shruti Malu, Michael S Cuoco, et al. Multimodal pooled Perturb-CITE-seq screens in patient models define mechanisms of cancer immune evasion. *Nature Genetics*, 53(3), 2021.

Charlie Frogner, Chiyuan Zhang, Hossein Mobahi, Mauricio Araya, and Tomaso A Poggio. Learning with a Wasserstein Loss. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 28, 2015.

Octavian-Eugen Ganea, Xinyuan Huang, Charlotte Bunne, Yatao Bian, Regina Barzilay, Tommi S. Jaakkola, and Andreas Krause. Independent SE(3)-Equivariant Models for End-to-End Rigid Protein Docking. In *International Conference on Learning Representations (ICLR)*, 2022.

Matthias Gelbrich. On a Formula for the ℓ_2 Wasserstein Metric between Measures on Euclidean and Hilbert Spaces. *Mathematische Nachrichten*, 147(1), 1990.

Aude Genevay, Gabriel Peyré, and Marco Cuturi. Learning Generative Models with Sinkhorn Divergences. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2018.

Aude Genevay, Lénaïc Chizat, Francis Bach, Marco Cuturi, and Gabriel Peyré. Sample Complexity of Sinkhorn Divergences. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 22, 2019.

Ivan Gentil, Christian Léonard, and Luigia Ripani. About the analogy between optimal transport and minimal entropy. In *Annales de la Faculté des sciences de Toulouse: Mathématiques*, volume 26, 2017.

Ivan Gentil, Christian Léonard, and Luigia Ripani. Dynamical aspects of the generalized Schrödinger problem via Otto calculus–A heuristic point of view. *Revista Matem'atica Iberoamericana*, 36(4), 2020.

Laura González-Silva, Laura Quevedo, and Ignacio Varela. Tumor functional heterogeneity unraveled by scRNA-seq technologies. *Trends in Cancer*, 6 (1), 2020.

Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A Kernel Two-Sample Test. *Journal of Machine Learning Research*, 13, 2012.

Gabriele Gut, Markus D Herrmann, and Lucas Pelkmans. Multiplexed protein maps link subcellular organization to cellular states. *Science*, 361 (6401), 2018.

David Ha, Andrew Dai, and Quoc V Le. Hypernetworks. *arXiv preprint arXiv:1609.09106*, 2016.

Tzachi Hagai, Xi Chen, Ricardo J Miragaia, Raghd Rostom, Tom'as Gomes, Natalia Kunowska, Johan Henriksson, Jong-Eun Park, Valentina Proserpio, Giacomo Donati, et al. Gene expression variability across cells and species shapes innate immunity. *Nature*, 563(7730), 2018.

Andi Han, Bamdev Mishra, Pratik Kumar Jawanpuria, and Junbin Gao. On Riemannian optimization over positive definite matrices with the Bures-Wasserstein geometry. *Advances in Neural Information Processing Systems (NeurIPS)*, 34, 2021.

Tatsunori Hashimoto, David Gifford, and Tommi Jaakkola. Learning Population-Level Diffusions with Generative Recurrent Networks. In *International Conference on Machine Learning (ICML)*, volume 33, 2016.

Christian M Hedrich and George C Tsokos. Epigenetic mechanisms in systemic lupus erythematosus and other autoimmune diseases. *Trends in Molecular Medicine*, 17(12), 2011.

Tiam Heydari, Matthew A. Langley, Cynthia L Fisher, Daniel Aguilar-Hidalgo, Shreya Shukla, Ayako Yachie-Kinoshita, Michael Hughes, Kelly M. McNagny, and Peter W Zandstra. IQCELL: A platform for predicting the effect of gene perturbations on developmental trajectories using single-cell RNA-seq data. *PLoS Computational Biology*, 18(2), 2022.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising Diffusion Probabilistic Models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, 2020.

Lars Holdijk, Yuanqi Du, Ferry Hooft, Priyank Jaini, Bernd Ensing, and Max Welling. Path Integral Stochastic Optimal Control for Sampling Transition Paths. *arXiv preprint arXiv:2207.02149*, 2022.

Chin-Wei Huang, Ricky T. Q. Chen, Christos Tsirigotis, and Aaron Courville. Convex Potential Flows: Universal Probability Distributions with Optimal Transport and Convex Optimization. In *International Conference on Learning Representations (ICLR)*, 2021a.

Chin-Wei Huang, Jae Hyun Lim, and Aaron Courville. A Variational Perspective on Diffusion-Based Generative Models and Score Matching. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021b.

Jian Huang, Yuling Jiao, Lican Kang, Xu Liao, Jin Liu, and Yanyan Liu. Schrödinger-Föllmer Sampler: Sampling without Ergodicity. *arXiv preprint arXiv:2106.10880*, 2021c.

Geert-Jan Huizing, Gabriel Peyré, and Laura Cantini. Optimal transport improves cell–cell similarity inference in single-cell omics data. *Bioinformatics*, 38(8), 2022.

Jan-Christian Hütter and Philippe Rigollet. Minimax estimation of smooth optimal transport maps. *The Annals of Statistics*, 49(2), 2021.

Hicham Janati, Thomas Bazeille, Bertrand Thirion, Marco Cuturi, and Alexandre Gramfort. Multi-subject MEG/EEG source imaging with sparse multi-task regression. *NeuroImage*, 220, 2020a.

Hicham Janati, Boris Muzellec, Gabriel Peyré, and Marco Cuturi. Entropic Optimal Transport between Unbalanced Gaussian Measures has a Closed Form. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, 2020b.

Richard Jordan, David Kinderlehrer, and Felix Otto. The Variational Formulation of the Fokker–Planck Equation. *SIAM Journal on Mathematical Analysis*, 29(1), 1998.

Hyun Min Kang, Meena Subramaniam, Sasha Targ, Michelle Nguyen, Lenka Maliskova, Elizabeth McCarthy, Eunice Wan, Simon Wong, Lauren Byrnes, Cristina M Lanata, et al. Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. *Nature Biotechnology*, 36(1), 2018.

L Kantorovich. On the transfer of masses (in Russian). In *Doklady Akademii Nauk*, volume 37, 1942.

Diederik P Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. In *International Conference on Learning Representations (ICLR)*, 2014.

Peter E Kloeden and Eckhard Platen. Stochastic Differential Equations. In *Numerical Solution of Stochastic Differential Equations*. Springer, 1992.

Martin Knott and Cyril S Smith. On the optimal mapping of distributions. *Journal of Optimization Theory and Applications*, 43(1), 1984.

Alexander Korotin, Vage Egiazarian, Arip Asadulaev, Alexander Safin, and Evgeny Burnaev. Wasserstein-2 Generative Networks. In *International Conference on Learning Representations (ICLR)*, 2021a.

Alexander Korotin, Lingxiao Li, Aude Genevay, Justin M Solomon, Alexander Filippov, and Evgeny Burnaev. Do Neural Optimal Transport Solvers Work? A Continuous Wasserstein-2 Benchmark. *Advances in Neural Information Processing Systems (NeurIPS)*, 34, 2021b.

Bernhard A Kramer, Jacobo Sarabia del Castillo, and Lucas Pelkmans. Multimodal perception links cellular state to decision-making in single cells. *Science*, 377(6606), 2022.

Ashwinikumar Kulkarni, Ashley G Anderson, Devin P Merullo, and Genevieve Konopka. Beyond bulk: a review of single cell transcriptomics methodologies and applications. *Current Opinion in Biotechnology*, 58, 2019.

Solomon Kullback. Probability densities with given marginals. *The Annals of Mathematical Statistics*, 39(4), 1968.

Shivanni Kummar, Helen X Chen, John Wright, Susan Holbeck, Myrtle Davis Millin, Joseph Tomaszewski, James Zweibel, Jerry Collins, and James H Doroshow. Utilizing targeted cancer therapeutic agents in combination: novel approaches and urgent requirements. *Nature Reviews Drug discovery*, 9(11), 2010.

E Kun, YTM Tsang, CW Ng, DM Gershenson, and KK Wong. MEK inhibitor resistance mechanisms and recent developments in combination trials. *Cancer Treatment Reviews*, 92:102137, 2021.

Hugo Lavenant, Stephen Zhang, Young-Heon Kim, and Geoffrey Schiebinger. Towards a mathematical theory of trajectory inference. *arXiv preprint arXiv:2102.09204*, 2021.

Christian Léonard. A survey of the Schrödinger problem and some of its connections with optimal transport. *arXiv preprint arXiv:1308.0215*, 2013.

Chunbo Li, Hao Wu, Luopei Guo, Danyang Liu, Shimin Yang, Shengli Li, and Keqin Hua. Single-cell transcriptomics reveals cellular heterogeneity and molecular stratification of cervical cancer. *Communications Biology*, 5 (1), 2022.

Prisca Liberali, Berend Snijder, and Lucas Pelkmans. A hierarchical map of regulatory genetic interactions in membrane trafficking. *Cell*, 157(6), 2014.

Guan-Horng Liu, Tianrong Chen, Oswin So, and Evangelos A Theodorou. Deep Generalized Schrödinger Bridge. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Xingchao Liu, Lemeng Wu, Mao Ye, and qiang Liu. Learning Diffusion Bridges on Constrained Domains. *International Conference on Learning Representations (ICLR)*, 2023.

Romain Lopez, Jeffrey Regier, Michael B Cole, Michael I Jordan, and Nir Yosef. Deep generative modeling for single-cell transcriptomics. *Nature Methods*, 15(12), 2018.

Jonathan Lorraine, Paul Vicol, and David Duvenaud. Optimizing Millions of Hyperparameters by Implicit Differentiation. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2020.

Mohammad Lotfollahi, F Alexander Wolf, and Fabian J Theis. scGen predicts single-cell perturbation responses. *Nature Methods*, 16(8), 2019.

Mohammad Lotfollahi, Anna Klimovskaia Susmelj, Carlo De Donno, Leon Hetzel, Yuge Ji, Ignacio L Ibarra, Sanjay R Srivatsan, Mohsen Naghipourfar, Riza M Daza, Beth Martin, et al. Predicting cellular responses to complex perturbations in high-throughput screens. *Molecular Systems Biology*, 2023.

Frederike Lübeck, Charlotte Bunne, Gabriele Gut, Jacobo Sarabia del Castillo, Lucas Pelkmans, and David Alvarez-Melis. Neural Unbalanced Optimal Transport via Cycle-Consistent Semi-Couplings. *arXiv preprint arXiv:2209.15621*, 2022.

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

Ashok Makkuva, Amirhossein Taghvaei, Sewoong Oh, and Jason Lee. Optimal transport mapping via input convex neural networks. In *International Conference on Machine Learning (ICML)*, volume 119, 2020.

Anton Mallasto, Augusto Gerolin, and H'a Quang Minh. Entropy-regularized 2-Wasserstein distance between Gaussian measures. *Information Geometry*, 2021.

Roger Mansuy and Marc Yor. *Aspects of Brownian motion*. Springer Science & Business Media, 2008.

Alexis Mathian, Miguel Hie, Fleur Cohen-Aubart, and Zahir Amoura. Targeting interferons in systemic lupus erythematosus: current and future prospects. *Drugs*, 75(8), 2015.

Robert J McCann. A Convexity Principle for Interacting Gases. *Advances in Mathematics*, 128(1), 1997.

Leland McInnes, John Healy, and James Melville. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv Preprint arXiv:1802.03426*, 2018.

Facundo Mémoli. Gromov-Wasserstein Distances and the Metric Approach to Object Matching. *Foundations of Computational Mathematics*, 11, 2011.

Luke Metz, Ben Poole, David Pfau, and Jascha Sohl-Dickstein. Unrolled Generative Adversarial Networks. In *International Conference on Learning Representations (ICLR)*, 2017.

Reza Bayat Mokhtari, Tina S Homayouni, Narges Baluch, Evgeniya Morgatskaya, Sushil Kumar, Bikul Das, and Herman Yeger. Combination therapy in combating cancer. *Oncotarget*, 8(23):38022, 2017.

Petr Mokrov, Alexander Korotin, Lingxiao Li, Aude Genevay, Justin Solomon, and Evgeny Burnaev. Large-Scale Wasserstein Gradient Flows. *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Gaspard Monge. Mémoire sur la théorie des déblais et des remblais. *Histoire de l'Académie Royale des Sciences*, 1781.

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

Noa Moriel, Enes Senel, Nir Friedman, Nikolaus Rajewsky, Nikos Karaiskos, and Mor Nitzan. NovoSpaRc: flexible spatial reconstruction of single-cell gene expression with optimal transport. *Nature Protocols*, 16(9), 2021.

Thomas M Norman, Max A Horlbeck, Joseph M Replogle, Alex Y Ge, Albert Xu, Marco Jost, Luke A Gilbert, and Jonathan S Weissman. Exploring genetic interaction manifolds constructed from rich single-cell phenotypes. *Science*, 365(6455), 2019.

Ingram Olkin and Friedrich Pukelsheim. The distance between two random vectors with given dispersion matrices. *Linear Algebra and its Applications*, 48, 1982.

Felix Otto. The geometry of dissipative evolution equations: the porous medium equation. *Taylor & Francis*, 2001.

François-Pierre Paty, Alexandre d'Aspremont, and Marco Cuturi. Regularity as Regularization: Smooth and Strongly Convex Brenier Potentials in Optimal Transport. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2020.

Danielle Perez-Bercoff, Hélène Laude, Morgane Lemaire, Oliver Hunewald, Valerie Thiers, Marco Vignuzzi, Hervé Blanc, Aurélie Poli, Zahir Amoura, Vincent Caval, et al. Sustained high expression of multiple APOBEC3 cytidine deaminases in systemic lupus erythematosus. *Scientific Reports*, 11(1), 2021.

Gabriel Peyré. Entropic Approximation of Wasserstein Gradient Flows. *SIAM Journal on Imaging Sciences*, 8(4), 2015.

Gabriel Peyré and Marco Cuturi. Computational Optimal Transport. *Foundations and Trends in Machine Learning*, 11(5-6), 2019.

Eckhard Platen and Nicola Bruti-Liberati. *Numerical Solution of Stochastic Differential Equations with Jumps in Finance*, volume 64. Springer Science & Business Media, 2010.

Aram-Alexandre Pooladian and Jonathan Niles-Weed. Entropic estimation of optimal transport maps. *arXiv preprint arXiv:2109.12004*, 2021.

Aram-Alexandre Pooladian, Vincent Divol, and Jonathan Niles-Weed. Minimax estimation of discontinuous optimal transport maps: The semi-discrete case. In *International Conference on Machine Learning (ICML)*, 2023.

Neha Prasad, Karren Yang, and Caroline Uhler. Optimal Transport using GANs for Lineage Tracing. *arXiv preprint arXiv:2007.12098*, 2020.

Philip E Protter. Stochastic Differential Equations. In *Stochastic Integration and Differential Equations*. Springer, 2005.

Marieke IG Raaijmakers, Daniel S Widmer, Melanie Maudrich, Tabea Koch, Alice Langer, Anna Flace, Claudia Schnyder, Reinhard Dummer, and

Mitchell P Levesque. A new live-cell biobank workflow efficiently recovers heterogeneous melanoma cells from native biopsies. *Experimental Dermatology*, 24(5), 2015.

Julien Rabin, Gabriel Peyré, Julie Delon, and Marc Bernot. Wasserstein Barycenter and Its Application to Texture Mixing. In *International Conference on Scale Space and Variational Methods in Computer Vision*. Springer, 2011.

Aaditya Ramdas, Nicol'as Garc'ia Trillos, and Marco Cuturi. On Wasserstein Two Sample Testing and Related Families of Nonparametric Tests. *Entropy*, 19(2):47, 2017.

MR Sandhya Rani, Graham R Foster, Stewart Leung, Douglas Leaman, George R Stark, and Richard M Ransohoff. Characterization of β-R1, a gene that is selectively induced by interferon β (IFN-β) compared with IFN-α. *Journal of Biological Chemistry*, 271(37), 1996.

Jack Richter-Powell, Jonathan Lorraine, and Brandon Amos. Input Convex Gradient Networks. *arXiv preprint arXiv:2111.12187*, 2021.

Philippe Rigollet and Austin J Stromme. On the sample complexity of entropic optimal transport. *arXiv preprint arXiv:2206.13472*, 2022.

David Rogers and Mathew Hahn. Extended-Connectivity Fingerprints. *Journal of Chemical Information and Modeling*, 50(5), 2010.

L Chris G Rogers and David Williams. *Diffusions, Markov Processes and Martingales: Volume 2, Itô Calculus*, volume 2. Cambridge University Press, 2000.

Yu Rong, Yatao Bian, Tingyang Xu, Weiyang Xie, Ying Wei, Wenbing Huang, and Junzhou Huang. Self-Supervised Graph Transformer on Large-Scale Molecular Data. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, 2020.

Tim Salimans, Han Zhang, Alec Radford, and Dimitris Metaxas. Improving GANs Using Optimal Transport. In *International Conference on Learning Representations (ICLR)*, 2018.

Filippo Santambrogio. Optimal Transport for Applied Mathematicians. *Birkhäuser*, 55(58-63):94, 2015.

Filippo Santambrogio. {Euclidean, metric, and Wasserstein} gradient flows: an overview. *Bulletin of Mathematical Sciences*, 7(1), 2017.

Kevin Scaman and Aladin Virmaux. Lipschitz regularity of deep neural networks: analysis and efficient estimation. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 31, 2018.

Meyer Scetbon and Marco Cuturi. Low-rank Optimal Transport: Approximation, Statistics and Debiasing. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, 2022.

Meyer Scetbon, Marco Cuturi, and Gabriel Peyré. Low-Rank Sinkhorn Factorization. In *International Conference on Machine Learning (ICML)*, 2021.

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), 2019.

Morgan A Schmitz, Matthieu Heitz, Nicolas Bonneel, Fred Ngole, David Coeurjolly, Marco Cuturi, Gabriel Peyré, and Jean-Luc Starck. Wasserstein dictionary learning: Optimal transport-based unsupervised nonlinear dictionary learning. *SIAM Journal on Imaging Sciences*, 11(1), 2018.

Erwin Schrödinger. *Über die Umkehrung der Naturgesetze*. Verlag der Akademie der Wissenschaften in Kommission bei Walter De Gruyter u. Company, 1931.

Philippe Schwaller, Alain C Vaucher, Ruben Laplaza, Charlotte Bunne, Andreas Krause, Clemence Corminboeuf, and Teodoro Laino. Machine intelligence for chemical reaction space. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 2022.

Sydney M Shaffer, Margaret C Dunagin, Stefan R Torborg, Eduardo A Torre, Benjamin Emert, Clemens Krepler, Marilda Beqiri, Katrin Sproesser, Patricia A Brafford, Min Xiao, et al. Rare cell variability and drug-induced reprogramming as a mode of cancer drug resistance. *Nature*, 546(7658), 2017.

Yuyang Shi, Valentin De Bortoli, Andrew Campbell, and Arnaud Doucet. Diffusion Schrödinger Bridge Matching. *arXiv preprint arXiv:2303.16852*, 2023.

Sidak Pal Singh and Martin Jaggi. Model Fusion via Optimal Transport. *Advances in Neural Information Processing Systems (NeurIPS)*, 33, 2020.

Michael P Smith, Holly Brunton, Emily J Rowling, Jennifer Ferguson, Imanol Arozarena, Zsofia Miskolczi, Jessica L Lee, Maria R Girotti, Richard Marais, Mitchell P Levesque, et al. Inhibiting drivers of non-mutational drug tolerance is a salvage strategy for targeted melanoma therapy. *Cancer Cell*, 29(3), 2016.

Berend Snijder, Raphael Sacher, Pauli Rämö, Eva-Maria Damm, Prisca Liberali, and Lucas Pelkmans. Population context determines cell-to-cell variability in endocytosis and virus infection. *Nature*, 461(7263), 2009.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep Unsupervised Learning using Nonequilibrium Thermodynamics. In *International Conference on Machine Learning (ICML)*, 2015.

Justin Solomon, Gabriel Peyré, Vladimir G Kim, and Suvrit Sra. Entropic metric alignment for correspondence problems. *ACM Transactions on Graphics (TOG)*, 35(4), 2016.

Vignesh Ram Somnath, Matteo Pariset, Ya-Ping Hsieh, Maria Rodriguez Martinez, Andreas Krause, and Charlotte Bunne. Aligned Diffusion Schrödinger Bridges. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, 2023.

Yang Song and Stefano Ermon. Generative Modeling by Estimating Gradients of the Data Distribution. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-Based Generative Modeling through Stochastic Differential Equations. In *International Conference on Learning Representations (ICLR)*, 2021.

Sanjay R Srivatsan, Jose L McFaline-Figueroa, Vijay Ramani, Lauren Saunders, Junyue Cao, Jonathan Packer, Hannah A Pliner, Dana L Jackson, Riza M Daza, Lena Christiansen, et al. Massively multiplex chemical transcriptomics at single-cell resolution. *Science*, 367(6473), 2020.

Asuka Takatsu. On Wasserstein geometry of Gaussian measures. In *Probabilistic Approach to Geometry*. Mathematical Society of Japan, 2010.

Guy Tennenholtz and Shie Mannor. The Natural Language of Actions. In *International Conference on Machine Learning (ICML)*, 2019.

James Thornton, Michael Hutchinson, Emile Mathieu, Valentin De Bortoli, Yee Whye Teh, and Arnaud Doucet. Riemannian Diffusion Schrödinger Bridge. *arXiv preprint arXiv:2207.03024*, 2022.

Alexander Tong, Jessie Huang, Guy Wolf, David Van Dijk, and Smita Krishnaswamy. TrajectoryNet: A Dynamic Optimal Transport Network for Modeling Cellular Dynamics. In *International Conference on Machine Learning (ICML)*, 2020.

Alexander Tong, Nikolay Malkin, Guillaume Huguet, Yanlei Zhang, Jarrid Rector-Brooks, Kilian Fatras, Guy Wolf, and Yoshua Bengio. Conditional Flow Matching: Simulation-Free Dynamic Optimal Transport. *arXiv preprint arXiv:2302.00482*, 2023.

Sophie Tritschler, Maren Büttner, David S Fischer, Marius Lange, Volker Bergen, Heiko Lickert, and Fabian J Theis. Concepts and limitations for learning developmental trajectories from single cell genomics. *Development*, 146(12), 2019.

Tomer Tsaban, Julia K Varga, Orly Avraham, Ziv Ben-Aharon, Alisa Khramushin, and Ora Schueler-Furman. Harnessing protein folding neural networks for peptide–protein docking. *Nature Communications*, 13(1):176, 2022.

Alexa B Turke, Youngchul Song, Carlotta Costa, Rebecca Cook, Carlos L Arteaga, John M Asara, and Jeffrey A Engelman. MEK inhibition leads to PI3K/AKT activation by relieving a negative feedback on ERBB receptors. *Cancer Research*, 72(13), 2012.

Théo Uscidda and Marco Cuturi. The Monge Gap: A Regularizer to Learn All Transport Maps. In *International Conference on Machine Learning (ICML)*, 2023.

user26872. Reference for Multidimensional Gaussian Integral. Mathematics Stack Exchange, 2012. URL https://math.stackexchange.com/q/126767.

Francisco Vargas, Pierre Thodoroff, Neil D Lawrence, and Austen Lamacraft. Solving Schrödinger Bridges via Maximum Likelihood. *Entropy*, 23(9), 2021.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is All you Need. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Titouan Vayer, Rémi Flamary, Romain Tavenard, Laetitia Chapel, and Nicolas Courty. Sliced Gromov-Wasserstein. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019.

Cédric Villani. *Topics in Optimal Transportation*, volume 58. American Mathematical Soc., 2003.

Cédric Villani. *Optimal Transport: Old and New*, volume 338. Springer, 2009.

Gefei Wang, Yuling Jiao, Qian Xu, Yang Wang, and Can Yang. Deep Generative Learning via Schrödinger Bridge. In *International Conference on Machine Learning (ICML)*, volume 139, 2021.

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 (PNAS)*, 115(10), 2018.

Caleb Weinreb, Alejo Rodriguez-Fraticelli, Fernando D Camargo, and Allon M Klein. Lineage tracing on transcriptional landscapes links state to fate during differentiation. *Science*, 367(6479), 2020.

Darren J Wilkinson. Stochastic modelling for quantitative description of heterogeneous biological systems. *Nature Reviews Genetics*, 10(2), 2009.

Ronald J Williams and David Zipser. A Learning Algorithm for Continually Running Fully Recurrent Neural Networks. *Neural Computation*, 1(2), 1989.

Fengying Wu, Jue Fan, Yayi He, Anwen Xiong, Jia Yu, Yixin Li, Yan Zhang, Wencheng Zhao, Fei Zhou, Wei Li, et al. Single-cell profiling of tumor heterogeneity and the microenvironment in advanced non-small cell lung cancer. *Nature Communications*, 12(1), 2021.

Karren D Yang and Caroline Uhler. Scalable Unbalanced Optimal Transport using Generative Adversarial Networks. *International Conference on Learning Representations (ICLR)*, 2019.

Karren Dai Yang, Karthik Damodaran, Saradha Venkatachalapathy, Ali C Soylemezoglu, GV Shivashankar, and Caroline Uhler. Predicting cell lineages using autoencoders and optimal transport. *PLoS Computational Biology*, 16(4), 2020.

Karren Dai Yang, Anastasiya Belyaeva, Saradha Venkatachalapathy, Karthik Damodaran, Abigail Katcoff, Adityanarayanan Radhakrishnan, GV Shivashankar, and Caroline Uhler. Multi-domain translation between single-cell imaging and sequencing data using autoencoders. *Nature Communications*, 12(1), 2021.

Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Russ R Salakhutdinov, and Alexander J Smola. Deep Sets. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, 2017.

Anthony Zee. *Quantum Field Theory in a Nutshell*, volume 7. Princeton University Press, 2010.

Qinsheng Zhang and Yongxin Chen. Path Integral Sampler: A Stochastic Control Approach For Sampling. In *International Conference on Learning Representations (ICLR)*, 2022.

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), 2021.

Wenting Zhao, Athanassios Dovas, Eleonora Francesca Spinazzi, Hanna Mendes Levitin, Matei Alexandru Banu, Pavan Upadhyayula, Tejaswi Sudhakar, Tamara Marie, Marc L Otten, Michael B Sisti, et al. Deconvolution of cell type-specific drug responses in human tumor tissue with single-cell RNA-seq. *Genome Medicine*, 13(1), 2021.

CURRICULUM VITAE

PERSONAL DATA

Name	Charlotte Bunne
Date of Birth	August 29, 1995
Place of Birth	Karlsruhe, Germany
Citizen of	Germany

EDUCATION

2022-2023	Broad Institute of MIT and Harvard, Cambridge (MA), USA Visiting Graduate Student
2016 – 2019	Eidgenössische Technische Hochschule (ETH), Zürich, Switzerland *Final Degree:* Master of Science
2018-2019	Massachusetts Institute of Technology (MIT), Cambridge (MA), USA Visiting Student
2013 – 2016	Heidelberg University Heidelberg, Germany *Final Degree:* Bachelor of Science

EMPLOYMENT

2022	Research Intern *Apple*, Paris, France
2020	Research Intern *Google Research*, Zürich, Switzerland
2017	Research Intern *IBM Research*, Zürich, Switzerland

PUBLICATIONS

Charlotte Bunne, David Alvarez-Melis, Andreas Krause, and Stefanie Jegelka. Learning Generative Models across Incomparable Spaces. In *International Conference on Machine Learning (ICML)*, 2019.

Charlotte Bunne, Andreas Krause, and Marco Cuturi. Supervised Training of Conditional Monge Maps. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022a.

Charlotte Bunne, Laetitia Meng-Papaxanthos, Andreas Krause, and Marco Cuturi. Proximal Optimal Transport Modeling of Population Dynamics. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 25, 2022b.

Charlotte Bunne, Ya-Ping Hsieh, Marci Cuturi, and Andreas Krause. The Schrödinger Bridge between Gaussian Measures has a Closed Form. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2023a.

Charlotte Bunne, Stefan G Stark, Gabriele Gut, Jacobo Sarabia del Castillo, Kjong-Van Lehmann, Lucas Pelkmans, Andreas Krause, and Gunnar Ratsch. Learning Single-Cell Perturbation Responses using Neural Optimal Transport. *Nature Methods*, 2023b.

Marco Cuturi, Laetitia Meng-Papaxanthos, Yingtao Tian, Charlotte Bunne, Geoff Davis, and Olivier Teboul. Optimal Transport Tools (OTT): A JAX Toolbox for all things Wasserstein. *arXiv Preprint arXiv:2201.12324*, 2022.

Octavian-Eugen Ganea, Xinyuan Huang, Charlotte Bunne, Yatao Bian, Regina Barzilay, Tommi S. Jaakkola, and Andreas Krause. Independent SE(3)-Equivariant Models for End-to-End Rigid Protein Docking. In *International Conference on Learning Representations (ICLR)*, 2022.

Frederike Lübeck, Charlotte Bunne, Gabriele Gut, Jacobo Sarabia del Castillo, Lucas Pelkmans, and David Alvarez-Melis. Neural Unbalanced Optimal Transport via Cycle-Consistent Semi-Couplings. *arXiv preprint arXiv:2209.15621*, 2022.

Philippe Schwaller, Alain C Vaucher, Ruben Laplaza, Charlotte Bunne, Andreas Krause, Clemence Corminboeuf, and Teodoro Laino. Machine intelligence for chemical reaction space. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 2022.

Vignesh Ram Somnath, Charlotte Bunne, Connor Coley, Andreas Krause, and Regina Barzilay. Learning Graph Models for Retrosynthesis Prediction. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021a.

Vignesh Ram Somnath, Charlotte Bunne, and Andreas Krause. Multi-Scale Representation Learning on Proteins. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021b.

Vignesh Ram Somnath, Matteo Pariset, Ya-Ping Hsieh, Maria Rodriguez Martinez, Andreas Krause, and Charlotte Bunne. Aligned Diffusion Schrödinger Bridges. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, 2023.

A

APPENDIX

A.1 FURTHER EMPIRICAL EVALUATION

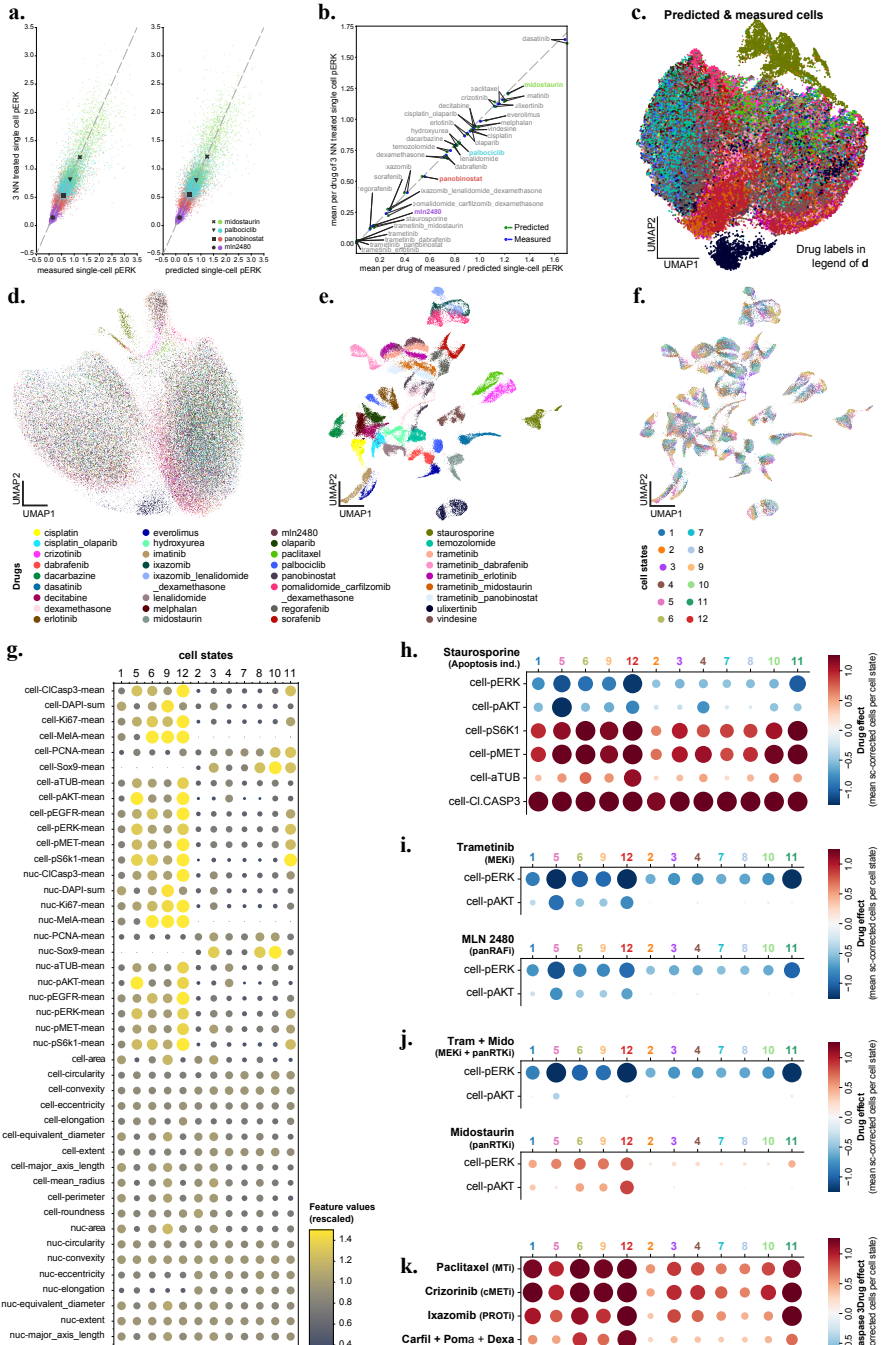


Figure A.1: **a.** High similarity of measured and CellOT-predicted single-cell pERK (phosphor ERK1/2) values at the single-cell level. Scatter plots compares the relationship between measured pERK values of cells (left) treated with Midostaurin (green dots), Palbociclib (blue dots), Panobinostat (red dots), and MLN2480 (purple dots) or (right) predicted for those drugs along the horizontal axis to their corresponding 3NN cells on the vertical axis. X mark, square, inverted triangle, and circle represent the mean of the respective measurements per drug. The dashed gray line indicates the diagonal along which the measurements would correlate perfectly. **b.** The high similarity of measured and CellOT-predicted single-cell pERK (phosphor ERK1/2) values at the population level across all drug perturbations. Drug average of measured (blue dots) and predicted (green dots) pERK values compared to their respective 3NN measurement. Drug treatments highlighted in color correspond the those presented in panel **a.**. The dashed gray line indicates the diagonal along which the measurements would correlate perfectly. **c.** Projection of measured perturbed and predicted perturbed cells in a shared UMAP space. Each cell is color-coded according to the perturbation from which it originates. **d.** Projection of mean-corrected measured perturbed cells in a UMAP space. Each cell is color-coded according to the perturbation from which it originates. Mean correction was achieved by subtracting calculating the mean of every feature for all cells in the control condition and subtracting the calculated feature means from the feature values of individual cells. **e.** Projection of single-cell corrected, predicted perturbed cells in a UMAP space. Each cell is color-coded according to the perturbation model with which it was predicted. **f.** Projection of single-cell corrected, predicted perturbed cells in a UMAP space. Each cell is color-coded according to its assignment to one of the 12 cell states. **g.** Feature value overview of the 12 identified cell states in DMSO-treated (control) cells. Each column represents a cell state, each row a feature. Circles are colored and scaled based on feature value, from small size in blue for low feature values, to large circles in yellow for high feature values. **h-j.** Drug effect overview of the 12 identified cell states in **h.** Staurosporine (apoptosis ind.m apoptosis inducer, **i.** Trametinib (MEKi, MEK inhibitor), MLN2480 (panRAFi, panRAF inhibitor), **j.** Trametinib + Midostaurin (Tram + Mido, MEK inhibitor + pan Receptor Tyrosine Kinase inhibitor (panRTK)), Midostaurin (panRTK). Each column represents a cell state, rows represent features. "cell-" stands for mean cell intensity. Circles are scaled based on drug effect, the larger the $\pm$ effect the larger the circles. Negative values are encoded in hues of blue, positive values in red hues of the respective circles. **k.** Effect of drug treatments on levels of cleaved Caspase 3 (cleaved Caspase 3) in the 12 identified cell states. Each column represents a cell state, each row a drug treatment.

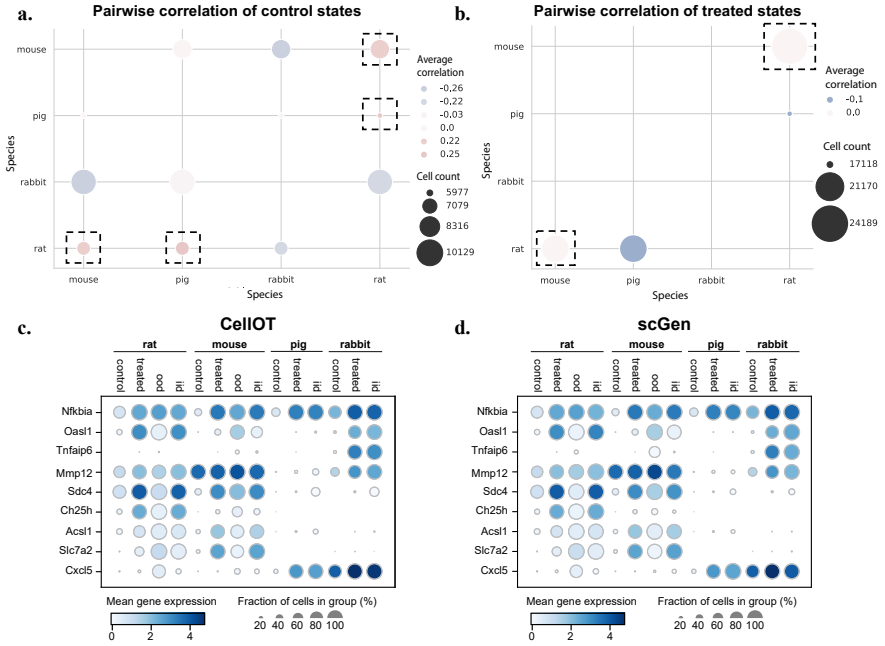


Figure A.2: Analysis and further results of the cross species dataset. **a.** Pairwise average correlation of the PCA embeddings of the control states between species. **b.** Pairwise average correlation of the PCA embeddings of the treated states between patients, masked to only those patient pairs that showed a positive correlation in the control states. Only rat and mouse show consistent responses, i.e., a positive correlation of the control states and non-negative correlation of the respective target cells, and are thus chosen for the o.o.d. analysis. I.i.d. and o.o.d. results measured in the average gene expression for both **d.** CellOT and **c.** scGen.

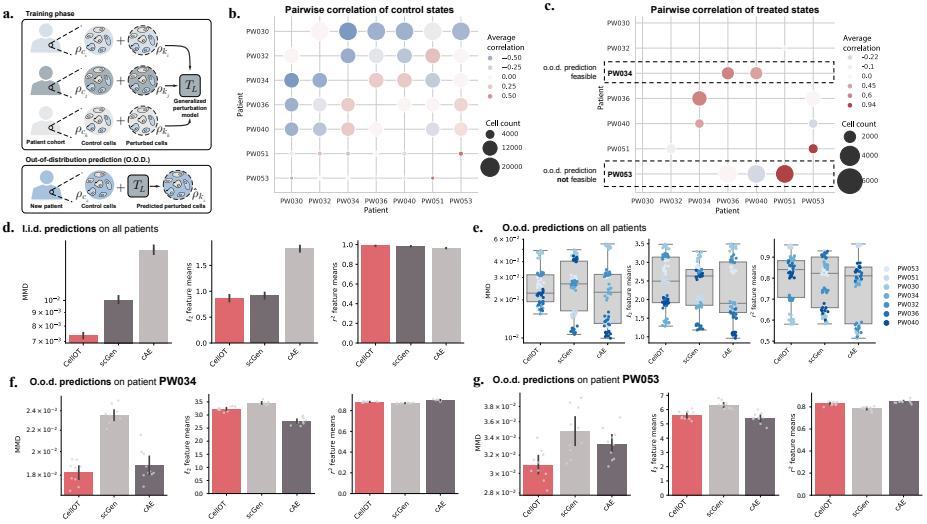


Figure A.3: Analysis and results of the glioblastoma dataset consisting of seven patients. **a.** Cells from seven glioblastoma patients are measured in an untreated and Panobinostat-treated state. For each sample, we train two models, an o.o.d. model trained on cells from all other samples but the holdout patient we test on and an i.i.d. model trained with additional access to half of the cells in the holdout sample. **b.** Pairwise average correlation of the PCA embeddings of the control states between patients. **c.** Pairwise average correlation of the PCA embeddings of the treated states between patients, masked to only those patient pairs that showed a positive correlation in the control states. Only patient PW034 positively correlates with all other patients. Other patients, such as PW053, correlate and anti-correlate with other patients in the treated state. Performance comparison between CellOT and baselines for different metrics in the **d.** i.i.d. setting (mean standard deviation across 7 samples, 10 bootstraps of the test set per sample), **e.** o.o.d. setting for all patients (box plots show median, minima, and maxima) **f.** o.o.d. setting for a patient positively correlating with all patients that are also similar in the control state, **g.** o.o.d. setting for a patient where similar patients in the control state show different responses (correlation and anti-correlation) in the treated states. Data in **f** and **g** are presented as the mean +/- standard deviation across n=10 bootstraps of the test set.

A.2 PROOF OF THEOREM 1

It is known that, for SBs, the optimal solution can be searched within the class of stochastic processes (Léonard, 2013)

$$X_t \sim \mathbb{P}_t : \quad \mathrm{d}X_t = (f_t(X_t) + w_t(X_t))\,\mathrm{d}t + g_t\,\mathrm{d}\mathbb{W}_t. \tag{A.2.1}$$

The Fokker-Planck equation for the SDE (A.2.1) is

$$\partial_t \rho_t = -\nabla_x \cdot (\rho_t(f_t + w_t)) + \frac{g_t^2}{2}\Delta\rho_t. \tag{A.2.2}$$

A simple application of the Girsanov's theorem then shows, up to a constant,

$$D_{\mathrm{KL}}\mathbb{P}_t Y_t = \mathbb{E}\left[\int_0^1 \frac{\|w_t\|^2}{2g_t^2}\,\mathrm{d}t\right]. \tag{A.2.3}$$

Using a change of variable $v_t = w_t - \frac{g_t^2}{2}\nabla \log \rho_t$, we see that (A.2.2) is equivalent to

$$\partial_t \rho_t = -\nabla_x \cdot (\rho_t(f_t + v_t)). \tag{A.2.4}$$

On the other hand, since $\|w_t\|^2 = \|v_t\|^2 + \frac{g_t^4}{4}\|\nabla \log \rho_t\|^2 + 2\left\langle v_t, \frac{g_t^2}{2}\nabla \log \rho_t \right\rangle$, the integrand in the objective of (A.2.3) becomes

$$\mathbb{E}\left[\int_0^1 \frac{\|v_t\|^2}{2g_t^2} + \frac{g_t^2}{8}\|\nabla \log \rho_t\|^2 + \frac{1}{2}\langle v_t, \nabla \log \rho_t\rangle\,\mathrm{d}t\right]. \tag{A.2.5}$$

Letting $H(\rho_t) := \int \rho_t \log \rho_t$ be the entropy, we have

$$\begin{aligned}
H(\rho_1) - H(\rho_0) &= \int_0^1 \partial_t H(\rho_t)\,\mathrm{d}t \\
&= \int_0^1 \int (1 + \log \rho_t)\partial_t \rho_t \mathrm{d}x\,\mathrm{d}t \\
&= \int_0^1 \int (1 + \log \rho_t) \cdot (-\nabla_x \cdot (\rho_t(f_t + v_t)))\mathrm{d}x\,\mathrm{d}t \qquad \text{by (A.2.2)} \\
&= \int_0^1 \int \rho_t \langle \nabla \log \rho_t, f_t + v_t\rangle \mathrm{d}x\,\mathrm{d}t
\end{aligned}$$

by integration by parts for the divergence operator. Therefore,

$$\mathbb{E}\left[\int_0^1 \langle \nabla \log \rho_t, v_t\rangle\,\mathrm{d}t\right] = H(\rho_1) - H(\rho_0) - \mathbb{E}\left[\int_0^1 \langle \nabla \log \rho_t, f_t\rangle\,\mathrm{d}t\right] \tag{A.2.6}$$

which concludes the proof. □

A.3 THE BURES-WASSERSTEIN GEOMETRY OF GAUSSIAN SCHRÖDINGER BRIDGES

A.3.1 *Review of Bures-Wasserstein Geometry*

Recall that the *metric tensor* $\langle \cdot, \cdot \rangle_\Sigma$ in the *Bures-Wasserstein geometry* (Takatsu, 2010) is defined in terms of the Lyapunov operator:

$$\forall\, U, V \in \mathcal{T}_\Sigma \mathbb{S}^d_{++}, \quad \langle U, V \rangle_\Sigma := \operatorname{tr} \mathcal{L}_\Sigma[U]\Sigma\mathcal{L}_\Sigma[V] = \frac{1}{2} \operatorname{tr} \mathcal{L}_\Sigma[U]V. \tag{A.3.1}$$

The corresponding Bures-Wasserstein norm is induced via $\|U\|^2_\Sigma := \langle U, U \rangle_\Sigma$. Another important operator is the Bures-Wasserstein *gradient*: For any function $F \colon \mathbb{S}^d_{++} \to \mathbb{R}$,

$$\mathcal{T}_\Sigma \mathbb{S}^d_{++} \ni \operatorname{grad} F(\Sigma) := 2\Big(\nabla F(\Sigma)\Sigma + \Sigma \nabla F(\Sigma)^\top\Big) \tag{A.3.2}$$

where ∇ is the usual Euclidean gradient of F, viewed as a function from $\mathbb{R}^{d\times d}$ to $\mathbb{R}$. Note that

$$\begin{aligned} \mathcal{L}_\Sigma[\operatorname{grad} F(\Sigma)] &= 2\mathcal{L}_\Sigma[\nabla F(\Sigma)\Sigma + \Sigma\nabla F(\Sigma)] && \text{(A.3.3)} \\ &= 2\nabla F(\Sigma) && \text{(A.3.4)} \end{aligned}$$

by definition of the Lyapunov operator. In other words,

$$\operatorname{grad} F(\Sigma) = \mathcal{L}_\Sigma^{-1}[2\nabla F]. \tag{A.3.5}$$

Lastly, we recall the Bures-Wasserstein *acceleration* of a curve $\Sigma_t : [0,1] \to \mathbb{S}^d_{++}$, which we denote by $\nabla_{\dot\Sigma_t}\dot\Sigma_t$:[1]

$$\nabla_{\dot\Sigma_t}\dot\Sigma_t = \ddot\Sigma_t - \Big(\mathcal{L}_{\Sigma_t}[\dot\Sigma_t]\dot\Sigma_t + \dot\Sigma_t\mathcal{L}_{\Sigma_t}[\dot\Sigma_t]\Big) + \Big(\Sigma_t\Big(\mathcal{L}_{\Sigma_t}[\dot\Sigma_t]\Big)^2 + \Big(\mathcal{L}_{\Sigma_t}[\dot\Sigma_t]\Big)^2\Sigma_t\Big). \tag{A.3.6}$$

A.3.2 *Proof of Theorem 2*

For convenience, we restate Theorem 2 in full below:

1 More formally, $\nabla_{\dot\Sigma_t}\dot\Sigma_t$ is the Bures-Wasserstein covariant derivative of $\dot\Sigma_t$ in the direction of $\dot\Sigma_t$.

Theorem 2. *The minimizer of* (7.18) *(and hence* (7.6)*) coincides with the solution of the action minimization problem:*

$$\min_{\Sigma_0=\Sigma,\Sigma_1=\Sigma'} \int_0^1 \frac{1}{2}\|\dot{\Sigma}_t\|^2_{\Sigma_t} - \mathcal{U}_\sigma(\Sigma_t)\,\mathrm{d}t \tag{7.19}$$

where $\mathcal{U}_\sigma(\Sigma_t) := -\frac{\sigma^4}{8}\operatorname{tr}\Sigma_t^{-1}$ *and the minimum is taken over all piecewise smooth curves in* $\mathbb{S}^d_{++}$*. In particular, the minimizer of* (7.18) *solves the Euler-Lagrange equation in the Bures-Wasserstein geometry:*

$$\nabla_{\dot{\Sigma}_t}\dot{\Sigma}_t = -\operatorname{grad}\mathcal{U}_\sigma(\Sigma_t), \quad \Sigma_0 = \Sigma, \quad \Sigma_1 = \Sigma', \tag{7.20}$$

where $\nabla_{\dot{\Sigma}_t}\dot{\Sigma}_t$ *denotes the Riemannian acceleration and* grad *the Riemannian gradient in the Bures-Wasserstein sense.*

The proof consists of verifying the Euler-Lagrange equation (7.20) for the curve (7.21).

A.3.2.1 *Verifying the Euler-Lagrange Equation* (7.20)

We begin by noting that the boundary conditions in (7.20) hold for the curve in (7.21).
We now compute the two sides of (7.20) separately:

THE RIGHT-HAND SIDE OF (7.20): $-\operatorname{grad}\mathcal{U}_\sigma(\Sigma_t)$. Since $\nabla\mathcal{U}_\sigma(\Sigma_t) = -\nabla\left(\operatorname{tr}\frac{\sigma^4}{8}\Sigma_t^{-1}\right) = \frac{\sigma^4}{8}\Sigma_t^{-1}\cdot\Sigma_t^{-1}$, we see from (A.3.2) that the negative Bures-Wasserstein gradient of $\mathcal{U}_\sigma(\Sigma_t)$ is

$$\begin{aligned} -\operatorname{grad}\mathcal{U}_\sigma(\Sigma_t) &= -2\left(\frac{\sigma^4}{8}\Sigma_t^{-1}\cdot\Sigma_t^{-1}\cdot\Sigma_t + \Sigma_t\cdot\frac{\sigma^4}{8}\Sigma_t^{-1}\cdot\Sigma_t^{-1}\right) \\ &= -\frac{\sigma^4}{2}\Sigma_t^{-1}. \end{aligned} \tag{A.3.7}$$

THE LEFT-HAND SIDE OF (7.20): $\nabla_{\dot{\Sigma}_t}\dot{\Sigma}_t$. Computing $\nabla_{\dot{\Sigma}_t}\dot{\Sigma}_t$ is significantly trickier than $-\operatorname{grad}\mathcal{U}_\sigma(\Sigma_t)$. The central piece of the proof is the following technical lemma:

Lemma A.3.1. *Define the matrix* $\tilde{S}_t$ *to be:*

$$\tilde{S}_t := t\Sigma\prime + \bar{t}C_\sigma - \bar{t}\Sigma - tC_\sigma^\top + \frac{\sigma^2}{2}(\bar{t}-t)I. \tag{A.3.8}$$

Then $\mathcal{L}_{\Sigma_t}[\dot{\Sigma}_t] = \tilde{S}_t^\top \Sigma_t^{-1}$. *In other words,* $\tilde{S}_t^\top \Sigma_t^{-1}$ *is symmetric and solves the Lyapunov equation:*

$$A: \quad A\Sigma_t + \Sigma_t A = \dot{\Sigma}_t. \tag{A.3.9}$$

Moreover, $\tilde{S}_t$ *satisfies the following identity:*

$$\dot{\tilde{S}}_t - \Sigma_t^{-1}\tilde{S}_t^2 = -\frac{\sigma^4}{4}\Sigma_t^{-1}. \tag{A.3.10}$$

Before commencing the proof of Lemma A.3.1, let us show how it readily leads us to (7.20).
Recall the definition of $\nabla_{\dot{\Sigma}_t}\dot{\Sigma}_t$ in (A.3.6). First, note that, by (7.21) and (A.3.8),

$$\begin{aligned}\frac{1}{2}\ddot{\Sigma}_t &= \Sigma + \Sigma\prime - \left(C_\sigma + C_\sigma^\top + \sigma^2 I\right) &\text{(A.3.11)}\\ &= \dot{\tilde{S}}_t. &\text{(A.3.12)}\end{aligned}$$

On the other hand, Lemma A.3.1 entails that

$$\begin{aligned}\Sigma_t\left(\mathcal{L}_{\Sigma_t}[\dot{\Sigma}_t]\right)^2 + \left(\mathcal{L}_{\Sigma_t}[\dot{\Sigma}_t]\right)^2\Sigma_t &= \Sigma_t\mathcal{L}_{\Sigma_t}[\dot{\Sigma}_t]\cdot\mathcal{L}_{\Sigma_t}[\dot{\Sigma}_t] + \mathcal{L}_{\Sigma_t}[\dot{\Sigma}_t]\cdot\mathcal{L}_{\Sigma_t}[\dot{\Sigma}_t]\Sigma_t\\ &= \Sigma_t\Sigma_t^{-1}\tilde{S}_t\cdot\mathcal{L}_{\Sigma_t}[\dot{\Sigma}_t] + \mathcal{L}_{\Sigma_t}[\dot{\Sigma}_t]\cdot\tilde{S}_t^\top\Sigma_t^{-1}\Sigma_t\\ &= \tilde{S}_t\mathcal{L}_{\Sigma_t}[\dot{\Sigma}_t] + \mathcal{L}_{\Sigma_t}[\dot{\Sigma}_t]\tilde{S}_t^\top. &\text{(A.3.13)}\end{aligned}$$

By noting, again from Lemma A.3.1,

$$\begin{aligned}\dot{\Sigma}_t &= \tilde{S}_t^\top\Sigma_t^{-1}\cdot\Sigma_t + \Sigma_t\cdot\Sigma_t^{-1}\tilde{S}_t\\ &= \tilde{S}_t + \tilde{S}_t^\top, &\text{(A.3.14)}\end{aligned}$$

we thus get

$$\begin{aligned}&\Sigma_t\left(\mathcal{L}_{\Sigma_t}[\dot{\Sigma}_t]\right)^2 + \left(\mathcal{L}_{\Sigma_t}[\dot{\Sigma}_t]\right)^2\Sigma_t - \left(\mathcal{L}_{\Sigma_t}[\dot{\Sigma}_t]\dot{\Sigma}_t + \dot{\Sigma}_t\mathcal{L}_{\Sigma_t}[\dot{\Sigma}_t]\right)\\ &\quad= \left(\tilde{S}_t - \dot{\Sigma}_t\right)\mathcal{L}_{\Sigma_t}[\dot{\Sigma}_t] + \mathcal{L}_{\Sigma_t}[\dot{\Sigma}_t]\left(\tilde{S}_t^\top - \dot{\Sigma}_t\right) &\text{(A.3.15)}\\ &\quad= -\left(\tilde{S}_t^\top\mathcal{L}_{\Sigma_t}[\dot{\Sigma}_t] + \mathcal{L}_{\Sigma_t}[\dot{\Sigma}_t]\tilde{S}_t\right)\end{aligned}$$

by (A.3.14). But $\tilde{S}_t^\top\mathcal{L}_{\Sigma_t}[\dot{\Sigma}_t] = \tilde{S}_t^\top\cdot\Sigma_t^{-1}\tilde{S}_t = \Sigma_t^{-1}\tilde{S}_t^2$ by symmetry of $\tilde{S}_t^\top\Sigma_t^{-1}$ and, similarly, we have $\mathcal{L}_{\Sigma_t}[\dot{\Sigma}_t]\tilde{S}_t = \Sigma_t^{-1}\tilde{S}_t^2$. As a result, (A.3.6) reduces to

$$\nabla_{\dot{\Sigma}_t}\dot{\Sigma}_t = 2\dot{\tilde{S}}_t - 2\Sigma_t^{-1}\tilde{S}_t^2. \tag{A.3.16}$$

In lieu of (7.20), (A.3.7), and (A.3.16), the proof of (7.19) can thus be reduced to showing

$$2\dot{\tilde{S}}_t - 2\Sigma_t^{-1}\tilde{S}_t^2 = -\frac{\sigma^4}{2}\Sigma_t^{-1} \tag{A.3.17}$$

which is exactly (A.3.10).

Proof of Lemma A.3.1. We now prove Lemma A.3.1. We begin by proving some useful identities that will inspire our proof for the general GSBs in Section 7.2.4.

USEFUL IDENTITIES. First, note that the definition of C_σ immediately implies $C_\sigma\Sigma = \Sigma C_\sigma^\top$. In addition, we have

$$\begin{aligned} C_\sigma^{-1}\Sigma &= 2\left(\Sigma^{\frac{1}{2}} D_\sigma \Sigma^{-\frac{1}{2}} - \sigma^2 I\right)^{-1}\Sigma \\ &= 2\left(\Sigma^{-\frac{1}{2}} D_\sigma \Sigma^{-\frac{1}{2}} - \sigma^2 \Sigma^{-1}\right)^{-1} \\ &= \Sigma C_\sigma^{-\top}. \end{aligned} \tag{A.3.18}$$

Recall from (Janati et al., 2020b) that C_σ solves the following matrix equation:

$$C_\sigma^2 + \sigma^2 C_\sigma = \Sigma\Sigma\prime. \tag{A.3.19}$$

We therefore have

$$\begin{aligned} C_\sigma &= C_\sigma^{-1}\Sigma\Sigma\prime - \sigma^2 I, \\ C_\sigma^\top &= \Sigma\prime\Sigma C_\sigma^{-\top} - \sigma^2 I, \end{aligned}$$

which, together with (A.3.18), implies

$$\begin{aligned} C_\sigma^\top \Sigma\prime &= \Sigma\prime\Sigma C_\sigma^{-\top}\Sigma\prime - \sigma^2\Sigma\prime \\ &= \Sigma\prime C_\sigma^{-1}\Sigma\Sigma\prime - \sigma^2\Sigma\prime \\ &= \Sigma\prime C_\sigma. \end{aligned} \tag{A.3.20}$$

Now, set $\tilde{S}_t = P_t - Q_t^\top + \frac{\sigma^2}{2}(\bar{t} - t)I$ where

$$P_t := t\Sigma\prime + \bar{t}C_\sigma, \quad Q_t := \bar{t}\Sigma + tC_\sigma. \tag{A.3.21}$$

Note that, by (A.3.20),

$$\begin{aligned}\Sigma\prime P_t^{-1} &= \left(P_t\Sigma\prime^{-1}\right)^{-1}\\ &= \left(tI + \bar{t}C_\sigma\Sigma\prime^{-1}\right)^{-1}\\ &= \left(tI + \bar{t}\Sigma\prime^{-1}C_\sigma^\top\right)^{-1}\\ &= \left(\Sigma\prime^{-1}P_t^\top\right)^{-1}\\ &= P_t^{-\top}\Sigma\prime. \end{aligned} \tag{A.3.22}$$

A similar calculation leading to (A.3.22) shows

$$Q_t^{-1}\Sigma = \Sigma Q_t^{-\top}. \tag{A.3.23}$$

PROOF OF SYMMETRY OF $\tilde{S}_t^\top\Sigma_t^{-1}$. We get, by (A.3.19) and (A.3.20),

$$\begin{aligned}P_t^2 + \sigma^2\bar{t}P_t &= t^2\Sigma\prime^2 + \bar{t}^2C_\sigma^2 + t\bar{t}(\Sigma\prime C_\sigma + C_\sigma\Sigma\prime) + \sigma^2t\bar{t}\Sigma\prime + \sigma^2\bar{t}^2C_\sigma\\ &= t^2\Sigma\prime^2 + \bar{t}^2\left(C_\sigma^2 + \sigma^2C_\sigma\right) + t\bar{t}\left(C_\sigma^\top\Sigma\prime + C_\sigma\Sigma\prime\right) + \sigma^2t\bar{t}\Sigma\prime\\ &= t^2\Sigma\prime^2 + \bar{t}^2\Sigma\Sigma\prime + t\bar{t}\left(C_\sigma^\top + C_\sigma + \sigma^2I\right)\Sigma\prime = \Sigma_t\Sigma\prime. \end{aligned} \tag{A.3.24}$$

It then follows from (A.3.24) that

$$P_t = \Sigma_t\Sigma\prime P_t^{-1} - \sigma^2\bar{t}I, \tag{A.3.25}$$

$$P_t^\top = P_t^{-\top}\Sigma\prime\Sigma_t - \sigma^2\bar{t}I. \tag{A.3.26}$$

As a result, we get, by (A.3.22) and (A.3.25)-(A.3.26),

$$\begin{aligned}\Sigma_t^{-1}P_t &= \Sigma\prime P_t^{-1} - \sigma^2\bar{t}\Sigma_t^{-1}\\ &= P_t^{-\top}\Sigma\prime - \sigma^2\bar{t}\Sigma_t^{-1}\\ &= P_t^\top\Sigma_t^{-1}. \end{aligned} \tag{A.3.27}$$

In exactly the same vein, we have

$$Q_t^2 + \sigma^2tQ_t = \Sigma\Sigma_t \tag{A.3.28}$$

as well as

$$\Sigma_t^{-1} Q_t^\top = Q_t \Sigma_t^{-1}. \tag{A.3.29}$$

The symmetry of $\tilde{S}_t^\top \Sigma_t^{-1}$ is then an immediate consequence of (A.3.27) and (A.3.29). In addition, we have

$$\begin{aligned} \dot{\Sigma}_t &= 2t\Sigma\prime - 2\bar{t}\Sigma + (\bar{t} - t)\left(C_\sigma + C_\sigma^\top + \sigma^2 I\right) \\ &= \tilde{S}_t + \tilde{S}_t^\top. \end{aligned} \tag{A.3.30}$$

Combining the symmetry of $\tilde{S}_t^\top \Sigma_t^{-1}$ and (A.3.30), we see that

$$\tilde{S}_t^\top \Sigma_t^{-1} \cdot \Sigma_t + \Sigma_t \cdot \Sigma_t^{-1} \tilde{S}_t = \tilde{S}_t + \tilde{S}_t^\top = \dot{\Sigma}_t,$$

i.e., $\mathcal{L}_{\Sigma_t}[\dot{\Sigma}_t] = \tilde{S}_t^\top \Sigma_t^{-1}$.

PROOF OF (A.3.10) We next compute

$$\begin{aligned} P_t Q_t^\top &= (t\Sigma\prime + \bar{t} C_\sigma)\left(\bar{t}\Sigma + t C_\sigma^\top\right) \\ &= t\bar{t}\Sigma\prime\Sigma + t^2 \Sigma\prime C_\sigma^\top + \bar{t}^2 C_\sigma \Sigma + t\bar{t} C_\sigma C_\sigma^\top \\ &= \bar{t}^2 \Sigma C_\sigma^\top + t^2 \Sigma\prime C_\sigma^\top + t\bar{t}\left(C_\sigma^{\top 2} + \sigma^2 C_\sigma^\top\right) + t\bar{t} C_\sigma C_\sigma^\top \\ &= \Sigma_t C_\sigma^\top \end{aligned} \tag{A.3.31}$$

where we have used (A.3.18) in the third equality of (A.3.31). A similar computation further shows

$$Q_t^\top P_t = \Sigma_t C_\sigma. \tag{A.3.32}$$

We thus get, by combining (A.3.24) (A.3.28)

$$\begin{aligned}\tilde{S}_t^2 &= P_t^2 - P_t Q_t^\top + \frac{\sigma^2}{2}(\bar{t}-t)P_t - Q_t^\top P_t + Q_t^{\top 2} - \frac{\sigma^2}{2}(\bar{t}-t)Q_t^\top + \frac{\sigma^2}{2}(\bar{t}-t)P_t \\ &\quad - \frac{\sigma^2}{2}(\bar{t}-t)Q_t^\top + \frac{\sigma^4}{4}(\bar{t}-t)^2 I \qquad \text{(A.3.33)} \\ &= P_t^2 + \sigma^2(\bar{t}-t)P_t + Q_t^{\top 2} - \sigma^2(\bar{t}-t)Q_t^\top - \left(P_t Q_t^\top + Q_t^\top P_t\right) + \frac{\sigma^4}{4}(\bar{t}-t)^2 I \\ &= \Sigma_t \Sigma\prime - \sigma^2 t P_t + \Sigma_t \Sigma - \sigma^2 \bar{t} Q_t^\top - \left(\Sigma_t C_\sigma^\top + \Sigma_t C_\sigma\right) + \frac{\sigma^4}{4}(\bar{t}-t)^2 I - \sigma^2 \Sigma_t + \sigma^2 \Sigma_t \\ &= \Sigma_t \left(\Sigma + \Sigma\prime - \left(C_\sigma + C_\sigma^\top + \sigma^2 I\right)\right) + \sigma^2\left(\Sigma_t - t P_t - \bar{t} Q_t^\top\right) + \frac{\sigma^4}{4}(\bar{t}-t)^2 I \\ &= \Sigma_t \dot{\tilde{S}}_t + \sigma^2 \cdot t\bar{t}\sigma^2 I + \frac{\sigma^4}{4}(\bar{t}-t)^2 I \\ &= \Sigma_t \dot{\tilde{S}}_t + \frac{\sigma^4}{4} I \qquad \text{(A.3.34)}\end{aligned}$$

where the third equality follows from (A.3.24), (A.3.28), and (A.3.31)-(A.3.32), and the fifth equality follows from (A.3.8). Multiplying both sides of (A.3.34) by Σ_t^{-1} from the right yields the desired (A.3.10). □

A.3.2.2 *Equivalence between* (7.18) *and* (7.19)

We first note that, by (A.3.1) and Lemma A.3.1,

$$\begin{aligned}\frac{1}{2}\|\dot{\Sigma}_t\|_{\Sigma_t}^2 &= \frac{1}{2}\operatorname{tr} \mathcal{L}_{\Sigma_t}[\dot{\Sigma}_t]\Sigma_t \mathcal{L}_{\Sigma_t}[\dot{\Sigma}_t] \\ &= \frac{1}{2}\operatorname{tr} \tilde{S}_t^\top \Sigma_t^{-1} \cdot \Sigma_t \cdot \Sigma_t^{-1} \tilde{S}_t \\ &= \frac{1}{2}\operatorname{tr} \tilde{S}_t^\top \Sigma_t^{-1} \tilde{S}_t, \qquad \text{(A.3.35)}\end{aligned}$$

and therefore the integrand in (7.19) is equal to

$$\operatorname{tr}\left(\frac{1}{2}\tilde{S}_t^\top \Sigma_t^{-1}\tilde{S}_t + \frac{\sigma^4}{8}\Sigma_t^{-1}\right). \qquad \text{(A.3.36)}$$

To proveed, we will need another formulation of (7.18), which is (Chen et al., 2016; Gentil et al., 2017) specialized to our case:

Lemma A.3.2. *Let* $\mathcal{N}_0 := \mathcal{N}(0, \Sigma)$ *and* $\mathcal{N}_1 := \mathcal{N}(0, \Sigma\prime)$. *Then* (7.18) *is equivalent to*

$$\min_{\rho_0 = \mathcal{N}_0, \rho_1 = \mathcal{N}_1} \int_0^1 \mathbb{E}\left[\frac{1}{2}\|\nabla \Phi_t\|^2 + \frac{\sigma^4}{8}\|\nabla \log \rho_t\|^2\right] \mathrm{d}t \qquad \text{(A.3.37)}$$

where the minimization is taken over all pairs $(\rho_t, \nabla\Phi_t)$ *such that* $\Phi_t : \mathbb{R}^d \to \mathbb{R}$ *are differentiable functions and the continuity equation holds:*

$$\partial_t \rho_t = -\nabla_x \cdot (\rho_t \nabla \Phi_t). \tag{A.3.38}$$

We will also need the *Jacobi formula*: Let $A(t) \colon \mathbb{R}^+ \to \mathbb{R}^{d\times d}$ be a differentiable matrix-valued function. Then

$$\frac{\mathrm{d}}{\mathrm{d}t} \det A(t) = \det A(t) \cdot \operatorname{tr} A^{-1}(t) \cdot \frac{\mathrm{d}}{\mathrm{d}t} A(t). \tag{A.3.39}$$

We are now ready to finish the proof of Theorem 2. By Léonard (2013), the optimal curve for (A.3.37) is Gaussian with zero mean. We denote by Σ_t the covariance of the solution at time t. By (A.3.39), we have

$$\begin{aligned}
\partial_t \rho_t(x) &= \partial_t \left((2\pi)^{\frac{d}{2}} (\det \Sigma_t)^{-\frac{1}{2}} \exp\left(-\frac{1}{2} x^\top \Sigma_t^{-1} x \right) \right) \\
&= (2\pi)^{\frac{d}{2}} \left(-\frac{1}{2} (\det \Sigma_t)^{-\frac{3}{2}} \right) \cdot \det \Sigma_t \cdot \operatorname{tr} \Sigma_t^{-1} \dot{\Sigma}_t \exp\left(-\frac{1}{2} x^\top \Sigma_t^{-1} x \right) \\
&\qquad + (2\pi)^{\frac{d}{2}} (\det \Sigma_t)^{-\frac{1}{2}} \exp\left(-\frac{1}{2} x^\top \Sigma_t^{-1} x \right) \cdot \left(\frac{1}{2} x^\top \Sigma_t^{-1} \dot{\Sigma}_t \Sigma_t^{-1} x \right) \\
&= \rho_t(x) \cdot \left(\frac{1}{2} x^\top \Sigma_t^{-1} \dot{\Sigma}_t \Sigma_t^{-1} x - \frac{1}{2} \operatorname{tr} \Sigma_t^{-1} \dot{\Sigma}_t \right).
\end{aligned} \tag{A.3.40}$$

On the other hand, by the chain rule for the divergence, we have

$$\nabla_x \cdot (\rho_t \nabla \Phi_t) = \langle \nabla \rho_t, \nabla \Phi_t \rangle + \rho_t \Delta \Phi_t. \tag{A.3.41}$$

Since $\nabla \rho_t = \rho_t \left(-\Sigma_t^{-1} x \right)$, the continuity equation (A.3.38) together with (A.3.40)-(A.3.41) implies that Σ_t must satisfy

$$\Delta \Phi_t = \frac{1}{2} \operatorname{tr} \Sigma_t^{-1} \dot{\Sigma}_t, \tag{A.3.42}$$

$$\langle \Sigma_t^{-1} x, \nabla \Phi_t(x) \rangle = \frac{1}{2} \langle \Sigma_t^{-1} x, \dot{\Sigma}_t \Sigma_t^{-1} x \rangle, \quad \forall x \in \mathbb{R}^d. \tag{A.3.43}$$

In other words, the optimal vector field is of the form $\nabla \Phi_t(x) = \tilde{S}_t^\top \Sigma_t^{-1} x$ for some matrix $\tilde{S}_t$ such that

$$\operatorname{tr} \tilde{S}_t^\top \Sigma_t^{-1} = \frac{1}{2} \operatorname{tr} \dot{\Sigma}_t \Sigma_t^{-1}, \tag{A.3.44}$$

$$\operatorname{tr} \Sigma_t^{-1} \tilde{S}_t^\top \Sigma_t^{-1} x x^\top = \frac{1}{2} \operatorname{tr} \Sigma_t^{-1} \dot{\Sigma}_t \Sigma_t^{-1} x x^\top, \quad \forall x \in \mathbb{R}^d. \tag{A.3.45}$$

Therefore, we see that

$$\begin{aligned}\mathbb{E}\left[\|\nabla\Phi_t\|^2\right] &= \mathbb{E}\left[\operatorname{tr}\tilde{S}_t^\top\Sigma_t^{-1}xx^\top\Sigma_t^{-1}\tilde{S}_t\right]\\ &= \operatorname{tr}\tilde{S}_t^\top\Sigma_t^{-1}\,\mathbb{E}[xx^\top]\Sigma_t^{-1}\tilde{S}_t\\ &= \operatorname{tr}\tilde{S}_t^\top\Sigma_t^{-1}\tilde{S}_t. \end{aligned}\tag{A.3.46}$$

Furthermore, we have

$$\begin{aligned}\mathbb{E}\left[\|\nabla\log\rho_t\|^2\right] &= \mathbb{E}\left[\operatorname{tr}\Sigma_t^{-1}xx^\top\Sigma_t^{-1}\right]\\ &= \operatorname{tr}\Sigma_t^{-1}. \end{aligned}\tag{A.3.47}$$

Finally, since the optimal vector field $\nabla\Phi_t$ is a gradient field, we must have $\tilde{S}_t^\top\Sigma_t^{-1} = \Sigma_t^{-1}\tilde{S}_t$. Combing all the above, we see that (A.3.37) is equivalent to

$$\min_{\substack{\Sigma_0=\Sigma,\Sigma_1=\Sigma\prime\\ \tilde{S}_t^\top\Sigma_t^{-1}=\Sigma_t^{-1}\tilde{S}_t}}\int_0^1\operatorname{tr}\left(\frac{1}{2}\tilde{S}_t^\top\Sigma_t^{-1}\tilde{S}_t+\frac{\sigma^4}{8}\Sigma_t^{-1}\right)\mathrm{d}t \tag{A.3.48}$$

which, in view of (A.3.36), is exactly the same as (7.19).

A.3.3 *Some Interesting Consequences of Theorem 2*

Here, we collect some interesting corollaries of Theorem 2, although they will not be used in the rest of the paper.

A.3.3.1 *Conservation of Hamiltonian*

The first result concerns the *Hamiltonian formulation* of the action minimization problem (7.19).

Corollary 2 (Conservation of Hamiltonian)**.** *Define the* ***Hamiltonian*** *associated with* (7.19) *to be*

$$\begin{aligned}\mathcal{H}(\Sigma_t) &:= \frac{1}{2}\|\dot{\Sigma}_t\|_{\Sigma_t}^2+\mathcal{U}_\sigma(\Sigma_t)\\ &= \operatorname{tr}\left(\frac{1}{2}\tilde{S}_t^\top\Sigma_t^{-1}\tilde{S}_t-\frac{\sigma^4}{8}\Sigma_t^{-1}\right).\end{aligned}\tag{H}$$

Then the Hamiltonian is conserved along Σ_t:

$$\dot{\mathcal{H}}\equiv 0,\ \textit{or, equivalently, } \mathcal{H}(\Sigma_t)=\operatorname{tr}(\Sigma+\Sigma\prime-D_\sigma)\ \text{for all } t. \tag{A.3.49}$$

The fact that the Hamiltonian, commonly interpreted as the *total energy*, is conserved is a well-known fact in physics (Villani, 2009) and directly follows from Theorem 2.

A.3.3.2 *Connection to Fisher Information*

The "potential energy" term $\mathcal{U}_\sigma(\Sigma_t)$ in (7.19) has an interesting origin: It is, up to a constant, the *entropy production rate*, i.e., the *Fisher information*.

Lemma A.3.3. *Let $\rho \sim \mathcal{N}(0, \Sigma)$, and let $H(\Sigma)$ be the (negative) Shannon entropy of ρ. Then*

$$\mathcal{U}_\sigma(\Sigma_t) = \frac{1}{2}\mathcal{I}_\sigma(\Sigma) \tag{A.3.50}$$

where

$$\mathcal{I}_\sigma(\Sigma) := \frac{\sigma^4}{4}\|\operatorname{grad} H(\Sigma)\|_\Sigma^2. \tag{A.3.51}$$

Proof. Recall that $\nabla H(\Sigma) = \nabla\left(-\frac{1}{2}\log\det\Sigma - \frac{d}{2}\log 2\pi e\right) = -\frac{1}{2}\Sigma^{-1}$. Therefore, by (A.3.1) and (A.3.5),

$$\begin{aligned}
\mathcal{I}_\sigma(\Sigma) &= \frac{\sigma^4}{4}\|\operatorname{grad} H(\Sigma)\|_\Sigma^2 \\
&= \frac{\sigma^4}{4}\langle \operatorname{grad} H(\Sigma), \operatorname{grad} H(\Sigma)\rangle_\Sigma \\
&= \frac{\sigma^4}{4}\operatorname{tr}\mathcal{L}_\Sigma[\operatorname{grad} H(\Sigma)]\Sigma\mathcal{L}_\Sigma[\operatorname{grad} H(\Sigma)] \\
&= \frac{\sigma^4}{4}\operatorname{tr}\mathcal{L}_\Sigma[\mathcal{L}_\Sigma^{-1}[2\nabla H(\Sigma)]]\Sigma\mathcal{L}_\Sigma[\mathcal{L}_\Sigma^{-1}[2\nabla H(\Sigma)]] \\
&= \frac{\sigma^4}{4}(-\Sigma^{-1})\Sigma(-\Sigma^{-1}) = \frac{\sigma^4}{4}\Sigma^{-1}.
\end{aligned}$$

□

An infinite-dimensional version of Lemma A.3.3 for non-Gaussian measures is proved in Chen et al. (2016); Gentil et al. (2017); the connection to the Bures-Wasserstein geometry here seems to be new.
The specific form of the potential energy in (A.3.51) has been shown to be intimately related to the *gradient flow* of entropy:

$$\dot{\Sigma}_t = -\operatorname{grad} H(\Sigma). \tag{A.3.52}$$

We refer the interested readers to (Gentil et al., 2020) for details.

A.3.3.3 *Solution of the Schrödinger Systems*

Another way of solving a system of the form (A.3.37) is via the so-called forward *Schrödinger system* (Chen et al., 2021b; Léonard, 2013):

$$\begin{cases} \partial_t \mu_t + \nabla_x \cdot (\mu_t \nabla \Phi_t) = \frac{\sigma^2}{2} \Delta \mu_t \\ \partial_t \Phi_t + \frac{\|\nabla \Phi_t\|^2}{2} + \frac{\sigma^2}{2} \Delta \Phi_t = 0 \end{cases} . \tag{A.3.53}$$

By the various identities we prove in Appendix A.3.2.1, one can easily show that the solution to (A.3.53) is given by

$$\Phi_t(x) = -\frac{\sigma^2}{4} \log \det \Sigma_t + \frac{\sigma^4}{4} \int_0^t \operatorname{tr} \Sigma_t^{-1} \, \mathrm{d}t + \frac{1}{2} \langle x, \left(\tilde{S}_t^\top - \frac{\sigma^2}{2} I \right) \Sigma_t^{-1} x \rangle + \text{const.} \tag{A.3.54}$$

This is in fact the same solution of the *fluid mechanical* problem

$$\min_{\substack{\rho_0 = \mathcal{N}_0, \rho_1 = \mathcal{N}_1 \\ \partial_t \rho_t + \nabla_x \cdot (\rho_t \nabla \Phi_t) = \Delta \rho_t}} \int_0^1 \mathbb{E} \left[\frac{1}{2} \|\nabla \Phi_t\|^2 \right] \mathrm{d}t \tag{A.3.55}$$

which is yet another equivalent formulation of (7.18).
There is also a backward Schrödinger system:

$$\begin{cases} -\partial_t \mu_t + \nabla_x \cdot (\mu_t \nabla \hat{\Phi}_t) = \frac{\sigma^2}{2} \Delta \mu_t \\ -\partial_t \hat{\Phi}_t + \frac{\|\nabla \hat{\Phi}_t\|^2}{2} + \frac{\sigma^2}{2} \Delta \hat{\Phi}_t = 0 \end{cases} , \tag{A.3.56}$$

whose solution is given by

$$\hat{\Phi}_t(x) = -\frac{\sigma^2}{4} \log \det \Sigma_t - \frac{\sigma^4}{4} \int_0^t \operatorname{tr} \Sigma_t^{-1} \, \mathrm{d}t - \frac{1}{2} \langle x, \left(\tilde{S}_t^\top + \frac{\sigma^2}{2} I \right) \Sigma_t^{-1} x \rangle + \text{const.} \tag{A.3.57}$$

Notice that

$$\Phi_t + \hat{\Phi}_t = \sigma^2 \log \rho_t \tag{A.3.58}$$

which is a well-known feature of the solutions to the forward and backward Schrödinger systems (Chen et al., 2021b; Léonard, 2013).

A.4 PROOF OF THE CLOSED-FORM SOLUTIONS FOR GAUSSIAN SCHRÖDINGER BRIDGES

A.4.1 *Preliminaries for the Proof of Theorem 3*

We need a technical lemma that is intimately related to the "central identity of quantum field theory" (Zee, 2010); the version below is adopted from (user26872, 2012), wherein the readers can find an easy proof.

Lemma A.4.1 (The central identity of Quantum Field Theory)**.** *The following identity holds for all matrix $M \succ 0$ and all sufficiently regular analytic function v (e.g., polynomials or $v \in \mathcal{C}^\infty(\mathbb{R}^d)$ with compact support):*

$$(2\pi)^{-\frac{d}{2}}(\det M)^{\frac{1}{2}} \int_{\mathbb{R}^d} v(x) \exp\left(-\frac{1}{2}x^\top M x\right) \mathrm{d}x = \exp\left(\frac{1}{2}\partial_x^\top M^{-1}\partial_x\right) v(x)\Big|_{x=0} \tag{A.4.1}$$

where $\exp\left(\frac{1}{2}\partial_x^\top M^{-1}\partial_x\right)$ is understood as a power series in the differential operators.

Lastly, we recall the elementary

Lemma A.4.2 (Conditional Gaussians are Gaussian)**.** *Let*

$$(Y_0, Y_1) \sim \mathcal{N}\left(\begin{bmatrix}\mu_0\\ \mu_1\end{bmatrix}, \begin{bmatrix}\Sigma_{00} & \Sigma_{01}\\ \Sigma_{10} & \Sigma_{11}\end{bmatrix}\right).$$

Then $Y_0|Y_1 = y \sim \mathcal{N}(\check{\mu}, \check{\Sigma})$ where

$$\begin{aligned}\check{\mu} &= \mu_0 + \Sigma_{01}\Sigma_{11}^{-1}(y - \mu_1),\\ \check{\Sigma} &= \Sigma_{00} - \Sigma_{01}\Sigma_{11}^{-1}\Sigma_{10}.\end{aligned} \tag{A.4.2}$$

A.4.2 *The Proof*

We are now ready for the proof. For convenience, we restate Theorem 3 below:

Theorem 3. *Denote by $\mathbb{P}_t$ the solution to Gaussian Schrödinger bridges (GSB). Set*

$$
\begin{aligned}
&r_t := \frac{\kappa(t,1)}{\kappa(1,1)}, \quad \bar{r}_t := \tau_t - r_t\tau_1, \quad \sigma_\star := \sqrt{\tau_1^{-1}\kappa(1,1)},\\
&\zeta(t) := \tau_t \int_0^t \tau_s^{-1}\alpha_s \,\mathrm{d}s, \rho_t := \frac{\int_0^t \tau_s^{-2} g_s^2 \,\mathrm{d}s}{\int_0^1 \tau_s^{-2} g_s^2 \,\mathrm{d}s},\\
&P_t := \dot{r}_t(r_t\Sigma_1 + \bar{r}_t C_{\sigma_\star}), \quad Q_t := -\dot{\bar{r}}_t(\bar{r}_t\Sigma_0 + r_t C_{\sigma_\star}),\\
&S_t := P_t - Q_t^\top + \Big[c_t\kappa(t,t)(1-\rho_t) - g_t^2\rho_t\Big] I. \qquad (7.27)
\end{aligned}
$$

Then the following holds:

1. *The solution $\mathbb{P}_t$ is a Markov Gaussian process whose marginal variable $X_t \sim \mathcal{N}(\mu_t, \Sigma_t)$, where*

$$
\begin{aligned}
\mu_t &:= \bar{r}_t\mu_0 + r_t\mu_1 + \zeta(t) - r_t\zeta(1), \qquad (7.28)\\
\Sigma_t &:= \bar{r}_t^2\Sigma_0 + r_t^2\Sigma_1 + r_t\bar{r}_t\Big(C_{\sigma_\star} + C_{\sigma_\star}^\top\Big) + \kappa(t,t)(1-\rho_t)I. \qquad (7.29)
\end{aligned}
$$

2. *X_t admits a closed-form representation as the SDE:*

$$
\mathrm{d}X_t = f_{\mathcal{N}}(t, X_t)\,\mathrm{d}t + g_t\,\mathrm{d}W_t \qquad (7.30)
$$

where

$$
f_{\mathcal{N}}(t,x) := S_t^\top \Sigma_t^{-1}(x - \mu_t) + \dot{\mu}_t. \qquad (7.31)
$$

Moreover, the matrix $S_t^\top\Sigma_t^{-1}$ is symmetric.

As the proof is quite complicated, we first outline the main steps below:

1. Leveraging existing results (Bojilov and Galichon, 2016; del Barrio and Loubes, 2020; Janati et al., 2020b; Mallasto et al., 2021), we first solve an appropriately chosen *static* GSB determined by the reference process $\mathbb{Q}_t$.
2. It can be shown from the disintegration formula (Léonard, 2013), the solution of the static GSBs (7.5), and properties of (7.24) that $\mathbb{P}_t$ is a Markov Gaussian process with mean (7.28) and covariance (7.29).
3. Invoking the *generator theory* (Protter, 2005), to prove (7.30), it suffices to show that X_t satisfies, for any sufficiently regular test function $u : \mathbb{R}^+ \times \mathbb{R}^d \to \mathbb{R}$,

$$
\lim_{h\to 0} \frac{\mathbb{E}\big[u\big(t+h, X_{t+h}\big) \,\big|\, X_t = x\big]}{h} = \mathcal{L}_t u(t,x), \qquad (A.4.3)
$$

where

$$\mathcal{L}_t u(t,x) := \frac{\partial}{\partial t} u(t,x) + \frac{g_t^2}{2} \Delta u(t,x) + \langle \nabla u(t,x), f_{\mathcal{N}}(t,x) \rangle \tag{A.4.4}$$

is the generator for the process (7.30).

4. Since the marginal/joint/conditional distributions of a Gaussian process are still Gaussian, the expectation in (A.4.3) requires to express Gaussian integrals as differential operators. To this end, the appropriate tool is the "central identity in quantum field theory" (Zee, 2010).
5. Proof concludes by matching terms in (A.4.3) and (A.4.4). □

Proof of Theorem 3. From now on, we will invoke the notations in (7.27) without explicit mentions.

THE STATIC GAUSSIAN SB. We begin by solving the *static* Gaussian SB

$$\min_{\mathbb{P}_{01}} D_{\mathrm{KL}} \mathbb{P}_{01} \mathbb{Q}_{01} \tag{A.4.5}$$

over all $\mathbb{P}_{01}$ having marginals $\mathcal{N}(\mu_0, \Sigma_0)$ and $\mathcal{N}(\mu_1, \Sigma_1)$.
Recall that, conditioned on Y_0, $Y_t \sim \mathbb{Q}_t$ is a Gaussian process with mean (7.25) and covariance (7.26). Thus, if we only consider the endpoint marginal distributions (Y_0, Y_1), it is easy to derive the transition probability:

$$\begin{aligned} \mathbb{Q}(Y_1 = y_1 | Y_0 = y_0) &= (2\pi)^{\frac{d}{2}} \det(\kappa(1,1) I)^{-\frac{1}{2}} \exp(&& \text{(A.4.6)} \\ &\quad - \frac{1}{2} (y_1 - \eta(1))^\top (\kappa(1,1) I)^{-1} (y_1 - \eta(1))) \\ &= (2\pi)^{\frac{d}{2}} \det(\kappa(1,1) I)^{-\frac{1}{2}} \exp(&& \text{(A.4.7)} \\ &\quad - \frac{1}{2\kappa(1,1)} \| y_1 - \tau_1 y_0 - \zeta(1) \|^2). \end{aligned}$$

Therefore, abusing the notation by continually writing $\mathbb{P}_{01}$ as the relative density of $\mathbb{P}_{01}$ with respect to the Lebesgue measure, we get

$$D_{\mathrm{KL}}\mathbb{P}_{01}\mathbb{Q}_{01} = \int_{\mathbb{R}^d\times\mathbb{R}^d} \log \frac{\mathrm{d}\mathbb{P}_{01}}{\mathrm{d}\mathbb{Q}_{01}}\,\mathrm{d}\mathbb{P}_{01} \tag{A.4.8}$$

$$= \text{const.} + \frac{1}{2\kappa(1,1)} \int_{\mathbb{R}^d\times\mathbb{R}^d} \|y\prime - \tau_1 y - \tau_1\zeta(1)\|^2\,\mathrm{d}\mathbb{P}_{01}(y,y\prime) \tag{A.4.9}$$

$$+ \int_{\mathbb{R}^d\times\mathbb{R}^d} \log \mathbb{P}_{01}\,\mathrm{d}\mathbb{P}_{01}.$$

If $\mathbb{P}_{01}$ is a joint distribution with marginals $Y \sim \mathcal{N}(\mu_0,\Sigma_0)$ and $Y\prime \sim \mathcal{N}(\mu_1,\Sigma_1)$, then the change of variable $\tilde{Y} = \tau_1 Y + \zeta(1)$ gives rise to a joint distribution $\tilde{\mathbb{P}}_{01}$ having marginals $\tilde{Y} \sim \mathcal{N}(\tilde{\mu}_0,\tilde{\Sigma}_0)$ and $Y\prime \sim \mathcal{N}(\mu_1,\Sigma_1)$, where

$$\tilde{\mu}_0 = \tau_1\mu_0 + \zeta(1), \tag{A.4.10}$$

$$\tilde{\Sigma}_0 = \tau_1^2\Sigma_0. \tag{A.4.11}$$

Obviously, there is a one-to-one correspondence between $\mathbb{P}_{01}$ and $\tilde{\mathbb{P}}_{01}$. The first integral in (A.4.9) is equal to $\mathbb{E}\left[\left\|Y\prime - \tilde{Y}\right\|^2\right]$. On the other hand, we always have

$$\int_{\mathbb{R}^d\times\mathbb{R}^d} \log \tilde{\mathbb{P}}_{01}\,\mathrm{d}\tilde{\mathbb{P}}_{01} = \int_{\mathbb{R}^d\times\mathbb{R}^d} \log \mathbb{P}_{01}\,\mathrm{d}\mathbb{P}_{01} + \text{const.}$$

Therefore, minimizing (A.4.8) over $\mathbb{P}_{01}$ is equivalent to

$$\min_{\tilde{\mathbb{P}}_{01}} D_{\mathrm{KL}}\tilde{\mathbb{P}}_{01}\mathbb{Q}_{01} \equiv \min_{\tilde{\mathbb{P}}_{01}} \int_{\mathbb{R}^d\times\mathbb{R}^d} \frac{\|y-y\prime\|_2^2}{2}\,\mathrm{d}\tilde{\mathbb{P}}_{01}(y,y\prime) \tag{A.4.12}$$

$$+ \kappa(1,1)\int_{\mathbb{R}^d\times\mathbb{R}^d} \log \tilde{\mathbb{P}}_{01}\,\mathrm{d}\tilde{\mathbb{P}}_{01}. \tag{A.4.13}$$

By (7.5), the solution to (A.4.12) is given by the joint Gaussian

$$\tilde{\mathbb{P}}_{01}^{\star} \sim \mathcal{N}\left(\begin{bmatrix}\tilde{\mu}_0\\ \mu_1\end{bmatrix}, \begin{bmatrix}\tilde{\Sigma}_0 & \tilde{C}_{\tilde{\sigma}}\\ \tilde{C}_{\tilde{\sigma}}^{\top} & \Sigma_1\end{bmatrix}\right) \tag{A.4.14}$$

where $\tilde{\sigma} = \sqrt{\kappa(1,1)}$ and

$$\tilde{C}_{\tilde{\sigma}} = \frac{1}{2}\left(\tilde{\Sigma}_0^{\frac{1}{2}}\tilde{D}_{\tilde{\sigma}}\tilde{\Sigma}_0^{-\frac{1}{2}} - \tilde{\sigma}^2 I\right), \tag{A.4.15}$$

$$\tilde{D}_{\tilde{\sigma}} = \left(4\tilde{\Sigma}_0^{\frac{1}{2}}\Sigma_1\tilde{\Sigma}_0^{\frac{1}{2}} + \tilde{\sigma}^4 I\right)^{\frac{1}{2}}. \tag{A.4.16}$$

The optimal static Gaussian SB $\mathbb{P}^\star_{01}$ is then given by the inverse transform $Y = \tau_1^{-1}\big(\tilde{Y} - \zeta(1)\big)$, i.e.,

$$\mathbb{P}^\star_{01} \sim \mathcal{N}\left(\begin{bmatrix}\mu_0\\ \mu_1\end{bmatrix}, \begin{bmatrix}\Sigma_0 & \tau_1^{-1}\tilde{C}_{\tilde{\sigma}}\\ \tau_1^{-1}\tilde{C}_{\tilde{\sigma}}^\top & \Sigma_1\end{bmatrix}\right). \tag{A.4.17}$$

Rearranging terms and using (A.4.15) and (A.4.16), we get

$$\tau_1^{-1}\tilde{C}_{\tilde{\sigma}} = C_{\sigma_\star} \tag{A.4.18}$$

where $\sigma_\star = \frac{\kappa(1,1)}{\tau_1}$.

THE $\mathbb{Q}$–BRIDGES For future use, we will need the distribution of Y_t conditioned on Y_0 and Y_1. When $Y_t \equiv \mathbb{W}_t$, the distribution is called the *Brownian bridge*, which is in itself an important subject in mathematics and financial engineering (Mansuy and Yor, 2008). We thus term the conditional distribution of Y_t the $\mathbb{Q}$–*Bridges*.
From (7.25) and (7.26), one can infer that, given Y_0, the joint distribution of (Y_t, Y_1) is

$$Y_t, Y_1 | Y_0 \sim \mathcal{N}\left(\begin{bmatrix}\eta(t)\\ \eta(1)\end{bmatrix}, \begin{bmatrix}\kappa(t,t)I & \kappa(t,1)I\\ \kappa(t,1)I & \kappa(1,1)I\end{bmatrix}\right). \tag{A.4.19}$$

Therefore, Lemma A.4.2 applied implies that, conditioned on Y_0 and Y_1, Y_t is Gaussian with mean

$$\begin{aligned}
\mathbb{E}[Y_t \,|\, Y_0, Y_1] &= \eta(t) + \frac{\kappa(t,1)}{\kappa(1,1)}(Y_1 - \eta(1))\\
&= \tau_t Y_0 + \zeta(t) + \frac{\kappa(t,1)}{\kappa(1,1)}(Y_1 - \tau_1 Y_0 - \zeta(1))\\
&= \left(\tau_t - \frac{\kappa(t,1)}{\kappa(1,1)}\tau_1\right)Y_0 + \frac{\kappa(t,1)}{\kappa(1,1)}Y_1 + \zeta(t) - \frac{\kappa(t,1)}{\kappa(1,1)}\zeta(1)\\
&= \bar{r}_t Y_0 + r_t Y_1 + \zeta(t) - \tau_t\zeta(1)
\end{aligned} \tag{A.4.20}$$

and covariance process (for any $t\prime \geq t$)

$$\begin{aligned}\mathbb{E}\Big[(Y_t - \mathbb{E}[Y_t \,|\, Y_0, Y_1])(Y_{t\prime} - \mathbb{E}[Y_{t\prime} \,|\, Y_0, Y_1])^\top \,\Big|\, Y_0, Y_1\Big] \\ = \left(\kappa(t,t') - \frac{\kappa(t,1)\kappa(t\prime,1)}{\kappa(1,1)}\right) I.\end{aligned} \tag{A.4.21}$$

Since a Gaussian process is uniquely determined by its mean and covariance processes, we have, for some Gaussian process ξ_t independent of Y_t having zero mean and covariance process (A.4.21),

$$Y_t | Y_0, Y_1 \overset{\text{law}}{=} \bar{r}_t Y_0 + r_t Y_1 + \zeta(t) - \tau_t \zeta(1) + \xi_t. \tag{A.4.22}$$

FROM $\mathbb{Q}$–BRIDGES TO μ_t AND Σ_t The disintegration formula of D_{KL}·· (Léonard, 2013) implies that the solution to (GSB) is given by first generating $(X_0, X_1) \sim \mathbb{P}^\star_{01}$ for $\mathbb{P}^\star_{01}$ in (A.4.17), and then connecting X_0 and X_1 using the $\mathbb{Q}$–bridges (A.4.22). Namely,

$$X_t \overset{\text{law}}{=} \bar{r}_t X_0 + r_t X_1 + \zeta(t) - r_t \zeta(t) + \xi_t \tag{A.4.23}$$

from which (7.28) and (7.29) follow by a straightforward calculation. Furthermore, in view of (A.4.17) and (A.4.23), X_t is obviously a Gaussian process. Finally, since $\mathbb{Q}_t$ is a Markov process, (Léonard, 2013, Theorem 2.12) implies that $\mathbb{P}_t$ is also Markov. This concludes the first half of Theorem 3.

THE SDE REPRESENTATION OF X_t The main idea of proving (7.31) is to compute

$$\lim_{h \to 0} \frac{\mathbb{E}\big[u\big(t+h, X_{t+h}\big) \,\big|\, X_t = x\big] - u(t,x)}{h} \tag{A.4.24}$$

and equate (A.4.24) with the *generator* of (7.30), which is (Protter, 2005)

$$\mathcal{L}_t u(t,x) \coloneqq \frac{\partial}{\partial t} u(t,x) + \frac{g_t^2}{2} \Delta u(t,x) + \langle \nabla u(t,x), f_{\mathcal{N}}(t,x) \rangle. \tag{A.4.25}$$

Since X_t is a Gaussian process, we may derive the conditional expectation in (A.4.24) using Lemma A.4.2. However, since eventually we will divide everything by h and drive $h \to 0$, we can ignore any term that is $o(h)$ during the computation. This simple observation will prove to be extremely useful in the sequel.

We first compute the first-order approximation of Σ_t. In view of (7.29), and since $r_t\kappa(t,1) = \kappa(t,t)\rho_t$ and $\dot{r}_t\kappa(t,1) = r_t\frac{\partial}{\partial t}\kappa(t,1)$, we have

$$\begin{aligned}
\dot{\Sigma}_t &= 2\dot{\bar{r}}_t\bar{r}_t\Sigma_0 + 2\dot{r}_t r_t\Sigma_1 + (\dot{r}_t\bar{r}_t + r_t\dot{\bar{r}}_t)\Big(C_{\sigma_\star} + C_{\sigma_\star}^\top\Big) \\
&\qquad + \Big(\frac{\partial}{\partial t}\kappa(t,t) - \dot{r}_t\kappa(t,1) - r_t\frac{\partial}{\partial t}\kappa(t,1)\Big)I \\
&= \dot{r}_t\Big(r_t\Sigma_1 + \bar{r}_t C_{\sigma_\star} + r_t\Sigma_1 + \bar{r}_t C_{\sigma_\star}^\top\Big) + \dot{\bar{r}}_t\Big(\bar{r}_t\Sigma_0 + r_t C_{\sigma_\star} + \bar{r}_t\Sigma_0 + r_t C_{\sigma_\star}^\top\Big) \\
&\qquad + \Big(\frac{\partial}{\partial t}\kappa(t,t) - 2\dot{r}_t\kappa(t,1)\Big)I \\
&= \Big(P_t + P_t^\top\Big) - \Big(Q_t + Q_t^\top\Big) + \Big(\frac{\partial}{\partial t}\kappa(t,t) - 2\dot{r}_t\kappa(t,1)\Big)I. \qquad \text{(A.4.26)}
\end{aligned}$$

Next, let $K_{t,t+h}$ denote the covariance process of X_t. We can estimate $K_{t,t+h}$ up to first order by computing:

$$\begin{aligned}
K_{t,t+h} &:= \mathbb{E}\Big[(X_t - \mu_t)\big(X_{t+h} - \mu_{t+h}\big)^\top\Big] \\
&= \bar{r}_t\bar{r}_{t+h}\Sigma_0 + r_t r_{t+h}\Sigma_1 + \bar{r}_t r_{t+h}C_{\sigma_\star} + r_t\bar{r}_{t+h}C_{\sigma_\star}^\top + \Big(\kappa(t,t+h) \\
&\qquad - r_{t+h}\kappa(t,1)\Big)I \\
&= \Sigma_t + \bar{r}_t(\bar{r}_{t+h} - \bar{r}_t)\Sigma_0 + r_t(r_{t+h} - r_t)\Sigma_1 + \bar{r}_t(r_{t+h} - r_t)C_{\sigma_\star} \\
&\qquad + r_t(\bar{r}_{t+h} - \bar{r}_t)C_{\sigma_\star}^\top + (\kappa(t,t+h) - \kappa(t,t) - r_{t+h}\kappa(t,1) + r_t\kappa(t,1))I \\
&= \Sigma_t + \frac{r_{t+h} - r_t}{\dot{r}_t}P_t - \frac{\bar{r}_{t+h} - \bar{r}_t}{\dot{\bar{r}}_t}Q_t^\top + \Big(\kappa(t,t+h) \\
&\qquad - \kappa(t,t) - r_{t+h}\kappa(t,1) + r_t\kappa(t,1)\Big)I \\
&= \Sigma_t + h\Big\{P_t - Q_t^\top + \Big[\Big(\frac{\partial}{\partial t\prime}\kappa\Big)(t,t) - \dot{r}_t\kappa(t,1)\Big]I\Big\} + o(h), \qquad \text{(A.4.27)}
\end{aligned}$$

where $\big(\frac{\partial}{\partial t\prime}\kappa\big)(t,t\prime) := \lim_{h\to 0}\frac{\kappa(t,t\prime+h) - \kappa(t,t\prime)}{h}$ denotes the derivative of the function $\kappa(t,\cdot)$. Using (7.26) and $\dot{\tau}_t = c_t\tau_t$, we have

$$\begin{aligned}
\Big(\frac{\partial}{\partial t\prime}\kappa\Big)(t,t) &= \frac{\partial}{\partial t\prime}\Big(\tau_t\tau_{t\prime}\int_0^t \tau_s^{-2}g_s^2\,\mathrm{d}s\Big)\Big|_{t\prime=t} \qquad \text{(A.4.28)} \\
&= \dot{\tau}_t\tau_t\int_0^t \tau_s^{-2}g_s^2\,\mathrm{d}s \\
&= c_t\kappa(t,t). \qquad \text{(A.4.29)}
\end{aligned}$$

On the other hand, we have

$$\begin{aligned}\dot{r}_t &= \frac{1}{\kappa(1,1)}\frac{\partial}{\partial t}\left(\tau_t\tau_1\int_0^t \tau_s^{-2}g_s^2\,\mathrm{d}s\right)\\ &= \frac{1}{\kappa(1,1)}\left(c_t\kappa(t,1)+\tau_t^{-1}\tau_1 g_t^2\right)\\ &= c_t r_t + \frac{\tau_1 g_t^2}{\tau_t\kappa(1,1)}.\end{aligned} \tag{A.4.30}$$

Combining (A.4.29) and (A.4.30), using the fact that $r_t\kappa(t,1)=\kappa(t,t)\rho_t$ and $\frac{\tau_1\kappa(t,1)}{\tau_t\kappa(1,1)}=\rho_t$, we may further write (A.4.27) as

$$\begin{aligned}K_{t,t+h} &= \Sigma_t + h\Big\{P_t - Q_t^\top + \Big[c_t\kappa(t,t)(1-\rho_t) - g_t^2\rho_t\Big]I\Big\} + o(h)\\ &= \Sigma_t + hS_t + o(h).\end{aligned} \tag{A.4.31}$$

We are now ready to derive (7.30). By Lemma A.4.2, the random variable X_{t+h} conditioned on $X_t = x$ follows $\mathcal{N}(\check{\mu}_{t+h}, \check{\Sigma}_{t+h})$ where, by (A.4.31),

$$\begin{aligned}\check{\mu}_{t+h} &= \mu_{t+h} + K_{t,t+h}^\top\Sigma_t^{-1}(x-\mu_t)\\ &= \mu_t + h\dot{\mu}_t + \left(I + hS_t^\top\Sigma_t^{-1}\right)(x-\mu_t) + o(h)\\ &= x + h\left(S_t^\top\Sigma_t^{-1}(x-\mu_t) + \dot{\mu}_t\right) + o(h),\end{aligned} \tag{A.4.32}$$

and, by (A.4.26) and (A.4.27),

$$
\begin{aligned}
\check{\Sigma}_{t+h} &= \Sigma_{t+h} - K_{t,t+h}^{\top}\Sigma_t^{-1}K_{t,t+h} \\
&= \Sigma_t + h\dot{\Sigma}_t - \Big(\Sigma_t + hS_t^{\top} + hS_t\Big) + o(h) \\
&= h\Bigg[P_t + P_t^{\top} - Q_t - Q_t^{\top} + \left(\frac{\partial}{\partial t}\kappa(t,t) - 2\dot{r}_t\kappa(t,1)\right)I \\
&\qquad - \left(P_t - Q_t^{\top} + \left[\left(\frac{\partial}{\partial t\prime}\kappa\right)(t,t) - \dot{r}_t\kappa(t,1)\right]I\right)^{\top} \\
&\qquad - \left(P_t - Q_t^{\top} + \left[\left(\frac{\partial}{\partial t\prime}\kappa\right)(t,t) - \dot{r}_t\kappa(t,1)\right]I\right)\Bigg] + o(h) \qquad \text{(A.4.33)} \\
&= h\left(\frac{\partial}{\partial t}\kappa(t,t) - 2\left(\frac{\partial}{\partial t\prime}\kappa\right)(t,t)\right)I + o(h). \qquad \text{(A.4.34)}
\end{aligned}
$$

However, by (7.26), we have

$$
\begin{aligned}
\frac{\partial}{\partial t}\kappa(t,t) &= \frac{\partial}{\partial t}\left(\tau_t^2\int_0^t \tau_s^{-2}g_s^2\,\mathrm{d}s\right) \\
&= 2\dot{\tau}_t\tau_t\int_0^t \tau_s^{-2}g_s^2\,\mathrm{d}s + g_t^2, \\
\left(\frac{\partial}{\partial t\prime}\kappa\right)(t,t) &= \frac{\partial}{\partial t\prime}\left(\tau_t\tau_{t\prime}\int_0^t \tau_s^{-2}g_s^2\,\mathrm{d}s\right)\Bigg|_{t\prime=t} \qquad \text{(A.4.35)} \\
&= \dot{\tau}_t\tau_t\int_0^t \tau_s^{-2}g_s^2\,\mathrm{d}s, \qquad \text{(A.4.36)}
\end{aligned}
$$

from which (A.4.34) simplifies to

$$
\check{\Sigma}_{t+h} = hg_t^2 I + o(h). \qquad \text{(A.4.37)}
$$

We can now compute $\mathbb{E}\left[u\left(t+h, X_{t+h}\right) \mid X_t = x\right]$ as follows:

$$\mathbb{E}\left[u\left(t+h, X_{t+h}\right) \mid X_t = x\right] = (2\pi)^{\frac{d}{2}} \left(\det \check{\Sigma}_{t+h}\right)^{-\frac{1}{2}} \int_{\mathbb{R}^d} u\left(t+h, x'\right) \exp\Big(- \frac{1}{2}\left(x' - \check{\mu}_{t+h}\right)^\top \check{\Sigma}_{t+h}^{-1}\left(x' - \check{\mu}_{t+h}\right)\Big)\, \mathrm{d}x' \tag{A.4.38}$$

$$= (2\pi)^{\frac{d}{2}} \left(\det \check{\Sigma}_{t+h}\right)^{-\frac{1}{2}} \int_{\mathbb{R}^d} u\left(t+h, x' + \check{\mu}_{t+h}\right) \exp\Big(- \frac{1}{2} x'^\top \check{\Sigma}_{t+h}^{-1} x'\Big)\, \mathrm{d}x'. \tag{A.4.39}$$

Invoking Lemma A.4.1, we see that (A.4.39) can be evaluated as

$$\mathbb{E}\left[u\left(t+h, X_{t+h}\right) \mid X_t = x\right] = \exp\left(\frac{1}{2}\partial_{x'}^\top \check{\Sigma}_{t+h} \partial_{x'}\right) u\left(t+h, x' + \check{\mu}_{t+h}\right)\Big|_{x'=0}. \tag{A.4.40}$$

Since $\check{\Sigma}_{t+h} = h g_t^2 I + o(h)$ by (A.4.37), expanding the power series $\exp\left(\frac{1}{2}\partial_{x'}^\top \check{\Sigma}_{t+h} \partial_{x'}\right)$ and ignoring every $o(h)$ terms, (A.4.40) becomes

$$\mathbb{E}\left[u\left(t+h, X_{t+h}\right) \mid X_t = x\right] = \left(u\left(t+h, x' + \check{\mu}_{t+h}\right) + \frac{h g_t^2}{2} \Delta u\left(t+h, x' + \check{\mu}_{t+h}\right)\right)\Big|_{x'=0} + o(h)$$

$$= u(t+h, \check{\mu}_{t+h}) + \frac{h g_t^2}{2} \Delta u(t+h, \check{\mu}_{t+h}) + o(h). \tag{A.4.41}$$

Recalling from (A.4.32) that $\check{\mu}_{t+h} = x + h\left(S_t^\top \Sigma_t^{-1}(x - \mu_t) + \dot{\mu}_t\right) + o(h)$, the Taylor expansion in the x variable for $u(t,x)$ shows that

$$\mathbb{E}\left[u\left(t+h, X_{t+h}\right) \mid X_t = x\right] = u(t+h, x) + h\Big(\frac{g_t^2}{2} \Delta u(t+h, x) + \left\langle \nabla u(t+h, x), S_t^\top \Sigma_t^{-1}(x - \mu_t) + \dot{\mu}_t \right\rangle\Big) + o(h) \tag{A.4.42}$$

whence

$$\lim_{h\to 0} \frac{\mathbb{E}\left[u\left(t+h, X_{t+h}\right) \mid X_t = x\right] - u(t,x)}{h} = \frac{\partial}{\partial t} u(t,x) + \frac{g_t^2}{2} \Delta u(t,x) + \left\langle \nabla u(t,x), S_t^\top \Sigma_t^{-1}(x-\mu_t) + \dot{\mu}_t \right\rangle. \tag{A.4.43}$$

This is exactly (A.4.25) with $f_{\mathcal{N}}(t,x) \leftarrow S_t^\top \Sigma_t^{-1}(x-\mu_t) + \dot{\mu}_t$, which concludes the proof for (7.30) and (7.31).
Finally, by (Léonard, 2013, (4.2)), the optimal drift $f_{\mathcal{N}}(t,x)$ is a *gradient field*:

$$f_{\mathcal{N}}(t,x) = \nabla \psi(t,x) \tag{A.4.44}$$

for some function $\psi : \mathbb{R}^+ \times \mathbb{R}^d \to \mathbb{R}$, implying that $S_t^\top \Sigma_t^{-1}$ must be symmetric. □