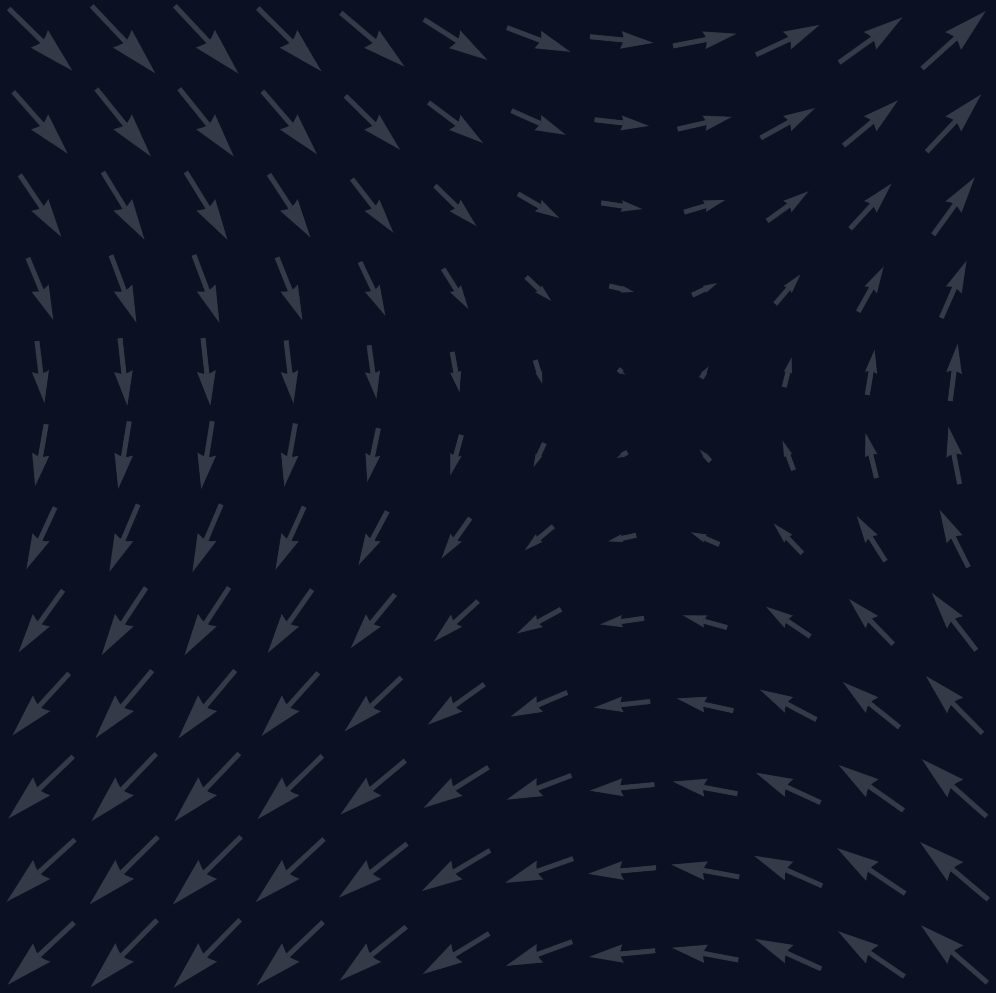
Neural Optimal Transport for Dynamical Systems

Methods and Applications in Biomedicine

Charlotte Bunne



Diss.-No. ETH 29594

CHARLOTTE BUNNE

NEURAL OPTIMAL TRANSPORT FOR DYNAMICAL SYSTEMS

METHODS AND APPLICATIONS IN BIOMEDICINE

DISS. ETH NO. 29594

NEURAL OPTIMAL TRANSPORT FOR DYNAMICAL SYSTEMS

METHODS AND APPLICATIONS IN BIOMEDICINE

A thesis submitted to attain the degree of

DOCTOR OF SCIENCES
(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

Modeling dynamical systems is a core subject of many scientific disciplines as it allows us to predict future states, understand complex interactions over time, and enable informed decision-making. Biological systems in particular are governed by dynamical processes, with their inherently complex and constantly changing patterns of interactions and behaviors. Single-cell biology has revolutionized biomedical research, as it allows to monitor such systems at unprecedented scales. At the same time, it presents us with formidable challenges: While single-cell high-throughput methods routinely produce millions of data points, they are destructive assays, such that the same cell cannot be observed twice nor profiled over time. Since many of the most pressing questions in the field involve modeling the dynamic responses of heterogeneous cell populations to various stimuli, such as therapeutic drugs or developmental signals, there is a pressing need to provide computational methods that can circumvent that limitation and re-align these unpaired measurements.

Optimal transport (OT) has emerged as a major opportunity to fill in that gap *in silico* as it allows us to reconstruct how a distribution evolves, given only access to *distinct snapshots* of *unaligned* data points. Classical OT methods, however, do not generalize to *unseen* samples. Yet, this is crucial when, for example, predicting treatment responses of incoming patient samples or extrapolating cellular dynamics beyond the measured horizon.

By harnessing the theoretical constructs of OT, this thesis explores and develops ***neural* static** and **dynamic optimal transport** schemes for elucidating the intricate dynamics of biological populations. It encapsulates an array of algorithmic frameworks, with contributions to both the *understanding* and *prediction* of population dynamics:

- First, we derive **static neural optimal transport** schemes capable of learning a map between the unpaired distributions of unperturbed and perturbed cells. These models excel at predicting single-cell responses to varying perturbations, such as cancer drug screens, and generalize the inference of treatment outcomes to *unobserved* cell types and patients. This has significant implications for personalized medicine, as it allows for the prediction of treatment responses for new patients in large-scale clinical studies.

- Second, we explore **dynamic neural optimal transport** formulations that leverage the connections of OT to partial differential equations and gradient flows through the Jordan-Kinderlehrer-Otto (JKO) scheme, as well as stochastic differential equations and optimal control through diffusion Schrödinger bridges. These methods then serve as robust tools for reconstructing stochastic and continuous-time dynamics from marginal observations, allowing us to dissect fine-grained and time-resolved cellular mechanisms.

This thesis connects a variety of seemingly unrelated concepts into a unified framework, and the presented methodologies offer a computational and mathematical foundation for modeling of cellular dynamics. This provides new avenues to understand cellular heterogeneity, plasticity, and response landscapes. Such neural parameterizations of static and dynamic OT that allow for out-of-sample inference lays the groundwork for exciting opportunities to make novel biological discoveries, infer personalized therapies from single-cell patient samples, and push the boundaries of regenerative medicine.

ZUSAMMENFASSUNG

Die Modellierung dynamischer Systeme bildet einen Schwerpunkt in vielen wissenschaftlichen Fachbereichen. Sie ermöglicht es uns, zukünftige Zustände vorherzusagen, komplexe zeitliche Interaktionen zu analysieren und fundierte Entscheidungen zu treffen. Insbesondere in biologischen Systemen, die von inhärent komplexen und ständig wechselnden Interaktions- und Verhaltensmustern gesteuert werden, kommt dieser Aspekt zum Tragen. Die Einzelzellbiologie hat dabei die biomedizinische Forschung revolutioniert, indem sie es ermöglicht, solche Systeme in beispielloser Größenordnung zu messen. Gleichzeitig stellt sie uns vor gewaltige Herausforderungen: Obwohl Einzelzell-Hochdurchsatzmethoden routinemäßig Millionen von Datenpunkten produzieren, sind es destruktive Assays, sodass dieselbe Zelle nicht wiederholt oder kontinuierlich über die Zeit gemessen werden kann. Angesichts dringender Fragen zur Modellierung der dynamischen Reaktionen heterogener Zellpopulationen auf verschiedene Perturbationen, wie therapeutische Medikamente oder Entwicklungssignale, besteht ein akuter Bedarf an der Entwicklung von Algorithmen, die diese Einschränkung überwinden und zeitliche Trajektorien einzelner Zellen rekonstruieren können.

Die mathematische Theorie des sogenannten optimalen Transports (OT) hat sich dabei als Schlüsselmethodik etabliert, um diese Lücke *in silico* zu schließen. OT erlaubt es uns zu rekonstruieren, wie sich eine Verteilung über die Zeit hinweg entwickelt hat, selbst aus *diskreten Messungen* von *nicht gekoppelten Datenpunkten*. Allerdings generalisieren klassische OT Methoden nicht auf unbekannte Proben. Dies ist jedoch eine entscheidende Fähigkeit für die Vorhersage von Behandlungsreaktionen verschiedener Patienten oder wenn zelluläre Dynamiken über den gemessenen Horizont hinaus extrapoliert werden sollen.

Diese Dissertation entwickelt, basierend auf den theoretischen Konzepten des optimalen Transports, sowohl ***neuronale* statische** als auch **dynamische optimale Transport** Systeme, um die komplexen Dynamiken biologischer Populationen zu modellieren und zu verstehen:

- **Statischer neuronaler optimaler Transport**: Diese Methoden sind in der Lage, eine Abbildung zwischen ungepaarten Verteilungen von unperturbierten und perturbierten Zellen zu erlernen. Sie erziehlen sehr gute

Ergebnisse in der Vorhersage von Einzelzellreaktionen auf verschiedene Perturbationen, wie zum Beispiel Krebsmedikamenten, und ermöglichen die Vorhersage von Behandlungsreaktionen für neue Patienten in groß angelegten klinischen Studien.

- **Dynamischer neuronaler optimaler Transport**: Durch das Verknüpfen von OT zu partiellen Differentialgleichungen und Gradientenflüssen durch das Jordan-Kinderlehrer-Otto (JKO) Schema, sowie stochastischer Differentialgleichungen und optimalen Steuerungen durch diffusive Schrödinger Brücken dienen diese Algorithmen als robuste Werkzeuge zur Rekonstruktion stochastischer und kontinuierlicher dynamischer Prozesse. Sie ermöglichen dadurch feinkörnige Analysen zeitlich aufgelöster zellulärer Mechanismen.

Diese Dissertation verbindet eine Vielzahl scheinbar nicht verwandter Konzepte in einem einheitlichen Rahmen, bietet eine methodische und mathematische Grundlage für die Modellierung zellulärer Dynamiken und erlaubt damit, zelluläre Heterogenität, Plastizität und Reaktionen zu Perturbationen besser zu verstehen. Die entwickelten neuronalen Parameterisierungen von statischen und dynamischen OT, die Inferenzen außerhalb der Stichprobe zulassen, eröffnen einen aufregenden Forschungsweg für die Zukunft. Sie bieten Möglichkeiten für neuartige biologische Entdeckungen und die Vorhersage personalisierter Therapien aus Einzelzell-Patientenproben und erweitern die Grenzen der regenerativen Medizin.

ACKNOWLEDGEMENTS

First and foremost, I would like to express my sincere gratitude to my advisor Andreas Krause for his excellent mentorship, continued support, for innovative ideas, precise remarks, and critical thinking. Thank you in particular for your trust and for finding the right balance between guidance and giving me the freedom to pursue my research interests. Your remarkable insights into which ideas are likely to succeed as well as how to communicate a scientific result never cease to amaze me and will remain with me as a particularly valuable lesson for the future.

I would like to extend my gratitude to Marco Cuturi for being my co-advisor from the very beginning. Without your invaluable mentorship, this thesis would certainly not be the same. Thank you for empowering me to acquire novel skills and develop and realize my own research agenda. Also, thank you for providing me with countless opportunities ranging from internships at Google Research and Apple to co-presenting the tutorial at ICML 2023 with me.
This combination of advisors has been key for my doctorate studies.

Further, I would also like to thank Jure Leskovec for participating in my thesis committee and for providing me with insightful feedback. Throughout my academic path, various scientists have left a mark on me as a researcher, chief among them Aviv Regev. While our academic collaboration is still at the very beginning, every interaction with you is intellectually rewarding. I am excited about the upcoming chapter with both of you at Stanford and Genentech.

The thesis would not be the same without the longstanding collaboration with Lucas Pelkmans and Gabriele Gut. First, thank you, Lucas, for serving on my thesis committee. I have particularly enjoyed our academic ventures at the intersection of molecular biology and machine learning. Working with you has taught me to think differently about problems, ask different questions, and had a lasting effect on this thesis.

I am very grateful to Anne Carpenter and Shantanu Singh for hosting me at the Broad Institute of MIT and Harvard and for the many research discussions we had. I will forever cherish my time in your lab.

I am profoundly indebted to all co-authors and fantastic collaborators with whom I worked on many projects throughout the past years, in particu-

lar, Gunnar Rätsch, Stefan Stark, Ya-Ping Hsieh, Matteo Pariset, Valentin De Bortoli, Octavian Ganea, Vignesh Ram Somnath, Frederike Lübeck, Laetitia Meng-Papaxanthos, Mojmír Mutný, Regina Barzilay, Tommi Jaakkola, Connor Coley, 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 to my first academic mentors, Stefanie Jegelka, David Alvarez Melis, Roland Eils, Thomas Höfer, and Lisa Buchauer for guiding me through my first actual research projects. Thank you, Stefanie, for being a mentor much beyond my time at MIT and for your continued advice. Thank you, David, for introducing me to optimal transport many years back, for every coffee during the frequent visits to Cambridge, and in particular, for navigating me through my faculty search and life in academia. I learned so much from you and will be forever grateful.

Lastly, I would not have started my scientific journey without the Life-Science Lab at the German Cancer Research Center. It is what sparked my interest in biology and engineering and set the course for what is here today. Participating in the "world championship in synthetic biology" was an eye-opener for the power and potential of interdisciplinary research.

Thanks to all members of the Learning and Adaptive Systems group at ETH Zurich and the ETH AI Center for creating an excellent research environment. Thanks to Lars Lorch, Mohammad Reza Karimi, and Lenart Treven for lengthy discussions on causal inference, sampling, or control theory. Thank you, Rita Klute, for the support, for entangling the jungle of bureaucracy, and for overcoming *any* administrative hurdle.

I am very grateful to my parents Nele and Egon and my siblings Kaspar, Henriette, and Frieder for their endless love, advice, encouragement, and support. Growing up in this environment of Freigeister has made me the person I am today.

Thank you to my friends around the globe for making life joyful and adventurous, every party fun, and the difficult moments bearable.

Lastly, thank you to Pol for expanding my world into unknown dimensions: "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

1 INTRODUCTION 1
1.1 Scope and Contributions . . . 3
1.2 Publications . . . 7
1.3 Collaborators . . . 9
2 DYNAMIC PROCESSES IN BIOMEDICINE 11
2.1 Single-Cell High-Throughput Technologies . . . 12
2.1.1 Sequencing-Based Screening . . . 12
2.1.2 Optical Phenotypic Screening . . . 13
2.2 Cellular Perturbation Responses to Drugs and Treatments . . . 14
2.3 Lineage Tracing in Developmental Biology . . . 15
2.4 Optimal Transport for Single-Cell Biology . . . 17
3 OPTIMAL TRANSPORT FOR DYNAMICAL SYSTEMS 19
3.1 Static Optimal Transport . . . 19
3.1.1 Monge Problem . . . 20
3.1.2 Kantorovich Relaxation . . . 21
3.1.3 Kantorovich Duality . . . 22
3.1.4 Geometry of Optimal Transport . . . 24
3.2 Dynamic Optimal Transport . . . 27
3.2.1 Monge-Ampère Equation . . . 28
3.2.2 Benamou-Brenier Formulation . . . 28
3.2.3 Jordan-Kinderlehrer-Otto Flows . . . 30
3.2.4 Stochastic Control Perspective . . . 33
3.2.5 Schrödinger Bridges . . . 34

I STATIC NEURAL OPTIMAL TRANSPORT

4 NEURAL OPTIMAL TRANSPORT 41
4.1 Neural Optimal Transport Solvers . . . 43
4.1.1 Input Convex Neural Networks . . . 45
4.1.2 Alternative Approaches . . . 45
4.2 Related Work . . . 46
4.3 CellOT: Predicting Perturbation Responses via Neural Monge Maps . . . 47
4.4 Empirical Evaluation . . . 49
4.4.1 Predicting Treatment Outcomes of Cancer Drugs . . . 49
4.4.2 Capturing Cell-to-Cell Variability in Drug Responses . . . 52

4.4.3 Disentangles Subpopulation-Specific Drug Effects . . . 54
4.4.4 Inferring Cellular Responses in Unseen Patients 57
4.4.5 Reconstructing Innate Immune Responses across Different Species . . . 58
4.4.6 Generalizing Developmental Fate Decisions from Multipotent to Oligopotent Cell Populations . . . 59
4.5 Discussion . . . 60
5 NEURAL OPTIMAL TRANSPORT WITH CONTEXT 63
5.1 CondOT: Supervised Training of Conditional Monge Maps . 64
5.1.1 A Regression Formulation for Conditional OT Estimation . . . 65
5.1.2 Integrating Context in Convex Architectures . . . 66
5.1.3 Conditional Monge Map Architecture . . . 67
5.2 Empirical Evaluation . . . 71
5.2.1 Modeling Dosage-Sensitive Treatment Responses to Cancer Drugs . . . 71
5.2.2 Predicting Cell Type-Specific Treatment Responses to Cancer Drugs . . . 73
5.2.3 Inferring Genetic Perturbation Responses . . . 73
5.3 Discussion . . . 76
II DYNAMIC NEURAL OPTIMAL TRANSPORT
6 LEARNING DYNAMICAL SYSTEMS VIA OT AND GRADIENT FLOWS 79
6.1 Population Dynamics as Gradient Flows . . . 80
6.2 JKOnet: A Proximal Optimal Transport Model . . . 83
6.2.1 Reformulation of JKO Flows via ICNNs . . . 83
6.2.2 Learning the Free Energy Functional . . . 85
6.2.3 Bilevel Formulation of JKOnet . . . 86
6.3 Empirical Evaluation . . . 88
6.3.1 Synthetic Dynamics . . . 88
6.3.2 Single-Cell Dynamics . . . 91
6.4 Discussion . . . 94
7 LEARNING DYNAMICAL SYSTEMS VIA OT AND STOCHASTIC CONTROL 95
7.1 Diffusion Schrödinger Bridges . . . 96
7.2 Data-Driven Priors for Diffusion Schrödinger Bridges . . . 98
7.2.1 Preliminaries on Gaussian Optimal Transport . . . 100
7.2.2 The Gaussian Schrödinger Bridge Problem . . . 101

7.2.3 The Bures-Wasserstein Geometry of $\sigma\mathbb{W}_t$-Gaussian Schrödinger Bridges . . . 104
7.2.4 Closed-Form Solutions of General Gaussian Schrödinger Bridges . . . 108
7.2.5 GSBFLOW: Recovering Stochastic Dynamics via Gaussian Schrödinger Bridges . . . 112
7.2.6 Empirical Evaluation . . . 113
7.2.7 Discussion . . . 118
7.3 Diffusion Schrödinger Bridges from Sparse Trajectories . . . 120
7.3.1 SBALIGN: Aligned Diffusion Schrödinger Bridges . . . 122
7.3.2 Aligned Schrödinger Bridges as Prior Processes . . . 127
7.3.3 Empirical Evaluation . . . 128
7.3.4 Discussion . . . 132
8 CONCLUSION AND FUTURE DIRECTIONS 133
A APPENDIX 167
A.1 Further Empirical Evaluation . . . 168
A.2 Datasets . . . 172
A.2.1 Srivatsan et al. (2020) . . . 173
A.2.2 Bunne et al. (2023b) . . . 174
A.2.3 Norman et al. (2019) . . . 175
A.2.4 Moon et al. (2019) . . . 176
A.2.5 Weinreb et al. (2020) . . . 178
A.2.6 Other Datasets . . . 179
A.3 Evaluation Metrics . . . 179
A.4 Proof of Theorem 3 . . . 181
A.5 The Bures-Wasserstein Geometry of Gaussian SBs . . . 182
A.5.1 Review of Bures-Wasserstein Geometry . . . 182
A.5.2 Proof of Theorem 4 . . . 182
A.5.3 Some Interesting Consequences of Theorem 4 . . . 190
A.6 Proof of the Closed-Form Solutions for Gaussian SBs . . . 193
A.6.1 Preliminaries for the Proof of Theorem 5 . . . 193
A.6.2 The Proof . . . 193

Σ_n	probability simplex of size n.
(μ, ν)	measures defined on spaces $(\mathcal{X}, \mathcal{Y})$.
(u, v)	histograms in the simplices $\Sigma_n \times \Sigma_m$.
$p_\mu = \frac{d\mu}{dx}$	density with respect to the Lebesgues measure.
$(\mu = \sum_i u_i \delta_{x_i}$, $\nu = \sum_j v_j \delta_{y_j})$	discrete measures defined on spaces $(x_1, \dots, x_n \in \mathcal{X}, y_1, \dots, y_m \in \mathcal{Y})$.
$\mathbf{1}$	matrix of $\mathbb{R}^{n \times m}$ with all entries identically set to 1.
$\mathbb{1}$	indicator function.
Id	identity map.
$c(x, y)$	ground cost, with associated pairwise cost matrix $(c(x_i, y_j))_{ij}$ evaluated on the support of μ, ν.
$\|\cdot\|_2^2$	squared Euclidean norm.
$\sharp$	pushforward operator.
$T : \mathcal{X} \times \mathcal{Y}$	Monge map, typically such that $T_\sharp \mu = \nu$.
π	coupling measure between μ and ν, for discrete measures $\pi = \sum_{ij} \mathbf{P}_{ij} \delta_{(x_i, y_j)}$.
$\Pi(\mu, \nu)$	set of couplings, for discrete measures $U(u, v)$.
$\text{supp}(\pi)$	support of π.
$W(\mu, \nu)$	Wasserstein distance between measures μ and ν.
$H(\pi)$	entropy of coupling π.
ε	regularization strength of the entropy regularization.
(f, g)	dual potentials.
f^*	Legendre transform of function f.
$f^\star$	optimum of function f.
φ	convex potential.
$(\mu_t)_{t=0}^T$	dynamic measures with $\mu_{t=0} = \mu_0$ and $\mu_{t=T} = \mu_T$.
v	speed in the dynamic Benamou-Brenier or control in the stochastic optimal control formulation.
Δ	Laplace operator.

τ	step size.
$\langle \cdot, \cdot \rangle$	Euclidean dot-product between vectors.
D_{H}	Hellinger distance.
D_{KL}	Kullback-Leibler divergence.
ℓ_1	Manhattan distance.
ℓ_2	Euclidean distance.
$\circ$	Hadamard element-wise product.
σ	noise level.
$\mathbb{P}_t$	stochastic process with $t \in [0, 1]$.
$\mathbb{W}_t$	standard Wiener process.
$\mathbb{Q}_t$	reference process, e.g., Brownian motion.
$Z_t, \hat{Z}_t$	time-indexed smooth vector fields indicating the forward and backward policy.
$\mathcal{N}(\mu, \Sigma)$	multivariate Gaussian distribution with mean vector μ and covariance matrix Σ.
$\dot{x}$	first derivative of x taken with respect to time.
$\ddot{x}$	second derivative of x taken with respect to time.
θ and ϕ	parameters of neural networks.
γ	learning rate.
ℓ	loss function.
ω	kernel function.

ACRONYMS

k-**NN** *k*-nearest neighbor.
4i iterative indirect immunofluorescence imaging.

BDT Black-Derman-Toy.
BM Brownian motion.

cDNA copy DNA.
CNN convolutional neural network.
CRISPR clustered regularly interspaced short palindromic repeats.

DDPM denoising diffusion probabilistic model.
DNA deoxyribonucleic acid.
DSB diffusion Schrödinger bridge.

EB embryoid bodies.
ESC embryonic stem cell.

FACS fluorescence activated cell sorting.
FBSDE forward-backward SDE.

GAN generative adversarial network.
GEO gene expression omnibus.
GNN graph neural network.
GSB Gaussian Schrödinger bridge.

HSPC hematopoietic stem and progenitor cells.
HVG highly variable genes.

i.i.d. independent and identically distributed.
ICNN input convex neural network.
IPF iterative proportional fitting.
iPSC induced pluripotent stem cell.

JKO Jordan-Kinderlehrer-Otto.

KKT Karush-Kuhn-Tucker.

l.s.c. lower semi-continuous.
LPS lipopolysaccharides.

MDS multidimensional scaling.
MEF mouse embryonic fibroblast.
MLP multi-layer perceptron.
MMD maximum-mean-discrepancy.
MOA mode of action.
mRNA messenger RNA.

NF normalizing flows.
NN neural network.

o.o.d. out-of-distribution.
o.o.s. out-of-sample.
ODE ordinary differential equation.
OHE one-hot encoding.
OT optimal transport.
OU Ornstein-Uhlenbeck.

PBMC peripheral blood mononuclear cells.
PCA principal component analysis.
PDE partial differential equation.
PICNN partially input convex neural network.
PS perturbation signatures.

RMSD root-mean-square deviation.
RNA ribonucleic acid.
RNA-seq RNA-sequencing.
RNN recurrent neural network.

SB Schrödinger bridge.
sc single-cell.
SDE stochastic differential equation.
SGM score-based generative model.
SMLD score matching with Langevin dynamics.

UMAP uniform manifold approximation and projection.

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)

Dynamical systems are governing every aspect of life, encapsulating the unifying principles and complex interactions that shape our world. They are the cornerstone of diverse scientific fields such as physics, chemistry, and biology. In physics, they assist in understanding the movement and interaction of objects large and small, from planetary orbits to particle motion. In the world of chemistry, they guide the interaction and transformation of molecules, offering insights into reaction dynamics and molecular behavior. In biology, it opens doors to insights ranging from the dynamic signaling activities in single cells to the ebb and flow of entire ecosystems.

Given the complexity and constant evolution inherent to living organisms, the application of dynamical systems in biology offers a particularly rich and challenging field of study. 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, advances in single-cell genomics and transcriptomics now provide a direct window into the molecular makeup of individual cells, with a resolution that was deemed unthinkable two decades ago, 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 that underlie health and disease.

The ability of single-cell high-throughput methods to produce routinely millions of data points holds multiple promises. They do, however, come with several limitations: Most prominently, they are typically destructive assays as cells are usually fixed and stained or chemically destroyed to obtain measurements. Thus, they produce data points that are not *aligned*, meaning that the same cell cannot be observed twice nor can we record

continuous time trajectories of it. Since many of the most pressing questions in the field involve modeling and understanding the dynamic responses of heterogeneous cell populations to various stimuli, such as environmental signals, developmental processes, genetic perturbations, or drug treatments, there is a pressing need to provide algorithms that can circumvent that limitation and (●) **reconstruct such dynamical processes**. This constraint 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.

Significantly, the diversity within a cell population, or *cellular heterogeneity*, plays a crucial role in determining how sensitive or resistant cells are to perturbations. Rather than resorting to population averages, we need to (●) **model the problem at the particle level**, e.g., based on distributions of single cells, in order to capture and then further analyze the distinct cells' responses to a perturbation. This requires scalable and principled algorithms that are well aligned with the constrained experimental settings of high-throughput methods and incorporate the inherent structure of biological processes.

Beyond, to effectively derive complex and nonlinear dynamic processes *from data*, it is essential to (●) **develop neural network-based parameterizations** of such processes. Traditional mathematical models might not be sufficient to capture all intricacies of a system. This holds particularly true in biology, where our understanding of the underlying mechanisms is incomplete and many factors contribute to the overall dynamics. Most importantly, neural networks (NNs) allow us to infer and forecast dynamical processes. This enables *out-of-sample* predictions of the evolution of previously unseen particles, such as unobserved cells from incoming patient samples and to extrapolate the inference to new cell types or beyond the recorded time horizon.

However, despite recent successes of deep learning in many areas such as computer vision (LeCun et al., 1998; Krizhevsky et al., 2012), natural language processing (Bengio et al., 2000; Vaswani et al., 2017), game playing (Mnih et al., 2015; Silver et al., 2016), and biochemistry (Jumper et al., 2021; Kipf and Welling, 2017; Jin et al., 2022), deep neural networks still fall short of a general ability to model dynamic and complex evolutions of

populations of particles. At the heart of this challenge lies the problem of inductive bias (Mitchell, 1980): How can models be constructed to learn the essential representations, abstractions, and skills that will enable them to generalize to unseen and unforeseeable situations?

The examples above highlight the effectiveness of crafting specific architectural inductive biases that are tailored to the object of study: convolutional neural networks (CNNs) in computer vision contain filters that process local features in an image, recurrent neural networks (RNNs) with attention layers in natural language processing facilitate the comprehension of long-range connections in sentences, or graph neural networks (GNNs) account for neighborhood structures to simulate chemical interactions within proteins and molecules (Ganea et al., 2021; Somnath et al., 2021). Introducing and developing (●) **inductive biases in deep learning architectures for population dynamics** is quintessential toward the success and ability of such methods to capture complex dynamics from *unaligned* snapshot data.

1.1 SCOPE AND CONTRIBUTIONS

In this thesis, we introduce and design novel deep learning architectures capable of modeling heterogeneous population dynamics from unaligned snapshot data. A framework that facilitates the design of such deep learning architectures that (●) operate on distributions, i.e., allows modeling heterogeneous populations of particles, (●) trained based on samples of *unaligned* distributions, is the **theory of optimal transport**.

The mathematical foundation of this work further builds on the intuition that perturbations incrementally alter the molecular profiles of cells. This underlying assumption aligns with the theory of optimal transport (OT) that studies the evolution of measures under the *minimum effort* principle. Thus, (●) OT serves naturally as an inductive bias for the deep learning architectures introduced in this thesis. By providing tools for modeling the evolution of distributions over time, OT allows us to reconstruct and predict the incremental changes in cell states upon perturbation such as therapeutic agents or developmental signals. In recent years, OT has enabled significant advances in single-cell biology problems (Schiebinger et al., 2019; Lavenant et al., 2021; Demetci et al., 2022). However, traditional OT methods do not enable predictions for unperturbed cells that have not been previously observed. They are thus unable to predict perturbation responses out-of-sample, e.g., of cells from new incoming samples, such as those from unseen

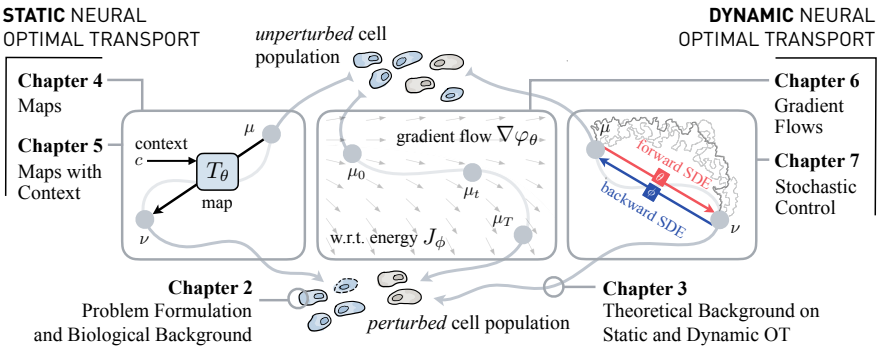


Figure 1.1: Structure of the thesis with references to the chapters.

patients. This thesis is thus concerned with the development of so-called (●) **neural optimal transport** methods that allow for out-of-sample *predictions* on unseen cells and *forecasting* of cellular dynamics (Makkuva et al., 2020; Tong et al., 2020; Korotin et al., 2021a; Bunne et al., 2022b,a, 2023b).

Optimal transport comes in many different flavors and has strong mathematical roots in dynamical systems theory:

The goal of the *static* formulation of OT is to identify a map that *optimally* morphs or transports a distribution onto another, and thus "best" explains the one-step evolution of a measure. The static framework, however, has an immediate dynamic interpretation. By introducing a time-variable into the transport, OT allows us to study the continuous evolution of a measure over time. Importantly, generalizations of this *dynamic* formulation can be formalized through ordinary, partial, or stochastic differential equations. By building on innovations in neural ordinary differential equations (ODEs) (Chen et al., 2018) and flow matching models (Lipman et al., 2023; Pooladian et al., 2023a; Liu et al., 2022b), physics-informed neural partial differential equations (PDEs) (Brandstetter et al., 2022; Raissi et al., 2019), as well as stochastic differential equation (SDE)-based score-based generative models (SGMs), neural OT methods can be equipped with various powerful backbone architectures.

The thesis is structured as follows (see outline in Fig. 1.1). After reviewing dynamical systems in biomedicine in Chapter 2 and providing a mathematical foundation of static and dynamic optimal transport in Chapter 3, the **thesis contributions** are organized in two parts:

PART I: STATIC NEURAL OPTIMAL TRANSPORT. At its core, static OT defines a so-called Monge *map*, which provides an actionable way to flow from one probability distribution onto another. The design of neural network-parameterizations of such maps and their application to predict perturbation responses of single-cells will form the first contribution of this thesis. This comprises approaches that are a direct consequence of the famous Brenier theorem, and allow us to model Monge maps as gradients of input convex functions (Bunne et al., 2023b, 2022a).

More concretely, Part I comprises two chapters: Chapter 4 proposes CELLOT, a framework leveraging optimal transport theory and convex neural architectures that aligns unpaired distributions of unperturbed and perturbed cells to predict individual cellular responses to various perturbations. CELLOT outperforms current methods in drug response predictions, and provides ways for interpretation and understanding of heterogeneous single-cell responses and patient-specific treatment outcomes (Bunne et al., 2023b).

Chapter 5 generalizes the above framework and introduces the CONDOT model, a multi-task approach that parameterizes a *family of OT maps* that can be conditioned on a context variable. Conditioning allows, for example, to predict the response of single-cells to combination therapies, i.e., the inference of the effect of genetic or therapeutic perturbations applied in combination (Bunne et al., 2022a).

PART II: DYNAMIC NEURAL OPTIMAL TRANSPORT. Beyond mappings, OT provides a mathematical link to geometric variational frameworks that allow studying flows of distributions on metric spaces. In particular, Brenier's dynamical formulation of OT has given rise to a flurry of applications studying PDEs as gradient flows or steepest descent in spaces of probability measures. This thesis highlights the 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.

Chapter 6 introduces a reformulation of the JKO objective that allows for gradient-based learning in an end-to-end fashion. Further, it proposes JKONET, a neural network architecture that models the collective dynamics of particle populations unsing a parameterized energy by learning

the causal JKO flow from measurement snapshots, demonstrating robust performance across a variety of settings (Bunne et al., 2022b).

Lastly, beyond PDEs, this thesis explores the link between the optimal transport problem and the Schrödinger bridge (SB) problem from stochastic control. It represents a key connection that has recently fueled the development of diffusion Schrödinger bridge (DSB) (De Bortoli et al., 2021b; Chen et al., 2021b; Liu et al., 2022a; Bunne et al., 2023a). Estimating such bridges, however, is notoriously difficult, motivating our contributions in Chapter 7 that propose novel adaptive schemes:

First, in Section 7.2 our goal is to rely on Gaussian approximations of the data to provide the reference stochastic process needed to estimate SBs. To that end, we solve the Schrödinger bridge problem with Gaussian marginals, for which we derive, as a central contribution, a closed-form solution and SDE-representation. We use these formulas to define the reference process used to estimate more complex SBs, and show that this does indeed help with its numerical solution of learning diffusion Schrödinger bridges. We obtain notable improvements when reconstructing single-cell genomics experiments of developmental embryoid body diffierentiation (Bunne et al., 2023a).

Second, parallel developments in bioengineering aim at overcoming the destructive nature of high-throughput methods and provide us with methodology that allows for partial tracking of single-cell trajectories over time. In Section 7.3, we thus introduce a novel algorithmic framework for solving diffusion Schrödinger bridges that respects data alignment and incorporates *sparse* trajectories by combining classical Schrödinger bridge theory with Doob's h-transform (Somnath et al., 2023). The overall framework comprises a much simpler training procedure and substantial improvements in tasks such as cellular differentiation processes of hematopoiesis (Weinreb et al., 2020).

1.2 PUBLICATIONS

All results presented in this thesis have been published in the following conference proceedings and journals:

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, 2022.

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

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

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.

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

FURTHER PUBLICATIONS. The following publications of the author and collaborators are more broadly relevant to the topic of this thesis but have not been directly included:

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

Matteo Manica, Charlotte Bunne, Roland Mathis, Joris Cadow, Mehmet Eren Ahsen, Gustavo A Stolovitzky, and Maria Rodriguez Martinez. COSIFER: a Python package for the consensus inference of molecular interaction networks. *Bioinformatics*, 37(14), 2020.

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)*, 2021.

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

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.

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.

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.

Matteo Pariset, Ya-Ping Hsieh, Charlotte Bunne, Andreas Krause, and Valentin De Bortoli. Unbalanced Diffusion Schrödinger Bridge. *arXiv Preprint arXiv: 2306.09099*, 2023.

1.3 COLLABORATORS

This thesis would not have been possible without my advisors, Andreas Krause and Marco Cuturi, and many of the ideas presented here have been shaped in our meetings. I further enjoyed collaborating with my colleagues on numerous ideas, and the results presented and not otherwise cited are by the author and collaborators. In particular, Chapter 4 contains material of a publication with shared first authorship between the author, Stefan Stark, and Gabriele Gut. Besides Andreas Krause, the corresponding authors are Kjong-Van Lehmann, Lucas Pelkmans, and Gunnar Rätsch. Section 7.2 is based on a joint first authorship project with Ya-Ping Hsieh who contributed the theoretical analysis of that work. Lastly, Section 7.3 contains material from a publication where Vignesh Ram Somnath and Matteo Pariset share the first authorship while the author serves as corresponding author.

2

DYNAMIC 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)

Dynamical processes, with their inherently complex and constantly changing patterns of interactions and behaviors, are fundamental to every function of life, from the oscillatory rhythms of cellular processes to the broader orchestration of biological systems. This involves an array of intricate interactions between molecules, genes, cells, tissues, and their biophysical environment. They span a multitude of scales and dimensions — spatial as well as temporal. Early events in cell signaling, for example, often start within seconds after the stimulus, followed by intracellular signaling and transcription changes over minutes to hours. In contrast, cell fate decisions like division, differentiation, or apoptosis can take many hours or days to manifest (Spiller et al., 2010). Measuring and modeling these inherently stochastic dynamics is critical to the effective understanding of biological systems and the subsequent development of diagnostic and therapeutic tools. However, they are equally daunting, demanding novel experimental and computational strategies.

After discussing current advances in high-throughput technologies for capturing dynamical systems in biomedicine (Section 2.1), we will discuss prime examples of such processes that are subject of this thesis. These are: the analysis of cellular responses to perturbations such as drugs or other therapeutic agents (Section 2.2) and secondly, the cell differentiation processes in developmental biology (Section 2.3). We will illuminate the critical role they play in biomedicine and the myriad of challenges and opportunities they present. This will be followed by a discussion on existing computational approaches that study these dynamical processes (Section 2.4).

2.1 SINGLE-CELL HIGH-THROUGHPUT TECHNOLOGIES

Until recently, our understanding of cellular dynamics was limited to "bulk" measurement technologies, which yield average measurements of a cell population. The advent of single-cell *omics* technologies has marked a significant shift from such bulk analyses, enabling a highly detailed examination of individual cells across various layers such as the gen*ome*, transcript*ome*, epigen*ome*, and prote*ome*. This shift has revealed previously hidden complexities, such as rare cell types, transitional states, and cell-to-cell variability. Transformative techniques like single-cell RNA sequencing (Section 2.1.1) have allowed researchers to profile gene expression in thousands of cells simultaneously, shedding light on new cellular subpopulations and responses (Jia et al., 2022). Additionally, modern imaging technologies have deepened our understanding of cell morphology and cellular signaling (Section 2.1.2). The integration of these diverse omics layers provides a holistic view of cellular diversity, opening new avenues in the study of health, disease, and personalized medicine (Baysoy et al., 2023).

Each point x of such single-cell measurements then represents the recorded features of a single cell. Each feature (dimension) of that point tracks the expression level of each studied gene, or morphological and signaling feature strength in that cell, at a given time. In this setting, a few thousand cells are sampled from a large population of cells, to obtain their features (high-throughput). This is done at distinct time points throughout cellular processes. Because of the destructive process of these measurements, each snapshot consists of different cells. Two consecutive snapshots can be seen as two point clouds or, alternatively, as two tabular datasets $X = [x_1, \ldots, x_n]$ and $Y = [y_1, \ldots, y_m]$: Each of the n or m rows contains a cell and its d-dimensional feature representation, where each column denotes a particular feature, e.g., the activity of a gene. In order to understand the temporal evolution of the cell population over time, we aim to provide an informed guess on an alignment or a map that relates the two cell populations X and Y.

2.1.1 *Sequencing-Based Screening*

Sequence-based profiling methods have emerged as a transformative tool for understanding the complexity of biological systems at the resolution of individual cells. Methods such as single-cell RNA-sequencing (scRNA-seq), enable us to characterize gene expression within a single cell. By mapping

the transcriptomic landscapes of individual cells, sequence-based profiling methods have allowed for unprecedented insights into developmental biology, tissue homeostasis, disease pathology, and therapeutic responses.

In the process, RNA serves a direct quantifier for gene activity: Genes in DNA are transcribed into messenger RNA (mRNA), which is then translated into proteins that carry out various functions within the cell. By measuring the amounts and types of mRNA present in a cell at a given time, i.e., the transcriptome, we can understand which genes are being actively expressed. This information is valuable as changes in gene expression can signal various biological processes, such as cell development, responses to environmental stimuli, disease states, and much more. In order to record the gene expression in single cells, scRNA-seq requires isolating individual cells, often using techniques such as fluorescence activated cell sorting (FACS) (Julius et al., 1972) or droplet-based technologies (Brouzes et al., 2009; Mazutis et al., 2013; Debs et al., 2012). Each cell's mRNA is reverse transcribed into copy DNA (cDNA), including a unique molecular identifier to correct for amplification bias. This cDNA is then amplified and prepared for high-throughput sequencing and the resultant data provide insights into gene expression of individual cells.

Crucially, this process is destructive: Cells are destroyed, i.e., lysed, in order to access the mRNA within. Once a cell is lysed, its structural integrity and function are lost, making it impossible to further manipulate or use that particular cell for subsequent experiments. This is a fundamental limitation, as the process of obtaining the high-resolution molecular data comes at the expense of the cell's viability.

2.1.2 *Optical Phenotypic Screening*

Optical phenotypic screening allows for the analysis of single cells based on their morphological and biochemical characteristics, with minimal perturbation to their natural state. Techniques such as fluorescence microscopy, time-lapse imaging, or flow cytometry typically involve labeling cells with fluorescent dyes or proteins that bind to or are expressed by specific cellular components of interest. Such targets tagged with markers can be potential important components of signaling pathways or crucial regulators and indicators of core cellular functions. The cells are then imaged or passed through a laser, and the fluorescence emitted is measured, providing information about the presence and quantity of the target molecules within each cell. The resulting images allow the extraction of morphological properties

and the fluorescence intensity of each tag in different parts of each cell (Carpenter et al., 2006).

However, it's important to note that while these methods are in general non-destructive in nature, some processes such as labeling or the use of certain dyes could potentially have some impact on cell viability or behavior. So while these techniques in general allow for dynamic tracking of cellular processes (Fischer et al., 2019; Hashimoto et al., 2016; Tvarusko et al., 1999; Busch et al., 2015), cells are usually fixed prior to staining to preserve cellular structures and allow for longer-term storage. Such procedures are lethal for cells, making longitudinal studies on the same cells impossible.

High-content imaging, particularly when augmented by multiplexing abilities, is ideally suited to study heterogeneous cell responses. In this thesis, we study heterogeneous cell line responses to various cancer drugs based on a measurement technology known as iterative indirect immunofluorescence imaging (4i) (Gut et al., 2018). With 4i, cells are fixed and fluorescently labeled antibodies are iteratively hybridized, imaged, and removed from a sample to measure the abundance and localization of proteins and their modifications. Thus, 4i quickly generates large, spatially resolved phenotypic datasets rich in molecular information from thousands of treated and untreated cells. Additionally to the multiplexed information generated by 4i, cellular and nuclear morphology are routinely extracted from microscopy images (without the need for 4i) by image analysis algorithms (Carpenter et al., 2006).

2.2 CELLULAR PERTURBATION RESPONSES TO DRUGS AND TREATMENTS

A fundamental task in personalized medicine involves predicting outcomes and responses of patients to potential treatments from the patient biopsies or tissue culture during screens. This would allow us to subsequently select the most effective therapy for each patient. A key aspect of biomedical research thus concerns the study of cellular perturbation responses to therapeutic agents, including drugs and other treatments. Perturbations such as small molecule drugs can have profound effects on the biological *phenotype* that responds in complex and often unpredictable ways. Such responses might range from the induction of apoptosis to changes in cellular proliferation, migration, differentiation, and metabolism, each significantly affecting the overall state of the organism. Perturbations can also trigger

cascade effects across interrelated signaling pathways, leading to a state of cellular disequilibrium.

Populations of cells are almost always *heterogeneous* in function and fate and different cell states exhibit distinct sensitivities toward a external stimuli (Spiller et al., 2010). Subsequently, their responses to a perturbation strongly vary. Heterogeneity is particular inherent in cancer and fuel for drug *resistance*: Cancer frequently results from damage to multiple genes controlling cell division and tumor suppressors. A defect in the regulation of any of these mechanisms often results in genomic instability, which involves single-base substitutions to whole-genome doublings and is critical to the emergence and progression of many cancers (Dagogo-Jack and Shaw, 1991). Understanding diverging behavior of tumor cells in response to a therapy is crucial in order to understand the underlying mechanisms of cellular sensitivities and resistance. At the same time it reveals potential targets for improved therapeutic strategies.

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. In order to understand how an unperturbed population μ responds and evolves into the perturbed population ν, we need to recover a map T that describes the perturbation effect, i.e., $T_\sharp \mu = \nu$. Parameterizing this through a neural network then allows us to predict how a cell x changes upon perturbation into the perturbed state $y = T(x)$ (see Fig. 2.1a).

2.3 LINEAGE TRACING IN DEVELOPMENTAL BIOLOGY

Complex cellular dynamics are not only initiated through external stimuli, but also during developmental processes, tissue regeneration, and formation. The spectacular journey of a single zygote in the embryonic development, for example, metamorphosing into a complex, multicellular organism, is largely governed by the mechanisms of cell differentiation, where pluripotent stem cells commit to specific lineages and mature into diverse cell types.

Understanding the molecular programs that guide such differentiation processes is a major challenge. However, approaches relying on the bulk

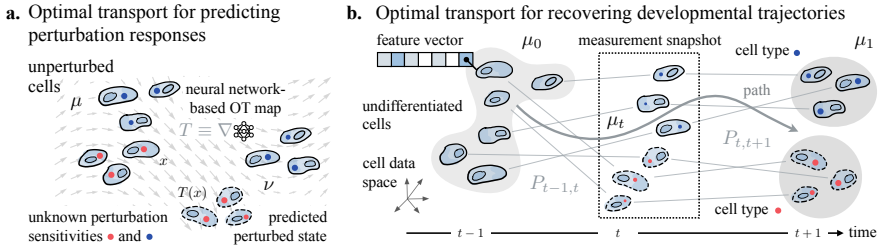


Figure 2.1: Overview on different dynamical processes in biomedicine. a. In order to understand and predict the response of an unperturbed cell population μ to a stimulus such as a therapeutic agent, we need to identify a map T that explains its evolution into the perturbed population ν. Then, the perturbed state of a cell x is given by $T(x)$. **b.** Developmental processes describe the evolution of a often homogeneous population of undifferentiated cells into various cell lineages. Reconstructing such processes from measurement snapshots can be achieved through sequential alignments $P_{t-1,t}$ that describe the evolution from μ_0 to μ_1, or by identifying the overall global path μ_t capturing the cell differentiation process.

analysis of cell populations fall short in tackling these issues, as they fail to offer comprehensive solutions to two key obstacles: identifying various cell types within a population and tracking the development of each of these types. By providing insights into the heterogeneity of cell populations, single-cell methods partially address the aforementioned challenges, but their destructive nature impedes the recording of the expression of the same cell and its direct descendants across time. Hence, such differentiation processes can only be measured through distinct snapshots of single-cell populations that are *not time-resolved* and *unaligned* (see Fig. 2.1b).

To understand differentiation —the continuous emergence of different cell types and branching events— we need to reconstruct such developmental processes from single-cell measurements that provide us with snapshots of the cell population evolving over time: Given that a cell has a specific expression profile at a time point, where will its descendants likely be at a later time point and where are its likely ancestors at an earlier time point? Schiebinger et al. (2019) for example study reprogramming of fibroblasts to induced pluripotent stem cell (iPSC) (Takahashi and Yamanaka, 2006), by measuring mouse embryonic fibroblast (MEF). Cells at time point t are connected to their ancestors at time $t-1$, by finding the corresponding transport plan $P_{t-1,t}$ between each pair of consecutive time steps (see Fig. 2.1b). To understand the overall dynamic process, however, this thesis aims at identifying the evolution of a measure through recovering a *continuous* path μ_t instead of resorting to the computation of distinct alignments

between consecutive measurement snapshots (Lavenant et al., 2021) (see Fig. 2.1b).

2.4 OPTIMAL TRANSPORT FOR SINGLE-CELL BIOLOGY

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, further cement 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 machine learning standards.

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) is a core element of 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; Bunne et al., 2023b), i.e., the subject of this thesis.

In this chapter, we primarily cast light on the static and dynamic formulations of optimal transport, and simultaneously establish their theoretical nexus by recalling its mathematical history from Monge and Kantorovich to modern Fields Medal winners Villani, Figalli, and Abel Prize recipient Caffarelli in order to provide a solid foundation for the discussion ahead.

3.1 STATIC OPTIMAL TRANSPORT

Optimal transport takes dual roles as it induces a mathematically well-characterized distance measure between distributions as well as provides a geometry-based approach to realize mappings between two probability distributions. In this section, we introduce the mathematical foundations of the **static** OT problem. Further, we provide an extended analysis of the Monge map, which gives an actionable way to flow from one probability distribution to another. We conclude with a complete proof of the cele-

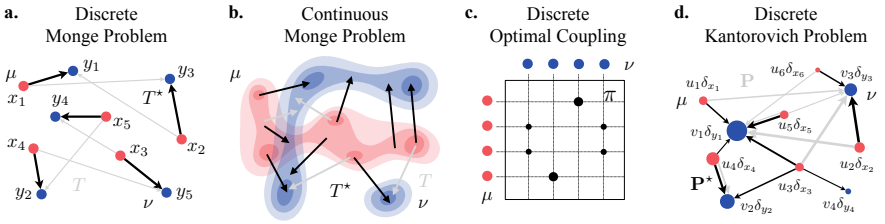


Figure 3.1: Overview on different formulations of the static OT problem for discrete and continuous measures. Monge map for **a.** discrete and **b.** continuous measures μ, ν. The optimal map $T^\star$ minimizes (3.1). **c.** Optimal coupling π (3.2) for discrete measures μ and ν. **d.** Mass splitting principle of the Kantorovich relaxation for discrete measures μ and ν of the optimal transport plan $\mathbf{P}^\star$ and a non-optimal plan $\mathbf{P}$. Figure adapted from Peyré and Cuturi (2019).

brated Brenier theorem. This quintessential result and its particularization to translation-invariant costs lay the foundation of the flurry of neural approaches proposed in the literature. This includes modeling Monge maps as gradients of convex functions parameterized through convex neural networks (Amos et al., 2017; Huang et al., 2021a; Makkuva et al., 2020; Korotin et al., 2021b; Lübeck et al., 2022; Bunne et al., 2022a), i.e., approaches that are a direct consequence of the Brenier theorem and the subject of this thesis, regularizers (Uscidda and Cuturi, 2023), amortized optimization (Amos, 2023; Amos et al., 2023), or entropic maps (Pooladian and Niles-Weed, 2021; Pooladian et al., 2023b; Divol et al., 2022; Cuturi et al., 2023).

3.1.1 *Monge Problem*

In the 18th century "Mémoire sur la théorie des déblais et des remblais", Gaspard Monge sets out to solve what is now known as the Monge problem, posing a seemingly simple, yet fundamentally complex question: Given two quantities of mass located at two different sites, what is the most efficient way to transport one into the other? In more formal terms, provided with two measures $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$, here restricted to measures supported on $\mathbb{R}^d$, Monge's initial approach was to find a map T that pushes one mass onto the other in a way that minimizes the total cost of transport. Given a measurable const function $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, the Monge problem then reads

$$T^\star := \arg \inf_{T\sharp\mu=\nu} \int_{\mathbb{R}^d} c(x, T(x)) d\mu(x) \,. \tag{3.1}$$

For two discrete measures $\mu = \sum_{i=1}^n u_i \delta_{x_i}, \nu = \sum_{j=1}^m v_j \delta_{y_j}$, it seeks a transport map $T : \mathcal{X} \to \mathcal{Y}$ associating each source point x_i to a target point y_j (see

Fig. 3.1a for the discrete and Fig. 3.1b for the continuous setting). The existence of $T^\star$ is guaranteed under fairly general conditions (Santambrogio, 2015, Theorem 1.22), which require that μ and ν have finite ℓ_2 norm, and that μ puts no mass on $(d-1)$ surfaces of class $\mathcal{C}_2$, i.e., the family of continuous functions that have both a continuous first and a continuous second derivative.

3.1.2 *Kantorovich Relaxation*

It was not until the 20th century, however, that the concept found a more tractable development. In 1942, Leonid Kantorovich provided a relaxation to this non-convex and difficult-to-solve problem. Instead of the deterministic matching proposed by Monge, Kantorovich considered probabilistic correspondences that allow for the transportation of mass from a single source point to various target points (mass splitting), resulting in the problem formulation

$$W(\mu,\nu) := \inf_{\pi\in\Pi(\mu,\nu)} \iint c(x,y)\pi(x,y)\mathrm{d}x\mathrm{d}y, \tag{3.2}$$

where $\Pi(\mu,\nu) := \left\{\pi \in \mathcal{P}(\mathcal{X}\times\mathcal{Y}) : P_{\mathcal{X}\sharp}\pi = \mu \text{ and } P_{\mathcal{Y}\sharp}\pi = \nu\right\}$ is the set of couplings on $\mathbb{R}^d \times \mathbb{R}^d$ with respective marginals μ, ν. Given the optimal transport coupling π, the resulting distance $W(\mu,\nu)$ between μ and ν is known as the Wasserstein distance. A visualization of the discrete setting is provided in Fig. 3.1c.

For his work, Kantorovich received the Nobel Prize in economics. The connection of OT to basic questions in economy becomes clear when interpreting μ as a density of resource units, and ν a density of factories, where the coupling π denotes the optimal transportation plan of distributing resources to factories.

Despite its elegance, the Wasserstein distance (3.2) presents a computationally challenging optimization problem. A partial remedy proposed by Cuturi (2013) is to solve regularized optimal transportation problems for an approximate solution. One example of an effective regularization is entropy regularization: For $\varepsilon \geq 0$, set

$$W_\varepsilon(\mu,\nu) = \inf_{\pi\in\Pi(\mu,\nu)} \iint c(x,y)\pi(x,y)\mathrm{d}x\mathrm{d}y \; - \varepsilon H(\pi), \tag{3.3}$$

where $H(\pi) := -\iint \pi(x,y)\log\pi(x,y)\mathrm{d}x\mathrm{d}y$ is the entropy of coupling π. Notice that the definition above reduces to the usual Wasserstein distance

(3.2) when $\varepsilon = 0$. When instantiated on finite discrete measures, such as $\mu = \sum_{i=1}^{n} u_i \delta_{x_i}$ and $\nu = \sum_{j=1}^{m} v_j \delta_{y_j}$, (3.2) translates to a linear program

$$W_\varepsilon(\mu, \nu) \coloneqq \min_{\mathbf{P} \in U(\mu,\nu)} \langle \mathbf{P}, [\|x_i - y_j\|^2]_{ij} \rangle - \varepsilon H(\mathbf{P}), \tag{3.4}$$

where $H(\mathbf{P}) \coloneqq -\sum_{ij} \mathbf{P}_{ij}(\log \mathbf{P}_{ij} - 1)$ and the polytope $U(\mu, \nu)$ is the set of $n \times m$ matrices $\{\mathbf{P} \in \mathbb{R}_+^{n \times m}, \mathbf{P}\mathbf{1}_m = \mu, \mathbf{P}^T \mathbf{1}_n = \nu\}$. Regularization with an entropy term results in a significantly more efficient optimization (Cuturi, 2013) and differentiability w.r.t. the inputs. As a consequence, 3.4 is commonly used as a loss function or evaluation metric in machine learning applications, e.g., for structured prediction (Frogner et al., 2015; Janati et al., 2020a) or generative model fitting (Arjovsky et al., 2017; Salimans et al., 2018; Genevay et al., 2018). While setting $\varepsilon > 0$ yields a faster and differentiable proxy to approximate W_0, it introduces a bias, since $W_\varepsilon(\mu, \mu) \neq 0$ in general.

3.1.3 *Kantorovich Duality*

The Kantorovich formulation (3.2) is a *convex* problem on $\mathcal{P}(\mathcal{X} \times \mathcal{Y})$ and thus admits a dual formulation introduced by Kantorovich (1942), i.e., a constrained concave maximization problem defined as

$$W(\mu, \nu) \coloneqq \sup_{(f,g) \in \Phi_c} \int f \, \mathrm{d}\mu + \int g \, \mathrm{d}\nu, \tag{3.5}$$

where the set of admissible dual potentials is given by $\Phi_c \coloneqq \{(f, g) \in L^1(\mu) \times L^1(\nu) : f(x) + g(y) \leq c(x, y), \forall (x, y) \, d\mu \otimes d\nu \text{ a.e.}\}$. (f, g) is thus a pair of continuous functions, often referred to as *Kantorovich potentials*. An informal interpretation of (3.5) was provided by Caffarelli (2003), revisiting the connection of OT to economics: A logistics company is concerned with transporting products from each resource unit x to a factory y. The transportation company charges $f(x)$ for loading resources at point x and $g(y)$ for unloading it at destination y but is constrained to charge $f(x) + g(y) \leq c(x, y)$. In order to arrange prizes f and g that increase profit, they thus maximize objective (3.5).

The Kantorovich duality (3.5) is a core pillar of optimal transport, powerful due to its generality and computationally attractive as it is easier to store two functions (f, g) than an entire coupling π. In the following, we will introduce the concept of c-transforms, a useful machinery to reduce (3.5) even further into an optimization problem over only one instead of two dual potentials.

Definition 1 (*c*-transform). *c-transforms (also called c-conjugate functions) are generalizations of the Legendre transform from convex analysis defined as*

$$\forall y \in \mathcal{Y}, \quad f^c(y) \coloneqq \inf_{x \in \mathcal{X}} c(x,y) - f(x)\,. \tag{c-transform}$$

The definition of f^c is also often referred to as a "Hopf-Lax formula". Similarly to the c-transform of $f : \mathcal{X} \to \mathbb{R}$, we can define the $\bar{c}$-transform of $g : \mathcal{Y} \to \bar{\mathbb{R}}$ by

$$\forall x \in \mathcal{X}, \quad g^{\bar{c}}(y) \coloneqq \inf_{y \in \mathcal{Y}} c(x,y) - g(y)\,,$$

where $\bar{c}(y,x) = c(x,y)$.

Further, f is a *$\bar{c}$-concave* function if there exists a g such that $f = g^{\bar{c}}$, and analogously, a function g is said to be *c-concave* if there is a function f such that $g = f^c$. When $\mathcal{X} = \mathcal{Y}$ and c is symmetric no distinction between c and $\bar{c}$ is necessary.

Using the concept of *c*-transforms, we can reduce (3.5) to a single potential: Assume we keep dual potential f fixed and given the constraint of the dual formulation(3.5)

$$\begin{aligned} f(x) + g(y) &\leq c(x,y) \\ g(y) &\leq c(x,y) - f(x)\,, \end{aligned}$$

we can see that the "best" potential g is given by the Definition 1 of f

$$g(y) \leq \inf_x c(x,y) - f(x).$$

Then, doing an alternate optimization on either f or g, we replace the dual potentials (f,g) with (f,f^c), and then $(f^{c\bar{c}},f^c)$, whilst preserving the constraints and increasing the value of the integrals of (3.5). Although one could continue this alternate optimization further, the invariance property $f^{c\bar{c}c} = f^c$ for any f shows that one can only "improve" once the dual potential using *c*-transforms, resulting in the semi-dual formulation of optimal transport

$$f^\star \coloneqq \arg \max_{f\ c\text{-concave}} \int f \mathrm{d}\mu + \int f^c \mathrm{d}\nu\,, \tag{3.6}$$

where $f^\star$ is the optimal dual function and *c*-concave.

3.1.4 *Geometry of Optimal Transport*

Following Gangbo and McCann (1996), $f^\star$ can be linked to the optimal transport map $T^\star$ via the following result:

Theorem 1 (Gangbo-McCann Theorem). *Given a cost function c, the relationship between the optimal transport map $T^\star : \mathcal{X} \to \mathcal{X}$ and the c-concave function $f^\star$ denoting the optimal dual potential is given by the expression*

$$T^\star(x) = \nabla_1 c(x, \cdot)^{-1} \circ \nabla f^\star(x)\,. \tag{3.7}$$

Thus, map T depends explicitly on the gradient of the cost, or rather on its inverse map (Gangbo and McCann, 1995).

Following Gangbo and McCann (1996) and considering translation-invariant costs[2] generated by a convex potential $h : \mathbb{R}^d \to \mathbb{R}$, i.e., $c(x, y)$=$h(x$-$y)$, this reduces to

$$T^\star(x) = x - \nabla h^* \circ \nabla f^\star(x)\,, \tag{3.8}$$

where h^* is the Legendre-transform of h given by

$$\forall z, \quad h^*(z) := \sup_x \langle x, z\rangle - h(x)\,. \qquad \text{(Legendre transform)}$$

Proof. Santambrogio (2015, Theorem 1.39) proves that the solutions of (3.2) and (3.5) are equivalent, i.e.,

$$\int_{\mathcal{X}} f \mathrm{d}\mu + \int_{\mathcal{Y}} f^c \,\mathrm{d}\nu = \int_{\mathcal{X}\times\mathcal{Y}} (f(x) + f^c(y))\,\mathrm{d}\pi(x,y) = \int_{\mathcal{X}\times\mathcal{Y}} c(x,y)\mathrm{d}\pi(x,y)\,.$$

Then, a point (x_0, y_0) in the coupling π, i.e., $(x_0, y_0) \in \mathrm{supp}(\pi)$, necessarily satisfies the constraint of the dual problem (3.5)

$$\pi^\star(x_0, y_0) > 0 \Leftrightarrow f^\star(x_0) + g^\star(y_0) = c(x_0, y_0)\,.$$

Replacing g by the c-transform of f, we have

$$\begin{aligned} &\Leftrightarrow f^\star(x_0) + f^{c\star}(y_0) = c(x_0, y_0) \\ &\Leftrightarrow f^{c\star}(y_0) = c(x_0, y_0) - f^\star(x_0)\,. \end{aligned}$$

Yet, by definition of the c-transform, $f^{c\star}$ is given by

$$\Leftrightarrow f^{c\star}(y_0) = \inf_z c(z, y_0) - f^\star(z)\,.$$

2 A cost is translation-invariant if $c(x, y) = h(x - y)$ for $h(z) = h(-z)$.

Thus, x_0 is a minimizer of the above expression and $\nabla_1 c\,(x_0, y_0) = \nabla f^\star\,(x_0)$. Therefore, after inversion, we have $y_0 = \nabla_1 c(x, \cdot)^{-1} \circ \nabla f^\star(x)$ and

$$T^\star(x) = x - \nabla h^* \circ \nabla f^\star(x)\,.$$

Applying this result to $c(x, y) = h(x - y)$, we get

$$\begin{aligned}\nabla_1 c(x, y) &= \nabla h(x - y)\\ \nabla_1 c(x, \cdot) &= \nabla h(x - \cdot)\\ \nabla_1 c(x, \cdot)^{-1} &= x - (\nabla h)^{-1}(\cdot)\,.\end{aligned}$$

Note that $(\nabla h)^{-1}$ is equivalent to ∇h^* with the convex conjugate h^* and thus,

$$\nabla_1 c(x, \cdot)^{-1} = x - \nabla h^*(\cdot)\,.$$

□

Alternative formulations that relate the Kantorovich setting (with general costs) (3.5) to that of Monge (3.1) were proposed by Rüschendorf (1991a,b); Caffarelli (1996).

The case of the squared Euclidean cost[3] $c(x, y) = \frac{1}{2}\|x - y\|^2$ in $\mathcal{X} = \mathcal{Y} = \mathbb{R}^d$ deserves a special attention. Taking advantage of the particular form of the quadratic cost function, we can expand the constraint of (3.5) such that

$$f(x) + g(y) \leq \frac{1}{2}\|x - y\|_2^2 \iff \left[\frac{1}{2}\|x\|_2^2 - f(x)\right] + \left[\frac{1}{2}\|y\|_2^2 - g(y)\right] \geq \langle x, y\rangle$$

and subsequently reparameterize $\varphi(x) \coloneqq \frac{1}{2}\|x\|_2^2 - f(x)$ and $\psi(y) \coloneqq \frac{1}{2}\|y\|_2^2 - g(y)$. Mirroring the same logic as for the c-transform, we derive the semi-dual in the Euclidean setting

$$\inf_{\varphi\,\text{convex}} \int \varphi \mathrm{d}\mu + \int \varphi^* \mathrm{d}\nu\,. \tag{3.9}$$

Following the double convexification trick as outlined in Villani (2003, Lemma 2.10), we see that applying the Legendre transform twice yields function pair $(\varphi^{**}, \varphi^*)$. As each of them is defined as the supremum of a family of linear functions, the result is an optimization problem over two convex lower semi-continuous functions.

3 For elegance, we consider $c(x, y) = \frac{1}{2}\|x - y\|^2$ instead of $c(x, y) = \|x - y\|^2$.

Similarly, for the special case of the squared Euclidean distance the Theorem 1 (3.8) with $h = \frac{1}{2}\|\cdot\|_2^2$ implies that $\nabla h = \nabla h^* = \mathrm{Id}$, and thus

$$T(x) = x - \nabla f(x) = \nabla\left(\frac{1}{2}\|x\|_2^2 - f(x)\right)(x) = \nabla\varphi(x),$$

where we again reparameterize $\varphi(x) := \frac{1}{2}\|x\|_2^2 - f(x)$ and $\varphi(x) = \frac{1}{2}\|x\|_2^2 - f(x)$ can be shown to be convex.

This connection presents a well known fact that has been investigated first by Brenier, establishing for the special case of the Euclidean distance the equivalence of the Monge (3.1) and Kantorovich formulation (3.2), the uniqueness of the optimal coupling π, and instantiating that there must exist a unique (up to the addition of a constant) potential $\varphi^\star : \mathbb{R}^d \to \mathbb{R}$ such that $T^\star = \nabla\varphi^\star$. This theorem has far-reaching implications: When seeking optimal transport maps, it is sufficient, to restrict the computational effort to seek a "good" convex potential φ, such that its gradient pushes μ towards ν. Let us state the celebrated Brenier theorem (1987) in more formal terms:

Theorem 2 (Brenier Theorem). *In the setting where both $\mathcal{X}$ and $\mathcal{Y}$ are equal to $\mathbb{R}^d$, and the cost function $c(x,y) = \|x-y\|^2$ is employed, and at least one of the two input measures μ possesses a density p_μ in relation to the Lebesgue measure, then there exists a unique optimal solution π in the Kantorovich formulation (3.2). This solution is exclusively supported on the graph $(x, T(x))$ of Monge map $T : \mathbb{R}^d \to \mathbb{R}^d$. In other terms, we can express π as $(\mathrm{Id}, T)_\sharp\alpha$, meaning that for any function h belonging to the set $\forall h \in \mathcal{C}(\mathcal{X}\times\mathcal{Y})$, the following equality holds*

$$\int_{\mathcal{X}\times\mathcal{Y}} h(x,y)\mathrm{d}\pi(x,y) = \int_{\mathcal{X}} h(x,T(x))\mathrm{d}\mu(x).$$

Moreover, this map T is uniquely determined by the gradient of a convex function φ, denoted as $T(x) = \nabla\varphi(x)$. The function φ is the unique convex function, up to an additional constant, for which $(\nabla\varphi)_\sharp\mu = \nu$.

Corollary 1. *Under the assumption of Theorem 2, $\nabla\varphi$ is the unique solution to the Monge transportation problem (3.1), i.e.,*

$$\int_{\mathbb{R}^d}\|x - \nabla\varphi(x)\|^2\mathrm{d}\mu(x) = \inf_{T_\sharp\mu=\nu}\int_{\mathbb{R}^d}\|x - T(x)\|^2\mathrm{d}\mu(x). \tag{3.10}$$

Theorem 2 by Brenier (1987, 1991) has been exploited to propose neural OT solvers (Taghvaei and Jalali, 2019; Makkuva et al., 2020; Korotin et al., 2021a; Bunne et al., 2022b; Alvarez-Melis et al., 2022; Mokrov et al., 2021; Amos, 2023) and recurrently permeates this thesis (see Chapter 4 and 5),

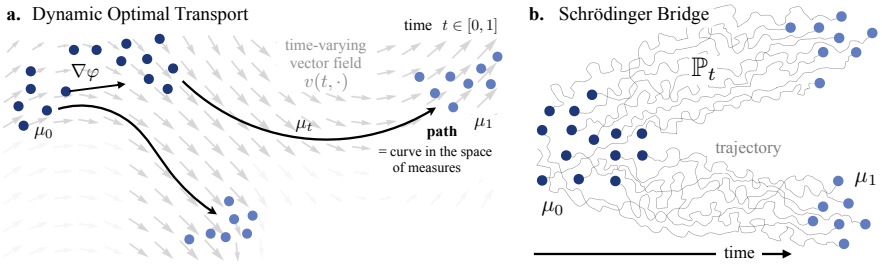


Figure 3.2: Overview on different formulations of the dynamic OT problem. a. We can model the evolution of a measure μ_t over time as minimal path on a time-varying vector field $v(t, \cdot)$ or according to the gradient of a convex potential $\nabla\varphi$. **b.** Alternatively, taking a stochastic perspective, we can study the dynamic formulation of the entropy-regularized OT problem (3.3) and find a stochastic process $\mathbb{P}_t$ that describes the particle dynamics from μ_0 to μ_1.

proving its essential nature in multiple instances and modern developments of optimal transport. Further, it presents an elegant way to solve the Monge problem in a geometric sense and has profound implications for the dynamic version of the problem, which we will study next.

3.2 DYNAMIC OPTIMAL TRANSPORT

We have hitherto engaged with the *static* optimal transport problem, establishing a solid foundation upon which to build more desirable dynamic formulations. In fact, the roots of these dynamic formulations are embedded within the static OT framework: As posited by Benamou and Brenier (2000), the dynamic formulation "was already implicitly contained in the original problem addressed by Monge", where "eliminating the time variable was just a clever way of reducing the dimension of the problem." When reintroducing time in the dynamic version, the optimal transport map becomes a time-dependent flow capable of describing the evolution of a measure over time.

In this section, we will cover several perspectives and frameworks of the **dynamic** OT problem: As mentioned earlier, the Brenier theorem forms a critical bridge that connects the static and dynamic formulation, perpetuated in the Monge-Ampère equation. Further, Benamou and Brenier (2000) introduce how the dynamic point of view offers an alternate and intuitive interpretation of optimal transport with links to fluid dynamics. The resulting framework surprisingly leads to a convex optimization problem that can be parameterized through continuous normalizing flows (Tong et al.,

2020; Chen et al., 2018) or flow matching frameworks (Lipman et al., 2023; Liu et al., 2022b; Pooladian et al., 2023a; Albergo et al., 2023). We further highlight the connections of OT to PDEs such as Fokker-Planck-like equations through the Jordan, Kinderlehrer, and Otto scheme. Lastly, moving beyond PDEs and taking a stochastic control perspective, we will introduce the notion of the Schrödinger bridge problem.

3.2.1 *Monge-Ampère Equation*

As a direct consequence of Theorem 2, if $T(x) = \nabla\varphi(x)$, φ being smooth and strictly convex, and μ and ν absolutely continuous with densities p_μ and p_ν, we can express $T_\sharp\mu = \nu$ in a nonlinear partial differential equation (PDE) form. More concretely, as a consequence of a simple change-of-variable computation, φ is a solution of the Monge-Ampère equation that reads

$$\det\left(\partial^2\varphi(x)\right) p_\nu(\nabla\varphi(x)) = p_\mu(x), \tag{3.11}$$

where $\partial^2\varphi(x) \in \mathbb{R}^{d\times d}$ is the Hessian of φ, describing the continuous evolution from μ to ν. First studied by Monge in 1781 and later by Ampère in 1819, this nonlinear partial differential equation arises in several problems from analysis to geometry, for example, in the Weyl and Minkowski problems in differential geometry of surfaces. The regularity of the solutions of (3.11), with implications on regularity results of the optimal transport map T, has been subject of a series of works by Caffarelli in the 1990s, for which he was awarded the Abel Prize in 2023, as well as more recently by Figalli, recognized with the Fields Medal in 2018.

3.2.2 *Benamou-Brenier Formulation*

Avoiding solving (3.11) directly, Benamou and Brenier (2000) introduce an alternative numerical framework by connecting the optimal mass transfer problem to continuum mechanic frameworks. Deviating from the previous notation of (μ, ν), in the following sections we study the dynamic problem via the evolution from measure μ_0 at time $t = 0$ to μ_1 at $t = 1$. In the setting $\mathcal{X} = \mathcal{Y} = \mathbb{R}^d$ with the squared Euclidean cost $c(x, y) = \|x - y\|^2$, the solution of (3.2) then coincides with finding the minimal path $(\mu_t)_{t=0}^1$, or more concretely, a curve in the space of measures, minimizing a total length. Such path μ_t can be described through a time-varying vector field

$v(t,\cdot)$ which moves particles around, satisfying the continuity equation in fluid dynamics or conservation of mass formula

$$\frac{\partial \mu_t}{\partial t} + \nabla \cdot (\mu_t v) = 0, \qquad \mu_{t=0} = \mu_0, \mu_{t=1} = \mu_1 \,, \tag{3.12}$$

where the vector field $v(t,\cdot)$ denotes the speed and $\mu_t v(t,\cdot) = J_t$ corresponds to the momentum. Reformulating the optimal transportation problem in a differential way, an "Eulerian" formulation inspired by fluid mechanics, will be crucial for the subsequent study of dynamical problems. Every curve μ_t describing the evolution of the measure over time can be interpreted as the fluid flow along a family of vector fields. We are searching for the vector field $v(t,\cdot)$ that (i.) satisfies the conservation of mass (3.12), and (ii.) minimized the path length. The infinitesimal length of such a vector field can be computed via

$$\|v\|_{\ell^2(\mu_t)} = \left(\int_{\mathbb{R}^d} \|v(t,x)\|^2 \, \mathrm{d}\mu_t(x) \right)^{1/2}. \tag{3.13}$$

resulting, in the case of $\mathcal{X} = \mathcal{Y} = \mathbb{R}^d$ and $c(x,y) = \|x-y\|^2$, in the minimal-path reformulation of (3.2)

$$\begin{aligned} W(\mu_0,\mu_1) = \inf_{(\mu_t,v)} \int_0^1 \int_{\mathbb{R}^n} \frac{1}{2} \|v(t,x)\|^2 d\mu_t(x) dt \\ \frac{\partial \mu_t}{\partial t} + \nabla \cdot (v\mu_t) = 0 \\ \mu_{t=0} = \mu_0, \mu_{t=1} = \mu_1 \,. \end{aligned} \tag{3.14}$$

Thus, path μ_t describes the time-evolving density of a set of particles moving continuously with velocity $v(t,\cdot)$. Taking the perspective of fluid dynamics, (3.13) can also be interpreted as the *kinetic energy* of the particles. The Benamou-Brenier formulation (3.14) then selects the vector field v that minimizes the total efforts or the total kinetic energy one has to spend in order to move particles around according to the vector field v.

A particularly important case occurs when there exists an optimal Monge map $T : \mathbb{R}^d \to \mathbb{R}^d$ with $T_\sharp \mu_0 = \mu_1$ (see Theorem 2): The solution of the time-dependent OT problem (3.14) then coincides with McCann's displacement interpolation between two measures. Reciting Theorem 2, with $T = \nabla\varphi$, μ_t is equal to McCann's interpolation between μ_0 and μ_1 given by

$$\mu_t = [(1-t)I + t\nabla\varphi]_\sharp \mu_0 = [(1-t)I + tT]_\sharp \mu_0 \,. \tag{3.15}$$

Despite their simplicity, this concept possesses remarkable applications beyond the realm of optimal transport (Bonneel et al., 2011). In particular, its interpretation as a geodesic formula in Riemannian geometry is thoroughly discussed in Gangbo and McCann (1996) and serves as a pivotal link to the subsequent discussion.

3.2.3 *Jordan-Kinderlehrer-Otto Flows*

The time-dependent Benamou-Brenier formulation (3.14) not only provides us with a more complete description of optimal transport but also the discovery that the resulting path $(\mu_t)_{t=0}^1$ may be seen as a constant-speed geodesic interpolating between population μ_0 and μ_1 in the space of measures, i.e., a Wasserstein geodesic. When studying dynamic processes in biomedicine, however, phenomena such as cellular differentiation in developmental processes, tissue formation, or cell migration involve intricate spatiotemporal dynamics that cannot be adequately captured by solely studying the interpolation between two measures μ_0 and μ_1. Instead, many phenomena in biology and physics can be modeled through an energy functional J such that the minimization of J describes the observed dynamics of the studied system —a concept known as gradient flows. At their core, gradient flows provide a powerful framework for understanding the evolution of functions or systems towards an optimal state through the direction of the steepest descent of a function J. More concretely, gradient flows capture the intuitive idea of objects moving in a direction that decreases their energy, seeking a state of minimum potential or maximum stability. In the following, we will study gradient flows in the Euclidean setting before considering generalizations to arbitrary measures that allow studying the evolution of populations over time.

EUCLIDEAN CASE. Considering the evolution of a vector x over time in Euclidean space. Provided with a smooth functional J, this can be realized through the standard gradient descent (forward) scheme

$$x_{t+1} \coloneqq x_t - \tau \nabla J(x_t),$$

where τ is the step size. The resulting sequence $x_0, \dots, x_{t-1}, x_t, x_{t+1}, \dots, x_T$ then describes the trajectory of a single particle x over time. For non-smooth functions, on can resort to a proximal scheme, i.e.,

$$x_{t+1} \coloneqq \operatorname{Prox}_{\tau J}^{\|\cdot\|}(x_t) \coloneqq \operatorname*{argmin}_x \frac{1}{2\tau} \|x - x_t\|^2 + J(x).$$

The proximal scheme can thus be seen as a *backward* Euler discretization of the gradient flow.

WASSERSTEIN CASE. When studying the evolution of a population or measure μ_t over time, however, we need to resort to optimal transport metrics $W(\cdot,\cdot)$ (3.2) instead of the ℓ_2-norm $\|\cdot\|^2$. Considering functionals J that take a measure or population as input, a gradient flow of μ w.r.t. to J can be similarly expressed through forward and backward schemes. Assuming $J(\mu) \coloneqq \int E(x)d\mu(x)$, i.e., ignoring particle interaction, the forward scheme reads

$$\mu_{t+1} \coloneqq (I - \tau\nabla E)_\sharp \mu_t$$

with the corresponding backward formulation defined as

$$\mu_{t+1} \coloneqq \operatorname*{argmin}_{\mu} \frac{1}{2\tau} W(\mu, \mu_t) + J(\mu). \tag{3.16}$$

This implicit time stepping is a useful tool to construct continuous flows: In the limit $\tau \to 0$ the resulting sequence $\{\mu_t\}_{t=0}^T$ approximates a continuous flow μ_t, i.e., a path in the Wasserstein space, and can thus be seen as the analogy of the usual proximal descent scheme, tailored for probability measures (Santambrogio, 2015, p.285)

Interest in Wasserstein gradient flows was sparked by the seminal work of Jordan, Kinderlehrer, and Otto (1998) who studied diffusion processes under the lens of the OT metric (see also Ambrosio et al., 2006). For a broad class of potentials, J and provided with an initial distribution μ_0, the resulting time-discrete, iterative variational scheme induced by the so-called *JKO step* (3.16) reconstructs the evolution of measure μ_t over time. As $\tau \to 0$, the solution of the time-discretized gradient flow converges to the solution of a corresponding PDE, and the resulting evolutions are often referred to as *JKO flows*.

In fact, following Otto (2001) on the calculus of optimal transport (Otto calculus), a large class of partial differential equation may then be viewed as gradient flows on the Wasserstein space (Jordan et al., 1998). For instance, the standard heat equation of physics, i.e.,

$$\frac{\partial \mu_t}{\partial t} = \Delta \mu_t,$$

with Δ being the spatial Laplacian, can be expressed as a gradient flow of the energy $J(\mu) = \int \mu_t(x) \log \mu_t(x) \mathrm{d}x$, i.e., Gibbs-Boltzmann's famous functional with the physical interpretation of the negative of an entropy.

Table 3.1: Equivalence between gradient flows and PDEs where the gradient flow of flow functional $J(\mu_t)$ in Wasserstein space satisfies the corresponding PDE (Alvarez-Melis et al., 2022; Villani, 2003). The function $f : \mathbb{R} \to \mathbb{R}$ is convex and superlinear and $V, W : \mathcal{X} \to \mathbb{R}$ are convex and sufficiently smooth.

Class	**PDE** $\frac{\partial \mu_t}{\partial t} =$	**Flow Functional** $J(\mu_t) =$
Heat Equation	$\Delta \mu_t$	$\int \mu_t(x) \log \mu_t(x) \mathrm{d}x$
Advection	$\nabla \cdot (\mu_t \nabla V)$	$\int V(x) \mu_t(x) \mathrm{d}x$
Fokker-Planck	$\Delta \mu_t + \nabla \cdot (\mu_t \nabla V)$	$\int \mu_t(x) \log \mu_t(x) \mathrm{d}x + \int V(x) \mu_t(x) \mathrm{d}x$
Porous Media	$\Delta \left(\mu_t^m\right) + \nabla \cdot (\mu_t \nabla V)$	$\frac{1}{m-1} \int \mu_t(x)^m \, \mathrm{d}x + \int V(x) \mu_t(x) \mathrm{d}x$
Advection, Diffusion, and Interaction	$\nabla \cdot [\mu_t \left(\nabla f'(\mu_t) + \nabla V + (\nabla W) * \mu_t\right)]$	$\int V(x) \mu_t(x) \mathrm{d}x + \int f(\mu_t(x)) \mathrm{d}x$ $+\frac{1}{2} \iint W\left(x - x'\right) \mu_t(x) \mu_t\left(x'\right) \mathrm{d}x \mathrm{d}x'$

Among further examples displayed in Table 3.1 (Alvarez-Melis et al., 2022; Villani, 2003), a classic subject of the theory of PDEs also comprises the linear Fokker-Planck equation

$$\frac{\partial \mu_t}{\partial t} = \Delta \mu_t + \nabla \cdot (\mu_t \nabla V), \tag{3.17}$$

that is connected to flow functional

$$J(\mu_t) = \int \mu_t(x) \log \mu_t(x) \mathrm{d}x + \int V(x) \mu_t(x) \mathrm{d}x \,.$$

Here, the first term again represents the negative Gibbs-Boltzmann entropy and the second term plays the role of an energy functional with a smooth potential function $V : \mathbb{R}^d \to \mathbb{R}$.

Thus, JKO flows have found application in inferring the evolution of populations over time, crucial in many scientific disciplines, for instance in biomedicine to reconstruct cellular dynamics from observations (Bunne et al., 2022b; Alvarez-Melis et al., 2022; Mokrov et al., 2021; Benamou et al., 2016a). In fact, this formulation is particularly interesting for studying dynamical systems in biomedicine, as the exact expression of the PDE corresponding to functional J does not need to be known. Instead, we can propose parameterizations of energy functional J that can be learned from data, an idea we will explore in Section 6.2.2. While providing a general framework for studying general and complex population dynamics, each step of the JKO scheme (3.16) is costly as it requires solving a minimization problem involving the Wasserstein distance (3.2). Beyond introducing learning schemes for functional J, we will thus introduce novel efficient and differentiable schemes for solving JKO flows in Section 6.2.1.

3.2.4 *Stochastic Control Perspective*

Benamou-Brenier motivated the introduction of the dynamic optimal transport problem from the perspective of fluid dynamics. As we shall see, both the OT problem (3.2) and its regularized version (3.3) can be viewed as stochastic *optimal control* problems. Control theory at the heart is concerned with finding optimal policies for dynamic systems subject to constraints. Despite wide-ranging progress on both the theory and applications, deploying control theory to large-scale and often unknown systems remains a grand challenge. As we will explore in the following, stochastic optimal control problems to regulate dynamic systems emerge from the theory of optimal transport (Santambrogio, 2015) that provides a geometric variational framework for studying flows of distributions on metric spaces (Chen et al., 2021a). These theoretical concepts build the foundation of recently developed deep learning architectures employed as generative models (Song et al., 2021; De Bortoli et al., 2021b) or for studying the evolution of dynamical systems over time (Chen et al., 2022a; Bunne et al., 2022b; Vargas et al., 2021). Further, celebrated control principles such as the Pontryagin maximum principle have been emphasized repeatedly in neural ordinary differential equation (ODE) (Chen et al., 2018) and stochastic differential equation (SDE) works (Jia and Benson, 2019).

... on Optimal Transport

Following Chen et al. (2021a,b), we will establish this stochastic control viewpoint by studying the Benamou-Brenier formulation using elementary control considerations. For this, we consider a system with state distribution $\mathrm{d}X_t = v\,(t, X_t)\,\mathrm{d}t$ and initial state $X_0 \sim \mu_0$. Provided with a time-dependent feedback control $v(t, \cdot)$, the objective of (3.14) has the following stochastic interpretation

$$\int_0^1 \int_{\mathbb{R}^n} \frac{1}{2}\|v(t,x)\|^2 \mathrm{d}\mu_t(x)\mathrm{d}t = \mathbb{E}\left\{\int_0^1 \frac{1}{2}\,\|v(t,X_t)\|^2\,\mathrm{d}t\right\},$$

resulting in the stochastic control formulation of the OT problem

$$\inf_{v\in\mathcal{V}} \mathbb{E}\left\{\int_0^1 \frac{1}{2}\|v(t,X_t)\|^2\mathrm{d}t\right\} \tag{3.18}$$

$$\mathrm{d}X_t = v\,(t, X_t)\,\mathrm{d}t \tag{3.19}$$

$$X_0 \sim \mu_0, \quad X_1 \sim \mu_1.$$

$\mathcal{V}$ here represents the family of admissible state feedback control strategies. Typically, in a density control problem, the objective is to guide a dynamical system from an initial state X_0 characterized by μ_0 to a desired state μ_1 with minimum cost and control that is a member of the set of admissible actions, i.e., $v \in \mathcal{V}$. The above strategy, however, differs from standard optimal control in the added constraint on the terminal state distribution and the absence of a terminal penalty.

... on Regularized Optimal Transport

Similarly, the regularized OT problem (3.3) can be cast as a stochastic control problem.

$$\inf_{v \in \mathcal{V}} \mathbb{E}\left\{ \int_0^1 \frac{1}{2} \|v(t, X_t)\|^2 \mathrm{d}t \right\} \tag{3.20}$$

$$\mathrm{d}X_t = v(t, X_t)\mathrm{d}t + \sigma \mathrm{d}\mathbb{W}_t \tag{3.21}$$

$$X_0 \sim \mu_0, \quad X_1 \sim \mu_1,$$

where $\mathbb{W}_t$ denotes a Wiener process, i.e., standard white noise. Different to (3.19), however, (3.21) is a stochastic diffusion process.

Besides, this problem exhibits a fluid-dynamic interpretation, i.e.,

$$\inf_{(\mu_t, v)} \int_0^1 \int_{\mathbb{R}^n} \frac{1}{2} \|v(t,x)\|^2 \mathrm{d}\mu_t(x)\mathrm{d}t \tag{3.22}$$

$$\frac{\partial \mu_t}{\partial t} + \nabla \cdot (v\mu_t) - \frac{1}{2}\sigma^2 \Delta \mu_t = 0 \tag{3.23}$$

$$\mu_{t=0} = \mu_0, \mu_{t=1} = \mu_1\,,$$

where Δ denotes the Laplace operator. (3.23) is the Fokker-Planck equation capturing the state distribution evolution. As $\sigma^2 \to 0$, the solution to this problem converges to the one of the Benamou-Brenier problem (3.14) (Mikami and Thieullen, 2008). For an extended discussion, see Dai Pra (1991); Mikami (2000, 2002).

3.2.5 *Schrödinger Bridges*

Interestingly, Eq. (3.20) first emerged in a very different setting. In his work "Über die Umkehrung der Naturgesetze" published in 1931, Schrödinger studied the most likely random evolution between two marginals, i.e., two point clouds of diffusive particles. His Gedankenexperiment is best

illustrated through a population of independent and identically distributed particles in $\mathbb{R}^d$ observed at $t = 0$ as the empirical distribution μ_0, and again at $t = 1$ as μ_1. To describe the mostly likely dynamics of these particles over time, we aim at finding the stochastic process $\mathbb{P}_t$ on $[0,1]$ such that $\mathbb{P}_0 = \mu_0, \mathbb{P}_1 = \mu_1$.

Provided with some prior knowledge of a reference process $\mathbb{Q}_t$, e.g., that the underlying dynamics follow a Brownian motion (BM), we aim to identify the stochastic process $\mathbb{P}_t$ that best describes the particles evolution, i.e., minimizes the overall relative entropy

$$\min_{\mathbb{P}_0=\mu_0,\, \mathbb{P}_1=\mu_1} D_{\mathrm{KL}}(\mathbb{P}_t \| \mathbb{Q}_t) = \int_{\mathcal{C}[0,1]} \log\left(\frac{d\mathbb{P}_t}{d\mathbb{Q}_t}\right) d\mathbb{P}_t, \tag{3.24}$$

where $\frac{d\mathbb{P}_t}{d\mathbb{Q}_t}$ denotes the Radon-Nikodym derivative and $\mathcal{C}[0,1]$ the continuous paths on $\mathbb{R}^d$ over the time interval $[0,1]$. More concretely, to find $\mathbb{P}_t$, Schrödinger (1931, 1932) considers the objective (3.24) as the "mostly likely process" that explains the marginal distributions $\mathbb{P}_0, \mathbb{P}_1$ relative to reference process $\mathbb{Q}_t$. This *KL-minimization* problem is thus called the (generalized) Schrödinger bridge. This idea generalizes verbatim to any reference process $\mathbb{Q}_t$. Unfortunately, in most applications, notably biology, we often have little to no prior information about the underlying process $\mathbb{P}_t$ (Liberali et al., 2014), a problem tackled in Chapter 7.

Recovering the stochastic calculus of variations formulation of the Schrödinger bridge (3.20) can be achieved via the Girsanov theorem which tells us how stochastic processes behave under changes in measure. The equivalence between both formulations can be then established via

$$\frac{\mathrm{d}\mathbb{P}_t}{\mathrm{d}\mathbb{P}_t^{v=0}} = \exp\left\{\int_0^1 \frac{1}{\sigma} v(t, X_t^v) \cdot \mathrm{d}\mathbb{W}_t + \int_0^1 \frac{1}{2\sigma^2} \|v(t, X_t^v)\|^2 dt\right\}$$

$$\text{and thus } D_{\mathrm{KL}}(\mathbb{P}_t \| \mathbb{P}_t^{v=0}) = \mathbb{E}\left\{\int_0^1 \frac{1}{2\sigma^2} \|v(t, X_t)\|^2 dt\right\},$$

where $\mathbb{P}_t$ and $\mathbb{P}_t^{v=0}$ denote a controlled process (with control v) and an uncontrolled process, i.e., with $v(t,\cdot) = 0$, respectively. In other words, the relative entropy between the stochastic process describing the particle dynamics and the reference process is equal to the control energy (scaled by $\frac{1}{\sigma^2}$).

OPTIMALITY CRITERIA. Classical strategies for solving (3.20) commonly replace the boundary constraint $X_1 \sim \mu_1$ with a penalty or artificial ter-

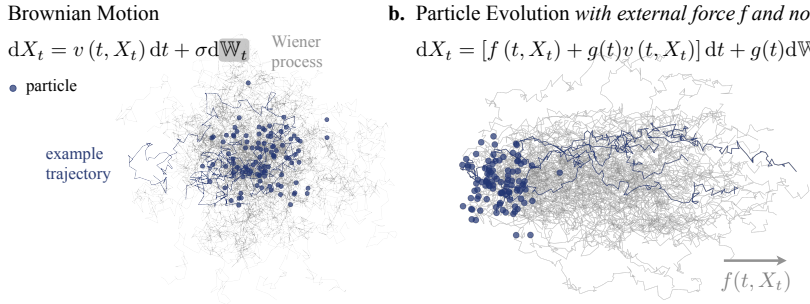


Figure 3.3: Comparison of different SDE classes. a. Particles evolve according to simple Brownian motion in all directions depending on the noise level σ. **b.** A particle evolution with an external speed f, here exemplified through a horizontal drift, and time-dependent noise g. The initial location of the particles is denoted as a blue dot, example trajectories are highlighted by blue lines.

minal cost, thus transforming (3.20) to standard stochastic optimal control formulations. The resulting optimality conditions are

$$\frac{\partial \psi}{\partial t} = -\frac{1}{2}\|\nabla \psi\|^2 - \frac{1}{2}\sigma^2 \Delta \psi \tag{3.25}$$

$$\frac{\partial \mu_t}{\partial t} = -\nabla \cdot (\mu_t \nabla \psi) + \frac{1}{2}\sigma^2 \Delta \mu_t \tag{3.26}$$

with value function $\psi(t, x)$ and μ_t being the associated optimal marginal density. Here, $v \equiv \nabla \psi$ and $\psi(1, \cdot)$ is equivalent to the terminal cost. Further, Eq. (3.25) is a second-order Hamilton-Jacobi-Bellman equation. After applying the Hopf-Cole transform $(\psi, \mu_t) \to (\Phi, \widehat{\Phi})$, we obtain the SB system associated to the SDE class in (3.21), which is given by

$$\begin{aligned} \frac{\partial \Phi}{\partial t} &= -\frac{1}{2}\sigma^2 \Delta \Phi \quad \text{s.t.} & \Phi(0, \cdot)\widehat{\Phi}(0, \cdot) &= \mu_0, \\ \frac{\partial \widehat{\Phi}}{\partial t} &= \frac{1}{2}\sigma^2 \Delta \widehat{\Phi} & \Phi(1, \cdot)\widehat{\Phi}(1, \cdot) &= \mu_1. \end{aligned} \tag{3.27}$$

i.e., a backward Kolmogorov and a Fokker-Planck equation, respectively. The optimal control is then given by $v(t, X_t) = \sigma^2 \nabla \log \Phi(t, x)$.

Generalizations to Other SDE Classes

To describe complex biological processes, however, we need to consider SDE classes comprising nonlinear drifts, affine control, and time-varying

diffusion. In the following, let us consider SDEs with an external speed $f(\cdot, t) : \mathbb{R}^d \to \mathbb{R}^d$, time-dependent diffusion $g(t) \in \mathbb{R}$, and standard Wiener process $\mathbb{W}_t \in \mathbb{R}^d$[4]. Caluya and Halder (2021); Chen et al. (2022a) provide a generalization of the above framework that reads

$$\inf_{v \in \mathcal{V}} \mathbb{E}\left\{ \int_0^1 \frac{1}{2} \|v(t, X_t)\|^2 dt \right\} \tag{3.28}$$

$$\mathrm{d}X_t = \left[f\left(t, X_t\right) + g(t) v(t, X_t) \right] \mathrm{d}t + \sigma g(t) \mathrm{d}\mathbb{W}_t \tag{3.29}$$

$$X_0 \sim \mu_0, \quad X_1 \sim \mu_1,$$

with $g(t)$ being uniformly lower-bounded and $f(t, X_t)$ satisfying Lipschitz conditions with at most linear growth in x. The effect of adding an external force f, here exemplified through a horizontal drift, compared to standard Brownian motion is visualized in Fig. 3.3.

OPTIMALITY CRITERIA. Again, we recover the optimality criteria via a Hopf-Cole transform of (3.28) resulting in

$$\frac{\partial \Phi}{\partial t} = -\nabla \Phi^T f - \frac{1}{2} \sigma^2 g^2 \Delta \Phi \qquad \text{s.t.} \qquad \Phi(0, \cdot) \widehat{\Phi}(0, \cdot) = \mu_0, \tag{3.30}$$

$$\frac{\partial \widehat{\Phi}}{\partial t} = -\nabla \cdot (\widehat{\Phi} f) + \frac{1}{2} \sigma^2 g^2 \Delta \widehat{\Phi} \qquad \Phi(1, \cdot) \widehat{\Phi}(1, \cdot) = \mu_1$$

with the optimal control $v(t, X_t) = \sigma^2 g(t) \nabla \log \Phi\left(t, X_t\right)$. The solution in (3.30) can be expressed through two coupled SDEs of the form (Léonard, 2013)

$$\mathrm{d}X_t = \left[f + g^2 \nabla \log \Phi\left(t, X_t\right) \right] \mathrm{d}t + g \, \mathrm{d}\mathbb{W}_t, \qquad X_0 \sim \mu_0, \tag{3.31}$$

$$\mathrm{d}X_t = \left[f - g^2 \nabla \log \widehat{\Phi}\left(t, X_t\right) \right] \mathrm{d}t + g \, \mathrm{d}\mathbb{W}_t, \qquad X_T \sim \mu_T, \tag{3.32}$$

where $T = 1$, $\sigma = 1$, and $\nabla \log \Phi\left(t, X_t\right)$ and $\nabla \log \widehat{\Phi}\left(t, X_t\right)$ are the optimal forward and backward drifts for the Schrödinger bridge.

Interestingly, the underlying SDEs (3.29) coincides with the dynamic systems considered in score-based generative models (Song et al., 2021), an emerging generative model class that has achieved remarkable results in synthesizing high-fidelity data (Song and Ermon, 2019; Kong et al., 2021). It also represents a key connection that has recently fueled the development of diffusion Schrödinger bridges (De Bortoli et al., 2021b; Chen et al., 2021b;

4 Hereafter, we will sometimes drop $f \equiv f(t, X_t)$ and $g \equiv g(t)$ for brevity.

Bunne et al., 2023a; Liu et al., 2022a), and will be subject of Chapter 7. 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; Somnath et al., 2023), and cell differentiation processes (Vargas et al., 2021; Bunne et al., 2023a; Tong et al., 2023).

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)

Contributions

Most of the material in this chapter has been already published in the following journal article:

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

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 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 chapter, 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 a static optimal transport formulations (see Section 3.1). As described in Section 2.4, it has found recent successes modeling cellular developmental processes (Lavenant et al., 2021; Schiebinger et al., 2019), albeit in a *non-parameterized* setting. Thus, such OT-based approaches are

unable to make predictions on unseen cells, such as those from unseen samples, e.g., new patients.

In the following, we propose a neural optimal transport-based approach for inferring single-cell perturbation responses. Our method, 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 (3.9) with input 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 lipopolysaccharides (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; Chen et al., 2020).

4.1 NEURAL OPTIMAL TRANSPORT SOLVERS

To generalize optimal transport formulations to the out-of-sample setting, recent efforts have concentrated on developing parameterizations of neural optimal transport schemes. While initial efforts concentrated on solving large-scale OT problems (Seguy et al., 2018), the focus quickly moved to generative adversarial network (GAN) (Arjovsky et al., 2017; Genevay et al., 2018). The majority of existing methods employ OT as a loss function to compute the discrepancy between the model and the data (target) distribution. More recently, however, the focus has shifted to parameterizing the OT map T (3.1) (Yang and Uhler, 2019; Rout et al., 2021; Daniels et al., 2021). Such parameterized OT maps not only function as generative models but extend to tasks that involve interpolations between distributions. More concretely, neural network-based parameterizations of T allow us to model the continuous evolution from a measure μ into a measure ν (Tong et al., 2020), which we will discuss in Part II of this thesis.

Multiple strategies exist on how to parameterize the optimal transport problem. In the following, we consider neural OT solvers that make direct use of the Brenier theorem (Theorem 2) and are based on the semi-dual (3.9).

For a review of alternative approaches, see Section 4.1.2. For convenience, let us restate the semi-dual formulation in (3.9)

$$\varphi^\star \leftarrow \arg \inf_{\varphi \text{ convex}} \int \varphi \mathrm{d}\mu + \int \varphi^* \mathrm{d}\nu \,,$$

where the optimization problem is concerned with finding a convex function φ and its convex conjugate (Legendre transform) φ^* given by

$$\forall y, \quad \varphi^*(y) := \sup_x \langle x, y \rangle - \varphi(y) \,. \tag{4.1}$$

As discussed in Section 3.1.4, the optimal transport map $T^\star$ for cost $c(x, y) = \|x - y\|_2^2$ and $\mathcal{X} = \mathcal{Y} = \mathbb{R}^d$ in (3.1) can be recovered via $\nabla\varphi^\star$.

As the convex conjugate φ^* is very hard to compute, Makkuva et al. (2020) propose to approximate it via another convex function g. Thus, to learn the optimal transport map, the approach builds upon celebrated results by Knott and Smith (1984) and Brenier (1991), which relate the optimal solutions for the primal (3.1) and the dual form (3.5), to derive a min-max formulation that approximates Monge map T (Makkuva et al., 2020, Theorem 3.3). The resulting objective reads

$$\arg \max_{\varphi \text{ convex}} \min_{g \text{ convex}} - \int \varphi(x) \mathrm{d}\mu(x) - \int \langle y, \nabla g(y) \rangle - \varphi(\nabla g(y)) \, \mathrm{d}\nu(y) \,. \tag{4.2}$$

The intuition behind the approach stems from the fact that

$$\int \varphi^* \mathrm{d}\nu = \sup_{g \text{ convex}} \int \langle y, \nabla g(y) \rangle - \varphi(\nabla g(y)) \, \mathrm{d}\nu(y) \,,$$

where we observe that in $\langle y, \nabla g(y) \rangle - \varphi(\nabla g(y)) \leq \varphi^*(y)$ for all functions g the equality is achieved with $g = \varphi^*$ (Makkuva et al., 2020, Theorem 3.3).

In order to solve the optimization problem stated in (4.2), Makkuva et al. (2020) parameterize both potentials φ and g using input convex neural network (ICNN) (Section 4.1.1), i.e., neural networks that parameterize the class of convex functions (Amos et al., 2017), such that

$$\begin{aligned} \theta^\star, \phi^\star \leftarrow \arg \max_{\varphi_\theta \text{ convex}} \min_{g_\phi \text{ convex}} & - \int \varphi_\theta(x) \mathrm{d}\mu(x) - \int \langle y, \nabla g_\phi(y) \rangle \\ & - \varphi_\theta(\nabla g_\phi(y)) \, \mathrm{d}\nu(y), \end{aligned} \tag{4.3}$$

where θ and ϕ are the parameters of each ICNN. Thus, the potentials φ and g can be learned via an alternate min-max optimization problem with loss functions

$$\ell_\varphi(\mu, \nu; \theta) = \mathbb{E}_{x \sim \mu}[\mathrm{ICNN}_\phi(x)] - \mathbb{E}_{y \sim \nu}[\mathrm{ICNN}_\theta(\nabla \mathrm{ICNN}_\phi(y))], \tag{4.4}$$

$$\ell_g(\mu, \nu; \phi) = -\mathbb{E}_{y \sim \nu}[\langle y, \nabla \mathrm{ICNN}_\phi(y) \rangle - \mathrm{ICNN}_\theta(\nabla \mathrm{ICNN}_\phi(y))]. \tag{4.5}$$

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

4.1.1 *Input Convex Neural Networks*

Input convex neural networks are neural networks $\varphi_\theta(x)$ 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 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 $l = 0, \dots, L-1$:

$$z_{l+1} = a_l(W_l^x x + W_l^z z_l + b_l) \text{ and } \varphi_\theta(x) = z_L, \tag{4.6}$$

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.

4.1.2 *Alternative Approaches*

Learning optimal transport problems based on neural networks is at the core of many machine learning applications, including normalizing flows (Rezende and Mohamed, 2015; Huang et al., 2021a) and generative models (Arjovsky et al., 2017; Genevay et al., 2018), albeit the transport map is not explicitly estimated. The formulation introduced in Section 4.1 proposes a neural optimal transport scheme via the semi-dual formulation. In fact, a stream of foundational papers has proposed methods to approximate the dual potential with a neural network. While a common strategies consists in parameterizing φ through an ICNN (Taghvaei and Jalali, 2019; Korotin et al., 2021a; Makkuva et al., 2020), other works explore learning φ using non-convex neural networks (Korotin et al., 2021b; Rout et al., 2021; Nhan Dam et al., 2019). Besides, as the convex conjugate φ^* is hard to compute, several strategies have been developed to solve the conjugation operation in (4.1): Taghvaei and Jalali (2019) propose to exactly computing the conjugate, which, however, is computationally challenging. Besides

approximations of the conjugate as considered in this thesis (Korotin et al., 2021a; Makkuva et al., 2020), more recent approaches suggest (near-)exact conjugate computations through amortized optimization (Amos, 2023).

Beyond such approaches that parameterize the dual potentials of optimal transport, several approaches consider parameterizing map T directly. This is done either without any regularization (Yang and Uhler, 2019), or by introducing regularizers that quantify how far a map T_θ deviates from the ideal properties we expect from a c-OT map. More concretely, Uscidda and Cuturi (2023) introduce the Monge gap regularizer defined as

$$\int c(x, T_\theta(x))\mathrm{d}\mu(x) - W_c(\mu, T_\theta\sharp\mu)$$

that, if 0, guarantees that T_θ is a c-OT map. For $c(x, y) = \|x - y\|_2^2$ and $\mathcal{X} = \mathcal{Y} = \mathbb{R}^d$, for example, T_θ corresponds to a gradient of a convex function (Theorem 2). The advantage of such methods is that they generalize to all cost functions c and are note restricted to the squared Euclidean distance.

4.2 RELATED WORK

With increasing data availability, a diverse set of approaches has been proposed to model cellular perturbation responses, ranging from mechanistic to current deep learning-based approaches. Mechanistic models (Yuan et al., 2021; Fröhlich et al., 2018) define mathematical models of molecular interactions to model the effect of perturbation. These methods, however, are restricted to simpler and well-understood systems as they do not capture highly nonlinear perturbation responses of a heterogeneous cell population. Further, these methods are limited in their applicability as they do not scale to genome-wide measurements (Snijder et al., 2012; Berchtold et al., 2018; Green and Pelkmans, 2016). Linear models (Dixit et al., 2016; Kamimoto et al., 2023), on the other hand, predict changes in cellular gene expression levels using regularized regression methods, where the model predicts a gene's expression level as a linear combination of effects of different perturbations, fitting the regulatory effect of each perturbation on each gene. Due to assuming only linear relationships of individual genes in response to a perturbation, these methods are similarly unable to capture complex and inhomogeneous population responses upon perturbation.

Heydari et al. (2022) predict perturbation responses by inferring the underlying gene regulatory network. Prediction of the perturbed states is achieved through a dynamic simulation of those logical gene networks.

Thus, the predicted perturbed states are restricted to only the selected set of genes used to build the corresponding regulatory network.

Lastly, current state-of-the-art methods (Lopez et al., 2018; Lotfollahi et al., 2019; Yang et al., 2020) aim to learn low-dimensional representations of inputs using autoencoders such that perturbation effects can be applied with simple linear interpolations in representation space. Thus, they predict perturbation responses via linear shifts in a learned low-dimensional latent space. These models are attractive because they are fully parameterized, enabling us to make predictions on unseen cells. By tackling the task of perturbation response predictions via the even more challenging task of learning a meaningful low-dimensional embedding, these methods can be expected to, at best, only perform moderately well. Therefore, we sought to learn a fully parameterized perturbation model that robustly describes the cellular dynamics upon intervention while accounting for underlying variability across samples.

4.3 CELLOT: PREDICTING PERTURBATION RESPONSES VIA NEURAL MONGE MAPS

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 μ consisting of n cells x_i for $i = 1, \ldots, n$, i.e., a dataset of n observations $\{x_1, \ldots, x_n\}, x_i \in \mathcal{X}$ drawn from $\mu \in \mathcal{P}(\mathcal{X})$. Upon perturbation k, the multivariate state of each cell x_i of the unperturbed population changes, which we observe as the perturbed population ν_k (Fig. 4.1a). To understand the mode of action and effect of perturbations, we seek to learn the transition and alignment between populations μ and ν_k via parameterizing a map T_k (see Fig. 4.1a-b), which explains the transition of each cell from the unperturbed cell population μ into their perturbed state ν_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

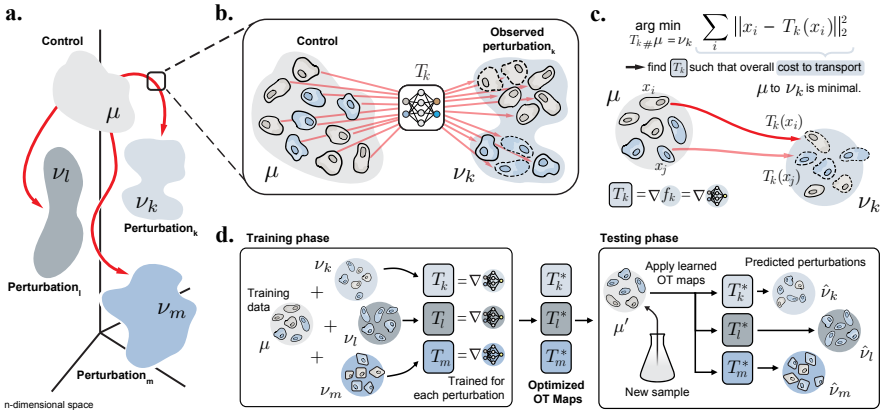


Figure 4.1: Overview of the CellOT Model. a. Distributions of single cells were measured in either an untreated control state (μ) or in one of several perturbed states ($\nu_k, \nu_l, \nu_m, \ldots$). These distributions lie in a high-dimensional space of profiled features. **b.** For a perturbation k, we aim to model it with a function T_k that maps untreated cells in μ to their treated counterparts in ν_k. **c.** Lacking paired measurements, we assume that the perturbation transforms μ into ν_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.

outcome of a perturbation k by applying the learned map T_k to a new unperturbed population μ' (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 μ and ν_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 ν_k and mass is moved from μ into ν_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 approximate semi-dual optimal transport problem φ and g (4.2). Given a set of perturbations K, and sample access to the control distribution μ as well as distributions ν_k for each perturbation $k \in K$, CellOT learns the optimal pair of dual potentials $(\varphi_{\theta_k^*}, g_{\phi_k^*})$ by solving (4.3). For this, each input convex neural networks (Amos et al., 2017) is trained with loss functions (4.4)-(4.4). We recover the optimal map T_k using the gradient of this convex function φ_k, i.e., $\nabla\varphi_k$.

4.4 EMPIRICAL EVALUATION

In the following, we evaluate CellOT on various tasks. For details on the datasets and the chosen evaluation metrics, see Appendix A.2 and Appendix A.3, respectively.

4.4.1 *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 scRNA-seq 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. To further test the hypothesis of the optimal transport modeling prior, we compare the learned OT map ∇f_k for each perturbation k with naive non-OT-based alignments. 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 scRNA-seq 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.

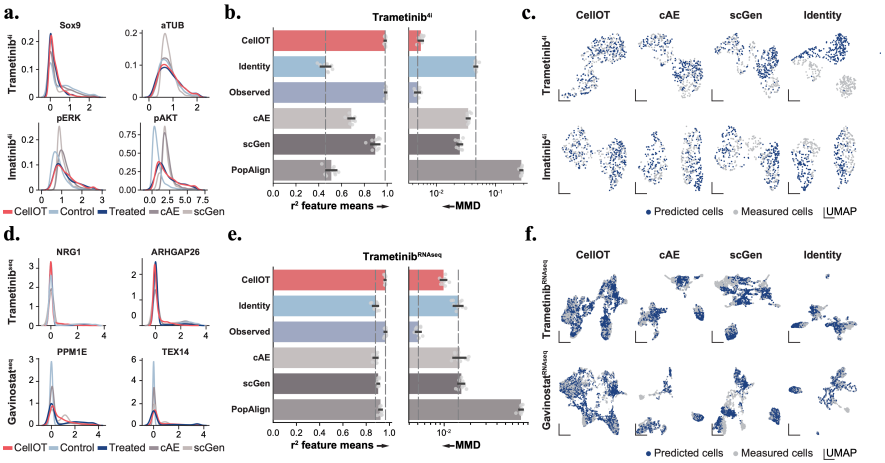


Figure 4.2: CellOT outperforms current state-of-the-art methods on different data modalities. Marginal distribution of marker gene expression (x-axis) of cells profiled by **a.** 4i and **d.** scRNA. Observed control and treated states are shown in light and dark blue. CellOT predictions are shown in 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 the drugs Trametinib, Imatinib, and Gavinostat. 4i data was generated using cell lines M130219 and M130429.

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

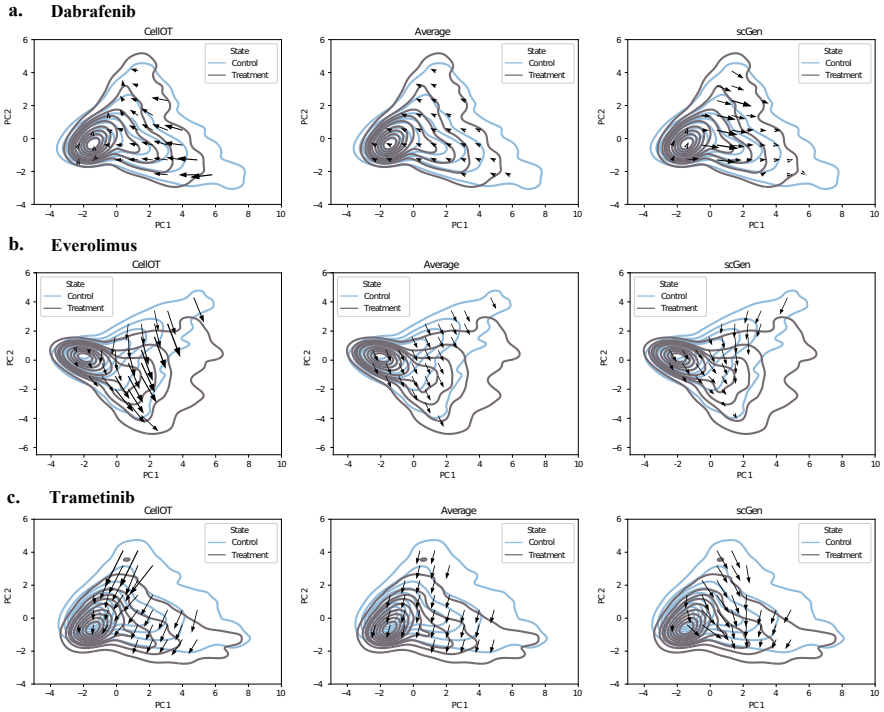


Figure 4.3: Visualization of the learned vector field describing the perturbation response on the single-cell level for **a.** Dabrafenib, **b.** Everolimus, and **c.** Trametinib of the 4i dataset for CellOT, the average effect, and scGen on the first two principal components. Cellular responses are 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.

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 uniform manifold approximation and projection (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.

4.4.2 *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 types or cell state-specific perturbation responses.

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).

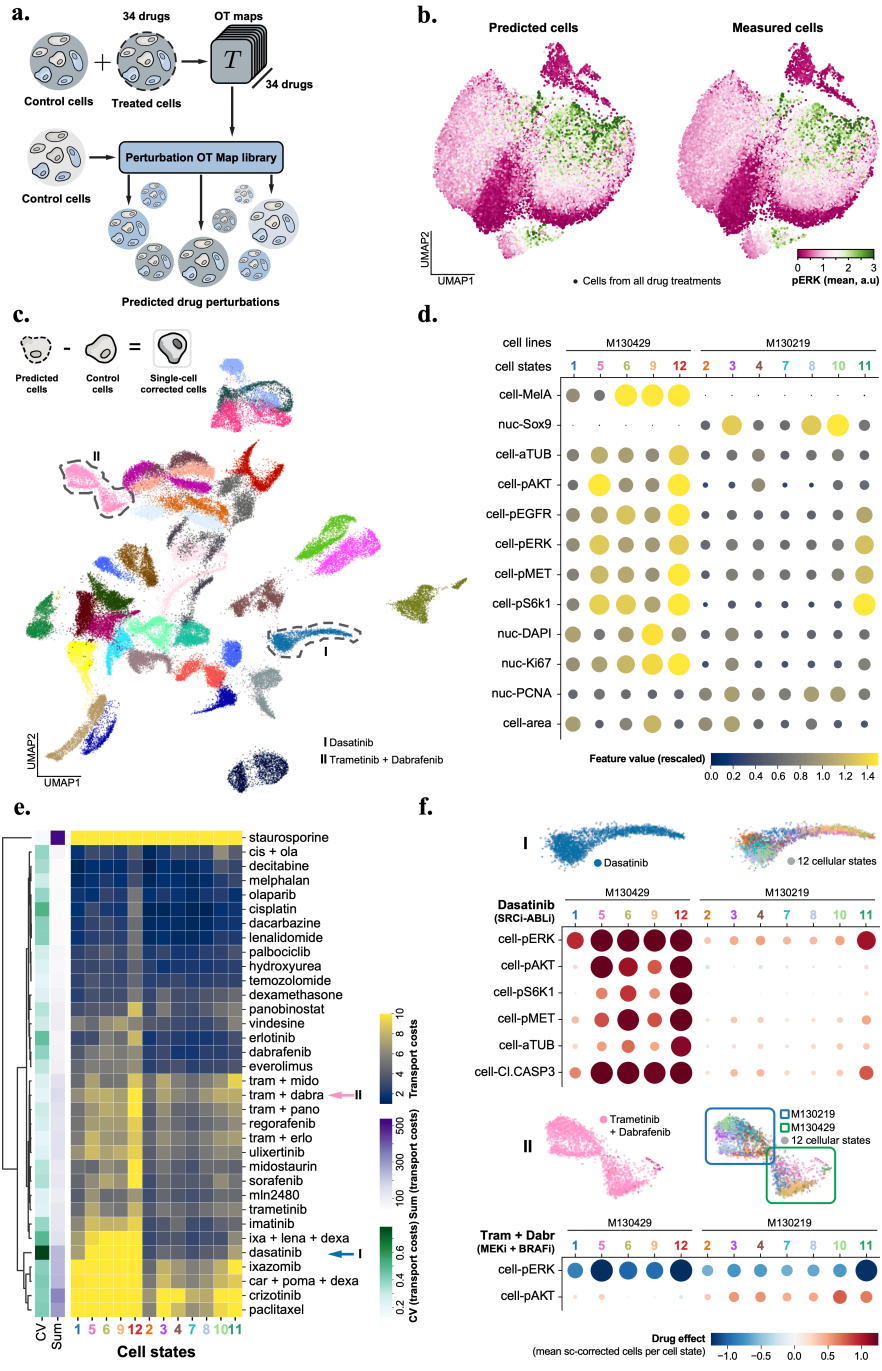

a.
34 drugs
OT maps
Control cells
Treated cells
34 drugs
Control cells
Perturbation OT Map library
Predicted drug perturbations
b.
Predicted cells
Measured cells
UMAP2
UMAP1
Cells from all drug treatments
pERK (mean, a.u)
c.
Predicted cells
Control cells
Single-cell corrected cells
I Dasatinib
II Trametinib + Dabrafenib
d.
cell lines
cell states
M130429
M130219
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
CV
Sum
Cell states
Transport costs
Sum (transport costs)
CV (transport costs)
f.
Dasatinib
12 cellular states
Dasatinib (SRCI-ABLI)
cell-pERK
cell-pAKT
cell-pS6K1
cell-pMET
cell-aTUB
cell-Cl.CASP3
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 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 a 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 are sorted based on their association with the cell lines M130219 and M130429. Vertical axis, cellular features (see Fig. A.1 for the full feature set). The 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, and positive values in red hues of the respective circles.

4.4.3 *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.

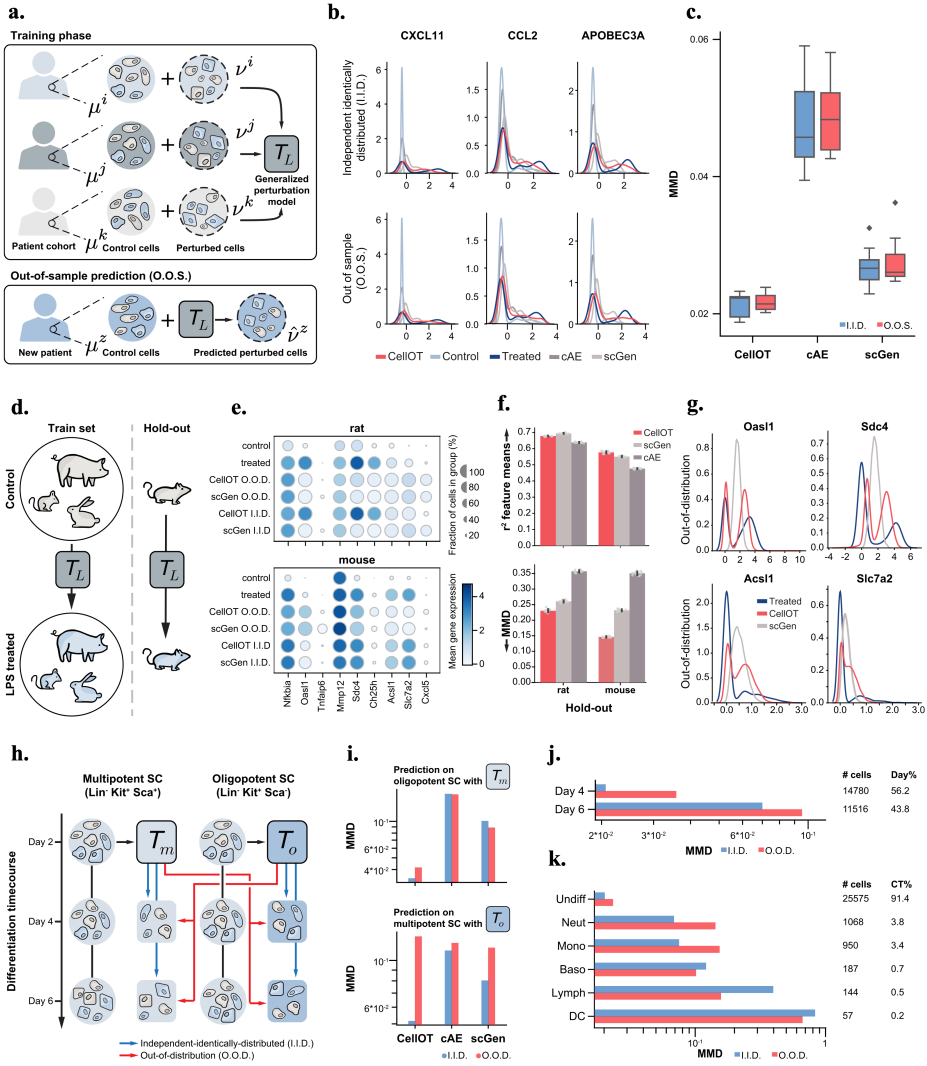

a.
Training phase
Patient cohort
Control cells
Perturbed cells
Generalized perturbation model
Out-of-sample prediction (O.O.S.)
New patient
Control cells
Predicted perturbed cells
b.
CXCL11
CCL2
APOBEC3A
Independent identically distributed (I.I.D.)
Out of sample (O.O.S.)
CellOT
Control
Treated
cAE
scGen
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.
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.
Undiff
Neut
Mono
Baso
Lymph
DC
cells
CT%
25575
91.4
1068
3.8
950
3.4
187
0.7
144
0.5
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 are 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.

4.4.4 *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.

To test the generalization capacity of CellOT in such an out-of-sample (o.o.s.) 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 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 μ_z of the unseen patient as input. This becomes evident when comparing the predicted population $\hat{\nu}_z$ with observations ν_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.2.

4.4.5 *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 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 consists of cells derived from either rat or mouse. See Fig. A.3a,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.3c-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).

4.4.6 *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 (HSPC), 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, and 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).

4.5 DISCUSSION

In this chapter, we proposed CellOT, a framework to model single-cell perturbation responses from unpaired treated and untreated cell states using neural optimal transport by learning a fully parameterized transport map. Previous methods (Rout et al., 2021; Yang and Uhler, 2019; Prasad et al., 2020) rely on an unconstrained parameterization of the 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 (Bunne et al., 2023b, Online Methods).

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 evaluate CellOT by analyzing its performance on unseen cells originating from the same distribution (o.o.s. setting), as well as on different sample compositions (o.o.d. setting). This involves a task on predicting treatment responses in unseen lupus patients, infer developmental trajectories on stem cells of lower potency, and translate innate immune responses across patients (see Fig. 4.5). 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 (i.) similar samples have been observed in the unperturbed setting, and (ii.) 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.2a-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.2f), but fail to capture perturbation effects for patients that exhibit unique responses (see Fig. A.2g). 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. In Chapter 5, we partially address this issue by deriving a neural optimal transport scheme that can be conditioned on a context, e.g., patient metadata, when predicting perturbation responses.

Lastly, concurrent developments in bioengineering aim at overcoming the technological limitation of destructive cell assays. Chen et al. (2022b), for example, 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. While CellOT is unable to incorporate such information into the learning process, in Section 7.3 we propose a novel algorithmic framework that is able to make use of such (partially) aligned datasets (Shi et al., 2023; Tong et al., 2023; Somnath et al., 2023; Liu et al., 2023a).

5

NEURAL OPTIMAL TRANSPORT WITH CONTEXT

Auch die Pause gehört zur Musik.
— Stefan Zweig, *Verwirrung der Gefühle* (1927)

Contributions

Most of the material in this chapter has been already published in the following conference proceedings:

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

A key challenge in the treatment of cancer is to predict the effect of drugs, or a combination thereof, on the 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 theory, and, in that process, develop tools that might be of interest to 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 $\mathcal{T}_\theta(c_{\text{new}})\sharp\mu$.

In this chapter, we introduce 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).

Chapter 4 exploited an explicit connection between OT and neural networks derived from the celebrated Brenier theorem (1987). In this chapter, 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. 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 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 information c_i. These ideas are summarized in an abstract regression model described below.

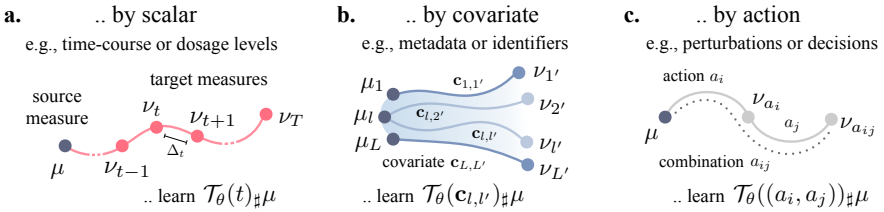


Figure 5.1: The evolution from a source μ to a target measure ν can depend on context variables c of various nature. This comprises **a.** scalars such as time or dosage t which determine the magnitude of the optimal transport, **b.** flow of measures into another one based on additional information (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.

5.1.1 *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 (3.9). 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 $\varphi_\theta(x, c) : \mathbb{R}^d \times \mathcal{C} \to \mathbb{R}$ that is convex w.r.t. x, namely $\mathcal{T}_\theta(c) := x \mapsto \nabla_1 \varphi_\theta(x, c)$; (ii.) estimate θ by maximizing *jointly* the dual

objectives (3.9) simultaneously for all N pairs of measures, in order to ensure that the maps are close to optimal, to form the aggregate problem

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

We detail in Section 4.1 how the Legendre transforms that appear in the energy terms $\mathcal{E}_{\mu_i,\nu_i}$ are handled with an auxiliary function.

5.1.2 *Integrating Context in Convex Architectures*

We propose to incorporate context variables, in order to modulate a family of convex functions $\varphi_\theta(x, c)$ 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 L-layer PICNN is defined as $\varphi_\theta(x, c) = z_L$, where, recursively for $0 \leq l \leq L-1$ one has

$$\begin{aligned} u_{l+1} &= a'_l\left(V_l u_l + v_l\right), \\ z_{l+1} &= a_l\left(W_l^z\left(z_l \circ [W_l^{zu} u_l + b_l^z]_+\right) + W_l^x\left(x \circ (W_l^{xu} u_l + b_l^x)\right) + W_l^u u_l + b_l^u\right), \end{aligned} \tag{5.3}$$

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

$$\theta = \{V_l, W_l^z, W_l^{zu}, W_l^x, W_l^{xu}, W_l^u, v_l, b_l^z, b_l^x, b_l^u\}.$$

Similar to ICNNs, the convexity w.r.t. input variable x is guaranteed as long as activation functions a_i are convex and non-decreasing, and the weight matrices W_l^z have non-negative entries. We parameterize this by storing them as element-wise 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 $L-1$ 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. $y^\star(x) = \arg\min_y \text{PICNN}_\theta(x, y)$. PICNNs 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 $\varphi_\theta(x, c)$ as a function of x, since x is an observation. Instead, our proposal rests on the property that $\nabla_1 \varphi_\theta(x, c)$ 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, PICNNs 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.

5.1.3 *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 denotes a strength or a temporal effect. For instance, McCann's interpolation (3.15) 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 φ_θ, a set of maps parameterized by t, $\mathcal{T}_\theta(t) := x \mapsto \nabla_x \left((1-t)\|x\|^2/2 + t\varphi_\theta(x)\right)$.
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.

5.1.3.1 *Embedding Module $\mathcal{E}$*

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 $\mathcal{E}_\phi$.

ONE-HOT ENCODING $\mathcal{E}_{\text{OHE}}$. When the set $\mathcal{C}$ is small, this can be done effectively using a one-hot encoding (OHE) $\mathcal{E}_{\text{ohe}}$. Thus, covariates, such as subpopulations or patient identifiers, can be simply embedded via one-hot encodings. These embeddings, however, are not able to capture unknown covariates after training.

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 of the context values. For molecular drugs, for example, we can learn molecular 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; Rogers and Hahn, 2010). However, often this domain knowledge is not available.

MODE OF ACTION EMBEDDING $\mathcal{E}_{\text{MOA}}$. In CONDOT, we thus construct so-called mode of action (MOA) embeddings, by computing an embedding $\mathcal{E}_{\text{moa}}$ that encourages actions a with a similar effect on target population ν to have a similar representation. Such mode-of-action embeddings map actions into a latent space based on their mechanism of action and effect on the target population. In the fashion of word embeddings (Mikolov et al., 2013a,b,c), we require actions with similar effects to be closely embedded in the learned representation. This means, however, that we require some sample access of target population particles, i.e., perturbed cells by individual compounds (not in combination). While several metric embeddings are possible (Chopra et al., 2005), we here test a simple multi-dimensional scaling-based embedding (Mead, 1992). For this, we compute the pairwise Wasserstein distance matrix between all target populations of different individual perturbations. We then compute a 10-dimensional MDS embedding based on the stress minimization using a majorization algorithm (`smacof`) (De Leeuw and Mair, 2009) of `sklearn` (Pedregosa et al., 2011), which serves as a descriptor for

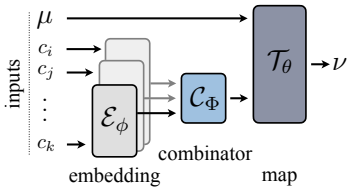


Figure 5.2: CondOT Architecture and Modules. The embedding module $\mathcal{E}_\phi$ embeds arbitrary conditions c, which are then combined via module $\mathcal{C}_\Phi$. Using the processed contexts c, the map $\mathcal{T}_\theta(c)$ acts on μ to predict the target measure ν.

each individual perturbation. In Section 5.2, we analyze several embedding types for different use cases.

5.1.3.2 *Combinator Module $\mathcal{C}$*

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 is 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 $\mathcal{C}_\Phi$. We consider the following combinator modules:

MULTI-HOT COMBINATOR $\mathcal{C}_+^{\text{OHE}}$. A naïve way of constructing the combinator is to combine different actions via multi-hot encodings. If all single perturbations are observed during training, each individual action can be represented via a one-hot encoding. The potential combination of different actions is then encoded by adding the respective one-hot encodings, resulting in a multi-hot encoding for each combination. A limitation of this embedding, however, is that it cannot generalize to unknown actions after training.

DEEP SET COMBINATOR $\mathcal{C}_\Phi^{\text{MOA}}$. When not considering one-hot-based embeddings and when aiming to generalize to unseen perturbations, we need a combinator module that learns how to associate different individual embeddings with each other to receive a joint embedding. As we, for now, do not make an assumption on the order of the perturbation, we consider a permutation-invariance network architecture such as deep sets (Zaheer et al., 2017) with parameters Φ. Taking a set of arbitrary size k containing individual context embeddings $\{\mathcal{E}_{\text{moa}}(c^1), \mathcal{E}_{\text{moa}}(c^2), \ldots, \mathcal{E}_{\text{moa}}(c^k)\}$, it returns a learned combination embedding $\hat{c}_i = \mathcal{C}_\Phi(\mathcal{E}_{\text{moa}}(c^1), \mathcal{E}_{\text{moa}}(c^2), \ldots, \mathcal{E}_{\text{moa}}(c^k))$. Receiving a flexible number of inputs from the embedding module $\mathcal{E}_\phi$, CONDOT allows for joint training of the PICNN parameters θ, embedding parameters ϕ, and combinator parameters Φ in a single, end-to-end differentiable architecture.

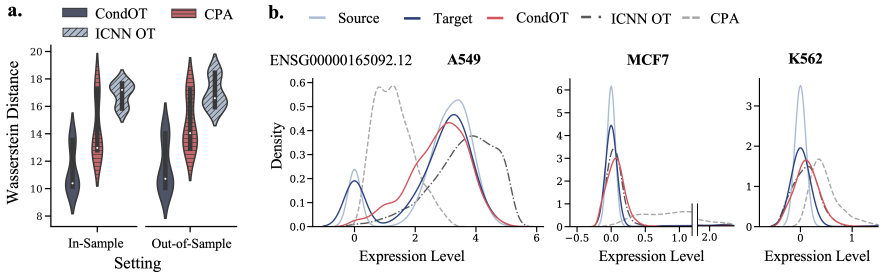


Figure 5.3: a. Predictive performance of CondOT and baselines w.r.t. the entropy-regularized Wasserstein distance on drug dosages *in-sample*, i.e., seen during training, and *out-of-sample*, i.e., unseen during training. **b.** Marginal distributions of observed source and target distributions, as well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

5.1.3.3 *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 follow a similar strategy as introduced in Section 4.1 to learn map $\mathcal{T}_\theta(c_i)$, i.e., by parameterizing potentials φ and g. 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 (3.9), we use the proxy dual objective defined in (4.2) (Makkuva et al., 2020) in place of (3.9) to minimize our overall loss (5.2). We then parameterize the potentials φ and g using two PICNNs, i.e., PICNN_θ and PICNN_ϕ, that already integrate an embedding and/or combinator module. The regularization in (4.2) promotes that for any c, $\text{PICNN}_\phi(\cdot, c)$ resembles the Legendre transform of the other network, i.e., $\text{PICNN}_{\theta^*}(\cdot, c)$. Parameters of all three modules are trained jointly through the alternate min-max optimization introduced in (4.3), replacing the ICNN architecture in loss functions (4.4)-(4.5) with PICNNs, i.e.,

$$\ell_\varphi(\mu, \nu, c; \theta) = \mathbb{E}_{x\sim\mu}[\text{PICNN}_\phi(x, c)] - \mathbb{E}_{y\sim\nu}[\text{PICNN}_\theta(\nabla\text{PICNN}_\phi(y, c), c)],$$
$$\ell_g(\mu, \nu, c; \phi) = -\mathbb{E}_{y\sim\nu}[\langle y, \nabla\text{PICNN}_\phi(y, c)\rangle - \text{PICNN}_\theta(\nabla\text{PICNN}_\phi(y, c), c)].$$

For more details, see Section 4.1.

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.

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

5.2 EMPIRICAL EVALUATION

In the following, we consider various tasks that involve predicting high-dimensional single-cell responses toward a perturbation, here cancer drugs or genetic perturbations. To evaluate the performance of CondOT versus other baselines, we will consider contexts of various nature ranging from the applied drug dosage (Section 5.2.1) and considered cell line (Section 5.2.2) to the selected genetic perturbation, possibly in combination (Section 5.2.3).

5.2.1 *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

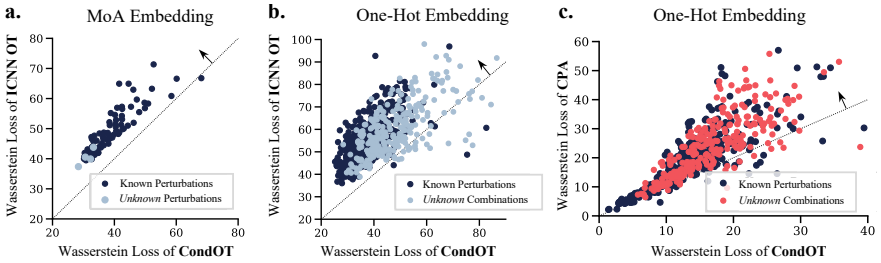


Figure 5.4: Comparison between **a.** CondOT and ICNN OT (Makkuva et al., 2020) based on embedding $\mathcal{E}_{\text{moa}}$ **b.** as well as $\mathcal{E}_{\text{ohe}}$, and **c.** CondOT and CPA (Lotfollahi et al., 2023) based on embedding $\mathcal{E}_{\text{ohe}}$ on *known* and *unknown* perturbations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

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). Details on the evaluation metrics and datasets are provided in Appendix A.3 and Appendix A.2, respectively. 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 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.

5.2.2 *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 encoding 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.

5.2.3 *Inferring Genetic Perturbation 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 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 generalize to *unknown* actions, as well as potential *combinations* of actions. We will discuss all three scenarios below.

5.2.3.1 *Known Actions*

In the following, we analyze CondOT's ability to accurately predict phenotypes of genetic perturbations based on single-cell RNA-sequencing pooled clustered regularly interspaced short palindromic repeats (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 the first step, we do not aim at generalizing beyond perturbations encountered during training, we utilize again a one-hot encoding $\mathcal{E}_{\text{ohe}}$ to

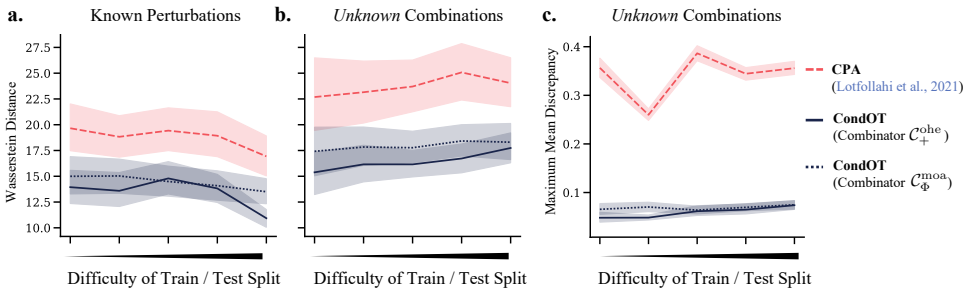


Figure 5.5: Predictive performance for **a.** known perturbations, **b.** unknown perturbations in combination w.r.t. regularized Wasserstein distance and **c.** MMD over different train/test splits of increasing difficulty for baseline CPA as well as CondOT with different combinators $\mathcal{C}_{+}^{\text{ohe}}$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

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 is 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.

5.2.3.2 *Unknown Actions*

With the emergence of new perturbations or drugs, we aim at inferring cellular responses to settings not explored during training. One-hot encodings, 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 have no meaningful embeddings as, for example, molecular fingerprints for drugs, we resort to mode-of-action embeddings introduced in Section 5.1.3. 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

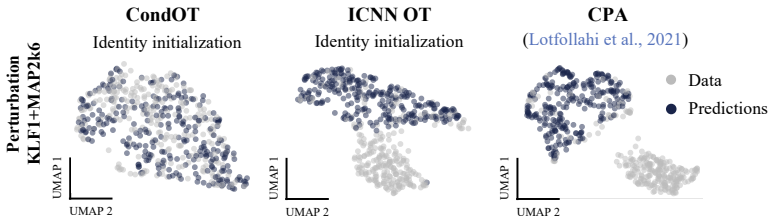


Figure 5.6: UMAP embeddings of cells perturbed by the combination KLF1+MAP2K6 (gray) and predictions of CondOT (ours), ICNN OT (Makkuva et al., 2020), and CPA (blue). While CondOT aligns well with observed perturbed cells, the baselines fail to capture subpopulations.

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.

5.2.3.3 *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 with 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 Section 5.1.3, 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).

5.3 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 neural network 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)

Contributions

Most of the material in this chapter has been already published in the following conference proceedings:

> 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, 2022.

In Part I we introduced *static* neural optimal transport schemes to model how a distribution μ morphs into distribution ν. Motivated by the task of predicting cellular responses to perturbation such as cancer drugs, Chapter 4 and Chapter 5 aimed at parameterizing the OT map T (conditioned on a context c) to enable inferring treatment outcomes to *unseen* cells, e.g., such as from a new patient. Perturbation responses of cells, however, are *dynamic*: After applying perturbation k, cell states evolve over time. Capturing and modeling such processes continuously in time is crucial to understand fine-grained mechanisms of cells. And while Part I assumed that we only have access to the distributions of cell states before μ and after injecting perturbation k, ν_k, many experimental techniques allow us to capture multiple snapshots of an evolving cell population μ_t over time t.
Despite the growing availability of live imaging, most measurement technologies such as scRNA-seq are destructive in nature (see Section 2.1). So while provided with multiple measurements $\{\mu_0, \mu_1, \ldots, \mu_T\}$, we are still dealing with *unaligned* snapshots. Part II will thus concentrate on developing **dynamic neural optimal transport scheme** that allow us to infer

cellular dynamics from a sequence of snapshots and subsequently follow continuous-time trajectories of cells evolving in time. Chapter 6 will concentrate on methods modeling cellular dynamics through connections of optimal transport to partial differential equations and gradient flows, while Chapter 7 takes a stochastic control perspective and elaborates on the link between static entropic OT (3.3) and stochastic differential equations.

6.1 POPULATION DYNAMICS AS GRADIENT FLOWS

Partial differential equations are a fundamental tool in the mathematical description of continuous phenomena, providing a way to capture how systems change over space and time (Risken, 1996). Specifically, the evolution of populations can be modeled by a drift-diffusion PDE, which represents a gradient flow in the space of probability measures. These PDEs encapsulate the rate of change in the population due to both local effects (diffusion) and global effects (drift), reflecting the driving forces in the biological landscape (Teschendorff and Feinberg, 2018; Weinreb et al., 2018). In the single-cell context, this translates into modeling how cell states evolve due to internal genetic factors and external environmental influences. Thus, the connection to PDEs and their links to gradient flow models offer a robust mathematical description for understanding the dynamics of populations at different scales, which we will explore next.

As the exact form of a PDE to model cellular dynamics is usually unknown, we propose in this chapter to model dynamics without necessarily having the PDE solutions in mind. Following in the footsteps of more general applications of the Jordan-Kinderlehrer-Otto scheme (Santambrogio, 2017, §4.8) introduced in Section 3.2.3, we instead interpret the JKO step as a more general parametric type of dynamic for probability measures, exclusively parameterized by the energy J itself. In developmental biology, for example, J might represent an epigenetic landscape: Drawing from the metaphor of Waddington's landscape, developmental biology commonly visualize cellular developmental pathways as marbles rolling down a complex landscape J (Waddington, 1957), e.g., transforming cells from pluripotent states (capable of becoming any cell type) to highly specialized ones (Schiebinger, 2021).

In order to learn such energy J using only snapshots, we propose JKONET, a neural architecture that computes (in end-to-end differentiable fashion) the JKO flow given a parametric energy J_ϕ and initial configuration of points.

RELATED WORK. When the observer only seeks to reconstruct particles' paths given starting and ending point cloud configurations, the machinery of optimal transport (Schiebinger et al., 2019; Yang et al., 2020; Yang and Uhler, 2019) or likelihood-based normalizing flowss (NFs) (Rezende and Mohamed, 2015; Grathwohl et al., 2019) can be used, either separately, or combined: Tong et al. (2020) use OT to motivate a regularizer (squared norm of displacements) in their NF estimation pipeline; Huang et al. (2021a) restrict their attention to flows expressed as gradients of convex functions. This choice is motivated by OT because it agrees with the Brenier (1987) principle that displacements arising from convex potentials give rise to optimal flows. When the observer seeks instead a *causal model*, namely one that is able to explain/predict future configurations of the point cloud (and not only interpolate between configurations), the parameters of that model can also be fitted with OT, as proposed by Hashimoto et al. (2016). Their model assumes a Langevin dynamic for the particles, driven by the gradient flow of a (neural) energy function; They fit the parameters of that network by minimizing regularized OT distances (3.4) (Cuturi, 2013) between their model's predictions and the corresponding ground truth snapshots.

In the following, we draw inspiration from both approaches above—the intuition from the recent normalizing flows 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 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 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 chapter 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

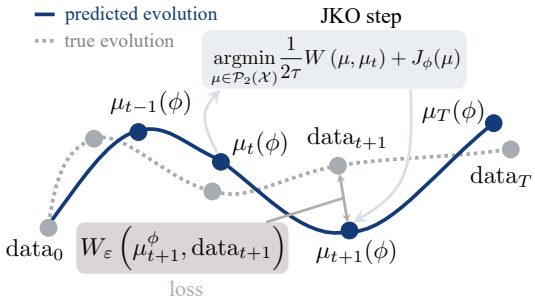


Figure 6.1: Given an observed trajectory $(\text{data}_0, \ldots, \text{data}_T)$ of point clouds (gray), we seek parameters ϕ for the energy J_ϕ such that the predictions $\mu_1, \ldots, \mu_T$ (blue) following a JKO flow from $\mu_0 = \mu_0$ are close the observed trajectory (gray), by minimizing (as a function of ϕ) the sum of Wasserstein distances between μ_{t+1}, the JKO step from μ_{t-1} using J_ϕ, and data μ_{t+1}.

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.

The contributions of this chapter 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 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 it to synthetic potential- and trajectory-based population dynamics, as well as developmental trajectories of human embryonic stem cells based on single-cell genomics data.

6.2 JKONET: A PROXIMAL OPTIMAL TRANSPORT MODEL

Given T discrete measures $\text{data}_0, \dots, \text{data}_T$ 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 ϕ and fit ϕ so that the JKO flow model matches the observed data.
Fitting parameter ϕ with a reconstruction loss requires, using the chain rule, being able to differentiate the JKO step's output w.r.t. ϕ (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 (Section 6.2.1), resulting in a bilevel optimization problem targeting the energy J_ϕ (Section 6.2.2).

6.2.1 *Reformulation of JKO Flows via ICNNs*

Given a starting condition μ_t and energy functional J_ϕ, the JKO step consists in producing a new measure μ_{t+1} implicitly defined as the minimizer of (3.16). Solving directly (3.16) 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 ϕ.
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 μ. Effectively, this is equivalent to making a change of variables in (3.16): Introduce a (variable) convex function φ, and replace the variable μ by the variable $\nabla\varphi_\sharp\mu_t$. Writing

$$\mathcal{E}_J(\mu, \nu) := J(\mu) + \frac{1}{2\tau} W_2^2(\mu, \nu), \tag{6.1}$$

Algorithm 1 JKONET

Input: Dataset $\mathcal{D} = \{\{\text{data}_t^0\}_{t=0}^T, \ldots, \{\text{data}_t^N\}_{t=0}^T\}$ of N population trajectories, ϕ^0 energy parameter initialization, θ^0 ICNN parameter initialization, learning rates γ_θ and γ_ϕ, step τ, regularizer ε, tolerance α, `TeacherForcing` flag

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

$\phi \leftarrow \phi^0$
for $\{\text{data}_t\}_{t=0}^T \in \mathcal{D}$ **do**
  for $t \leftarrow 0$ **to** $T-1$ **do**
    $\theta \leftarrow \theta^0$
    if `TeacherForcing` **or** $t = 0$ **then**
      $\nu \leftarrow \text{data}_t$
    else
      $\nu \leftarrow \mu_t(\phi)$
    while $\frac{\sum_i \|\nabla_{\theta_i} \mathcal{F}_{J_\phi}(\theta)\|_2}{\sum_i \text{count}(\theta_i)} \geq \alpha$ **do**
      $\theta \leftarrow \theta - \gamma_\theta \times \nabla_\theta \mathcal{F}_{J_\phi,\nu}(\theta)$
    $\mu_{t+1}(\phi) \leftarrow \nabla\varphi_{\theta\sharp}\nu$
    $\phi \leftarrow \phi - \gamma_\phi \times \nabla_\phi W_\varepsilon(\mu_{t+1}(\phi), \text{data}_{t+1})$

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

$$\min_\mu \mathcal{E}_J(\mu,\nu) = \min_{\varphi \text{ convex}} \mathcal{F}_J(\varphi,\nu) := \mathcal{E}_J(\nabla\varphi_\sharp\nu,\nu)\,,$$

simplifying the Wasserstein term in (6.1), using the assumption that φ is convex and Brenier's theorem Theorem 2:

$$\mathcal{F}_J(\varphi,\nu) = J(\nabla\varphi_\sharp\nu) + \frac{1}{2\tau}\int \|x - \nabla\varphi(x)\|^2 d\nu(x) \tag{6.2}$$

We pick an ICNN architecture to optimize over a restricted family of convex functions, $\{\varphi_\theta\}$, and define, starting from $\mu_0(\phi) := \text{data}_0$, the recursive sequence for $t \geq 0$,

$$\mu_{t+1}(\phi) := \nabla\varphi_{\theta^\star(\phi,\mu_t(\phi))\,\sharp}\,\mu_t(\phi)\,, \tag{6.3}$$

with $\theta^\star(\phi,\mu_t)$ defined implicitly using ϕ and any ν as

$$\theta^\star(\phi,\nu) := \arg\min_\theta \mathcal{F}_J(\varphi_\theta,\nu) \tag{6.4}$$

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

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

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 $\nabla\varphi(x)$. 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 $J_\phi : \mathcal{P}(\mathbb{R}^d) \to \mathbb{R}$ can be any parameterized function taking a measure as input. Since our model assumes that the observed dynamic is parameterized entirely by that energy (and the initial observation μ_0), the more complex this dynamic, the more complex one would expect the energy J_ϕ to be. We focus in this first attempt on linear functions in the space of measures, that is expectations over μ of a vector-input neural network E_ϕ

$$J_\phi(\mu) := \int E_\phi(x) d\mu(x), \tag{6.5}$$

where $E_\phi : \mathbb{R}^d \to \mathbb{R}$ 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.
To address slow convergence and instabilities for dynamics with many snapshots, we use teacher forcing (Williams and Zipser, 1989) to learn J_ϕ through time. In those settings, during training, J_ϕ uses the ground truth as input instead of predictions from the previous time step. At test time, we do not use teacher forcing.

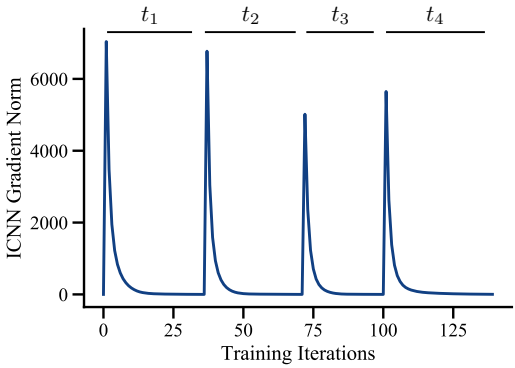


Figure 6.2: Optimization of the ICNN used in JKO steps. The bumps correspond to a change in the outer iteration, and the smooth decrease in between corresponds to a single minimization (6.4) of a time step t_i.

6.2.3 *Bilevel Formulation of* JKONET

Learning the free energy functional J_ϕ while solving each JKO step via an ICNN results in a challenging bilevel optimization problem. At each time step, the predicted dynamics are compared to the ground truth trajectory $(\text{data}_0, \text{data}_1, \ldots, \text{data}_T)$ with the entropy-regularized OT loss (see 3.4),

$$\begin{aligned} &\min_{\phi} \sum_{t=0}^{T-1} W_\varepsilon(\mu_{t+1}(\phi), \text{data}_{t+1}), \\ &\text{s.t. } \mu_0(\phi) := \text{data}_0, \\ &\quad \mu_{t+1}(\phi) := \nabla\varphi_{\theta^\star} \sharp\, \mu_t(\phi)\,, \\ &\quad \theta^\star := \arg\min_{\theta} \mathcal{F}_{J_\phi}(\varphi_\theta, \mu_t(\phi)) \end{aligned} \tag{6.6}$$

The dependence of the Sinkhorn divergence losses in (6.6) on ϕ only appears in the fact that the predictions $\mu_{t+1}(\phi)$ are themselves implicitly defined as solving a JKO step parameterized with the energy J_ϕ.
Learning J_ϕ 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).

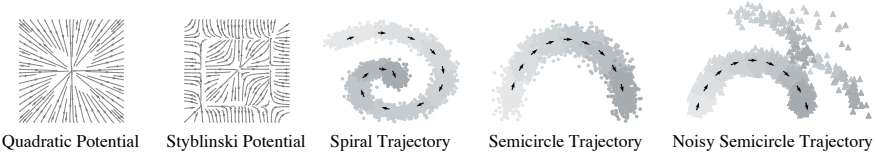


Figure 6.3: Overview on different tasks including trajectory- and potential-based dynamics.

INNER LOOP TERMINATION. A question that arises when defining $\mu_{t+1}(\phi)$ 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 ϕ 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 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., $\sum_i \|\nabla_{\theta_i} \mathcal{F}_{J_\phi}(\theta_i, \phi)\|_2 / \sum_i \text{count}(\theta_i)$. 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 $\partial \mu_{t+1} / \partial \phi$ arising when computing the gradient $\nabla_\phi W_\varepsilon(\mu_{t+1}(\phi), \text{data}_{t+1})$ 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).

SETTING τ IN (6.2). In usual JKO applications, τ needs to be tuned manually. Here, the energy J_ϕ is not fixed, but trained to fit data. Since we put no constraints on the scaling of J_ϕ, τ can be set to 1 without loss of generality, as the parameter ϕ will automatically adjust so that the scale of J_ϕ 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).

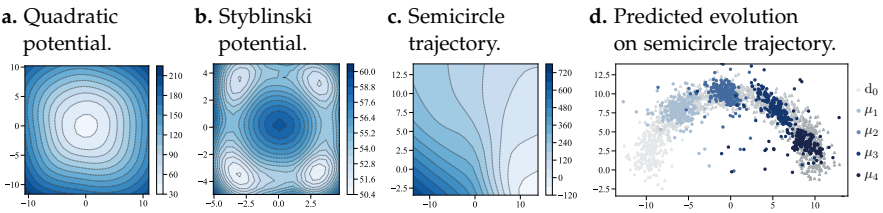


Figure 6.4: Results of JKOnet on potential- and trajectory-based dynamics. (a)-(c) Contour plots of the energy functionals J_ϕ of JKONET on potential- and trajectory-based population dynamics, color gradients depict the magnitude of J_ϕ. (d) Predicted population snapshots $(\mu_1, \ldots, \mu_4)$ (blue) and data trajectory $(d_0, \ldots, d_4)$ (gray).

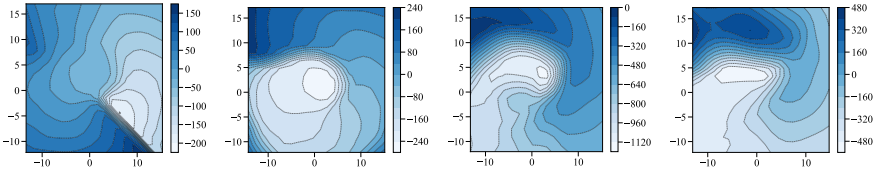


Figure 6.5: Comparison between energy functionals J_ϕ of the spiral trajectory task (see 6.3) between the forward method and JKONET, trained with or without teacher forcing Section 6.2.2). When using teacher forcing, the forward method overfits a gap in the lower-right corner of the spiral, outputting 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.

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

6.3.1 *Synthetic Dynamics*

ENERGY-DRIVEN TRAJECTORIES. The first task involves the evolution 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

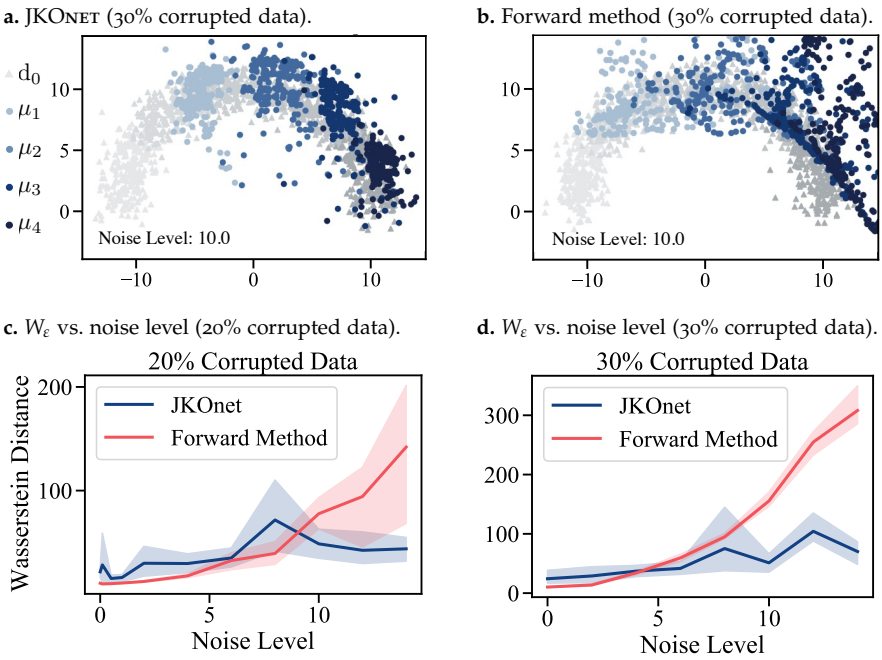


Figure 6.6: Comparison between JKONET and the forward method in settings of increasing noise on corrupted data on the semicircle trajectory task.

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_ϕ and ICNN φ_θ with linear layers ($\varepsilon = 1.0$, $\tau = 1.0$. Figure 6.4a-b demonstrate JKONET's ability to recover convex and nonconvex potentials via energy J_ϕ.

ARBITRARY TRAJECTORIES. As a sanity check, we evaluate if JKONET can recover an energy functional J_ϕ 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 6.5c-d (10 snapshots), JKONET learns energy functionals J_ϕ that can then model the ground truth trajectories. These trajectory-based dynamics are learned using the strong convexity regularizer ($\ell = 0.8$, see Section 6.2.1).

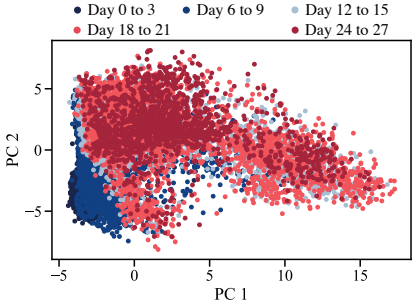
Table 6.1: Evaluation of predictive performance w.r.t. the entropy-regularized Wasserstein distance W_ε (3.4) of JKONET and the forward method on the embryoid body scRNA-seq data per time step (using 3 runs).

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	$\mathbf{0.148 \pm 0.015}$	$\mathbf{0.144 \pm 0.013}$	$\mathbf{0.154 \pm 0.024}$	$\mathbf{0.138 \pm 0.034}$

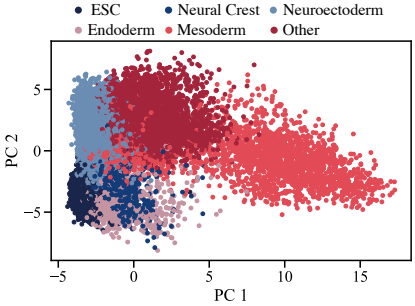
COMPARISON TO FORWARD METHODS. Instead of parameterizing the next iteration $\mu_{t+1}(\phi)$ as we do in the JKONET formulation (3.16), the *forward* scheme states that the prediction at time $t+1$, η_{t+1}, can be obtained as $(\nabla F_\phi)_\sharp \eta_t(\phi)$, where F_ϕ is any arbitrary neural network, as considered in Hashimoto et al. (2016), namely $\eta_0 := \mu_0$ and subsequently $\eta_{t+1}(\phi) := (\nabla F_\phi)_\sharp \eta_t(\phi)$. Although OT still plays an important role in that paper, since the potential F is estimated by minimizing a Sinkhorn loss $W_\varepsilon(\eta_{t+1}, \text{data}_{t+1})$, as we do in (6.6), the forward displacement operator $(\nabla F_\phi)_\sharp$ 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. 6.6c,d). As noise increases, the performance of the forward method deteriorates (Fig. 6.6b), while JKONET, constrained to move points with OT maps, is robust (Fig. 6.6a).

Second, we compare the resulting energy functionals F_ϕ and J_ϕ of the forward method and JKONET, respectively, on the spiral trajectory (see Fig. 6.5). When learning long and complex population dynamics, teacher forcing improves training (see Fig. 6.4c-d). While facilitating the training of

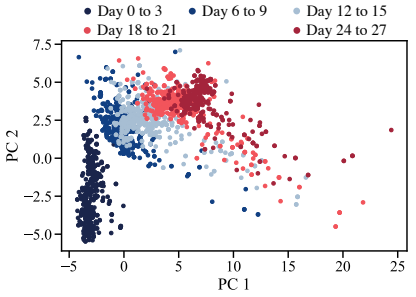
a. PCA embedding of the embryoid body scRNA-seq data colored by the snapshot time.



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.

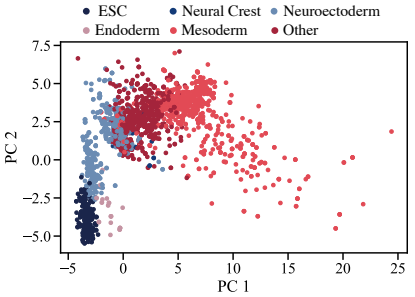
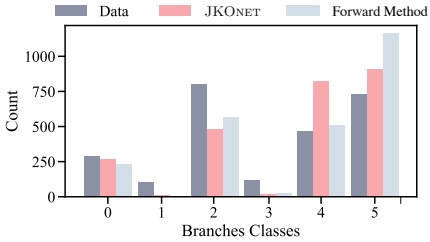


Figure 6.7: Analysis of population dynamics predictions of JKONET on the embryoid body scRNA-seq data.

the forward method in some settings, it likewise results in wrong energy functionals F_ϕ (Fig. 6.5a). JKONET, on the other hand, is able to globally learn the energy functional J_ϕ, despite being only exposed to a one-step history of snapshots during training with teacher forcing (see Fig. 6.5c).

6.3.2 *Single-Cell Dynamics*

We investigate the ability of JKONET to predict the evolution of cellular and molecular processes through time. In the following, we are provided with a single-cell dataset of independent samples of distinct *unaligned* distributions at each snapshot *without* access to ground-truth single-cell trajectories. More concretely, we apply JKONET to embryoid body scRNA-seq data (Moon



Method	Cell Lineage Classification	
	ℓ_1	D_{H}
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 D_{H} (6.7) of the predicted branch class distributions (using 3 runs).

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 4,000 highly differentiable genes. More details on the dataset are provided in Appendix A.2.4. We split the dataset into train and test data ($\sim 15\%$) and parameterize both energy J_ϕ and ICNN φ_θ with linear layers ($\varepsilon = 1.0$, $\tau = 1.0$).

CAPTURING SPATIO-TEMPORAL DYNAMICS. Given the samples from the cell population at day 1 to 3 (i.e., data_0), JKONET learns the underlying spatiotemporal 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 (3.4), 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. For details on the selected evaluation metrics, see Appendix A.3. JKONET outperforms the forward method in terms of W_ε (3.4) distance for both one-step ahead and all-steps ahead predictions for all time steps. The performance of both methods is relatively stable even until days 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 (~ 500 cells per snapshot, Fig. 6.7d).

CAPTURING BIOLOGICAL HETEROGENEITY. Besides measuring the ability of JKONET to model and predict the spatiotemporal 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 body 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

$$D_{\mathrm{H}}(a, b) := \frac{1}{2} \sum_{i=1}^{k} \left(\sqrt{a_i / \|a\|_1} - \sqrt{b_i / \|b\|_1} \right)^2 \tag{6.7}$$

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 that 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 RNA-seq. 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)

Contributions

Most of the material in this chapter has been already published in the following conference proceedings:

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

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.

One of the main challenges in continuously modeling cellular dynamics concerns the innate randomness and fluctuations that exist in biological systems. These random events, such as the timing of a gene being transcribed, or the inherent noise in protein production, make cellular dynamics inherently stochastic. Stochastic processes provide an apt mathematical framework to encapsulate these random phenomena. They enable us to model not just a smoothed behavior, but also the variance and distribution of outcomes, and thus allow us to capture rare but biologically significant events. Therefore, modeling cellular dynamics as stochastic processes is essential to capture the full spectrum of biological behaviors and understand the underlying mechanisms driving these intricate processes.
Consequently, the following chapter deals with the question of learning and identifying a stochastic process $\mathbb{P}_t$ that describes the evolution of a population μ_0 at time point $t = 0$ into a population μ_T at time $t = T$. To draw parallels to Chapter 2, μ_0 and μ_T potentially represent gene expression samples of cells at time 0 and $T = 1$. In this context, recovering the dynamics

from $\mathbb{P}_0 = \mu_0$ to $\mathbb{P}_1 = \mu_1$ might provide us with an understanding of how and why tumor cells evade cancer therapies (Frangieh et al., 2021) or to reconstruct developmental trajectories (Schiebinger et al., 2019).
More concretely, the following chapter concerns Schrödinger bridges, a framework that combines the theory of optimal transport with stochastic optimal control formulations (see Section 3.2.5) and identifies a stochastic process $\mathbb{P}_t$ that represents the evolution of a thermodynamic system at almost equilibrium.
The chapter is divided into two parts: Section 7.2 builds on the realization that estimating such bridges is notoriously difficult. Here, we hypothesize that Gaussian approximations of the data can be used to construct a data-driven reference stochastic process needed to estimate SB. To that end, we solve the SB problem with Gaussian marginals, for which we provide, as a central contribution, a closed-form solution and SDE representation. We use these formulas to define the reference process used to estimate more complex SBs and show that this does indeed help with its numerical solution.
Section 7.3 similarly proposes a new training objective for learning and parameterizing SBs by utilizing the structure of *aligned* data, which naturally arises in many biological phenomena. While most single-cell measurement technologies are destructive (see Section 2.1), recent developments in molecular biology 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) clonal tracing evolving cells, allowing us to connect them to their progenitors. Here we propose a novel algorithmic framework that, for the first time, learns SBs while respecting the data alignment hinging on a combination of two decades-old ideas: The classical Schrödinger bridge theory and Doob's *h-transform*.

7.1 DIFFUSION SCHRÖDINGER BRIDGES

As described above, in order to learn the evolution of a population from μ_0 to μ_1, we will invoke the framework of SBs. Given two marginals μ_0 and μ_1, we select a reference process $\mathbb{Q}_t$ based on prior knowledge, for instance, simple Brownian motion. As discussed in Section 3.2.5, it turns out that the solution to the general SB problem (3.24) is itself given by two coupled SDEs (3.31)-(3.32) (Léonard, 2013), which we restate here for convenience

$$\mathrm{d}X_t = (f + gZ_t)\,\mathrm{d}t + g\,\mathrm{d}W_t, \quad X_0 \sim \mu_0, \tag{7.1}$$

$$\mathrm{d}X_t = (f - g\hat{Z}_t)\,\mathrm{d}t + g\,\mathrm{d}W_t, \quad X_1 \sim \mu_1\,, \tag{7.2}$$

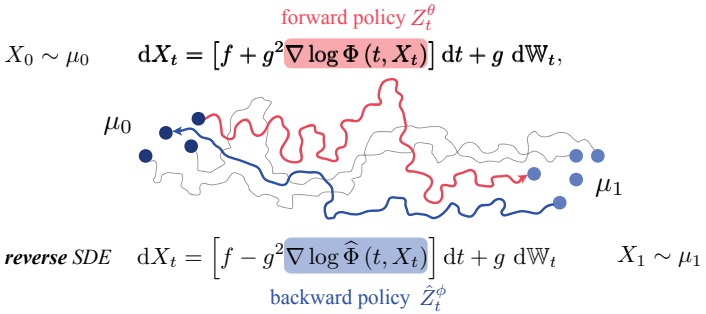


Figure 7.1: Parameterization of diffusion Schrödinger bridges. The forward SDE with forward policy Z_t steers particles $X_0 \sim \mu_0$ from $t = 0$ to μ_1 at $t = 1$. The reverse SDE runs backward in time. Here, backward policy $\hat{Z}_t$ determines the evolution of particles $X_1 \sim \mu_1$ at $t = 1$ to μ_0 at $t = 0$.

where we replaced $\nabla \log \Phi\,(t, X_t)$ in (3.31) and $\nabla \log \widehat{\Phi}\,(t, X_t)$ in (3.32) with $Z_t, \hat{Z}_t \colon \mathbb{R}^d \to \mathbb{R}^d$, i.e., two time-indexed smooth *vector fields* called the optimal forward and backward drift, respectively. Note that (7.2) runs backward in time, i.e., from $1 \to 0$ (Anderson, 1982). Choosing f and g depending on the considered SDE class, the forward and backward policies $Z_t, \hat{Z}_t$ are generally *unknown*. Similar as in score-based generative models (SGMs) (Song et al., 2021; Hyvärinen and Dayan, 2005) which parameterize the score function, in order to *estimate* the resulting SB from data, we learn the forward and backward drift through neural networks with parameters θ, ϕ, i.e., $Z_t^\theta(x)$ and $\hat{Z}_t^\phi(x)$. For a visualization of the resulting parameterization, see Fig. 7.1.

Several estimators and training procedures for the so-called diffusion Schrödinger bridge (DSB), i.e., for learning Z_t and $\hat{Z}_t$, have been proposed based on either Gaussian processes (Vargas et al., 2021), dual potentials (Finlay et al., 2020), or neural networks (De Bortoli et al., 2021b; Chen et al., 2022a). In this thesis, we consider the likelihood training framework by Chen et al. (2022a) grounded on forward-backward SDE (FBSDE) theory (Ma and Yong, 1999; Exarchos and Theodorou, 2018). Crucially, these FBSDEs can be used to construct the likelihood objectives for SBs that, surprisingly, generalize the ones for SGMs as special cases.

The negative likelihood functions for θ and ϕ are then given by

$$\ell(x_0;\phi) = \int_0^1 \mathbb{E}_{(7.1)}\left[\frac{1}{2}\|\hat{Z}_t^\phi\|^2 + g\nabla_x\cdot\hat{Z}_t^\phi + \langle Z_t^\theta, \hat{Z}_t^\phi\rangle\,\mathrm{d}t\middle| X_0 = x_0\right], \quad (7.3a)$$

$$\ell(x_1;\theta) = \int_0^1 \mathbb{E}_{(7.2)}\left[\frac{1}{2}\|Z_t^\theta\|^2 + g\nabla_x\cdot Z_t^\theta + \langle \hat{Z}_t^\phi, Z_t^\theta\rangle\,\mathrm{d}t\middle| X_1 = x_1\right]. \quad (7.3b)$$

and serve as loss functions for likelihood-based training of DSBs. Here, $\nabla_x\cdot$ denotes the divergence operator w.r.t. the x variable: For any $v\colon \mathbb{R}^d \to \mathbb{R}^d$, $\nabla_x\cdot v(x) := \sum_{i=1}^d \frac{\partial}{\partial x_i} v_i(x)$.

Unfortunately, such frameworks necessitate a forward-backward learning process known as the iterative proportional fitting (IPF) procedure (Fortet, 1940; Kullback, 1968). As both policies $Z_t, \hat{Z}_t$ are initially unknown and randomly parameterized, training DSBs often results in numerical and scalability issues.

Further, none of these approaches is capable of incorporating *alignment* of the data. This can be seen by inspecting the objective (3.24), in which the coupling information (x_0^i, x_1^i) is completely lost as only its individual marginals μ_0, μ_1 play a role therein.

In this chapter, we tackle both of these limitations: First, in Section 7.2 we derive a data-driven reference process that provides a more robust initialization of DSBs. Second, in Section 7.3, we devise the first algorithmic framework that solves (7.1)-(7.2) in settings where sparse trajectories, or partially aligned data, are available *without* resorting to IPF.

7.2 DATA-DRIVEN PRIORS FOR DIFFUSION SCHRÖDINGER BRIDGES

The Schrödinger bridge (Léonard, 2013; Chen et al., 2021b), alternatively known as the *dynamic* entropy-regularized optimal transport, 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., 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., 2022a).

Despite these impressive achievements, a common limitation of the existing works is that the SBs are typically solved in a purely numerical fashion.

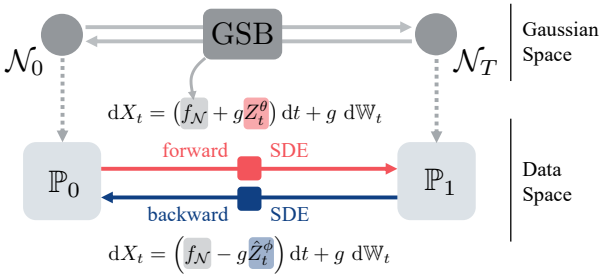


Figure 7.2: Solving the SB problem between μ_0 and μ_1 is notoriously difficult because it requires learning the time-dependent drifts of two SDEs that respect the desired marginals, and a random initialization for these drifts is usually extremely far from satisfying that constraint. We propose a data-dependent procedure that relies first on Gaussian approximations of the data measures, which provide a closed-form drift $f_{\mathcal{N}}$ in (7.32) (the GSB). We show that this facilitates the training of forward/backward drifts $\hat{Z}_t^\theta, \hat{Z}_t^\phi$.

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).
The goal of this section is to continue this pursuit of closed-form solutions and extend these advantages to SB-based learning methods. For an overview of the method, see Fig. 7.2. 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 expres-

sions 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 overall better performance.

7.2.1 *Preliminaries on Gaussian Optimal Transport*

Throughout this chapter, 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):

$$\min_{\pi \in \Pi(\xi, \xi')} \int \|x - x'\|^2 \mathrm{d}\pi(x, x') + 2\sigma^2 D_{\mathrm{KL}}(\pi \| \xi \otimes \xi'), \tag{7.4}$$

where $\xi \otimes \xi'$ denotes the product measure of ξ and ξ', and $\sigma \geq 0$ is a regularization parameter. When $\sigma = 0$, (7.4) 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

$$D_\sigma := (4\Sigma^{\frac{1}{2}}\Sigma'\Sigma^{\frac{1}{2}} + \sigma^4 I)^{\frac{1}{2}}, \quad C_\sigma := \frac{1}{2}(\Sigma^{\frac{1}{2}} D_\sigma \Sigma^{-\frac{1}{2}} - \sigma^2 I), \tag{7.5}$$

then the solution $\pi^\star$ to (7.4) is itself a Gaussian:

$$\pi^\star \sim \mathcal{N}\left(\begin{bmatrix} \mu \\ \mu' \end{bmatrix}, \begin{bmatrix} \Sigma & C_\sigma \\ C_\sigma^T & \Sigma' \end{bmatrix} \right). \tag{7.6}$$

7.2.1.2 *Dynamic Gaussian Optimal Transport*

In the literature, (7.4) 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:

$$\min_{\rho_0=\xi,\ \rho_1=\xi'} \mathbb{E}\left[\int_0^1 \frac{1}{2}\|v_t\|^2 + \frac{\sigma^4}{8}\|\nabla \log \rho_t\|^2 \,\mathrm{d}t\right]. \tag{7.7}$$

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:

$$\partial_t \rho_t = -\nabla_x \cdot (\rho_t v_t), \tag{7.8}$$

where $(\nabla_x \cdot v_t)(x) := \sum_{i=1}^d \frac{\partial}{\partial x_i} v^i(x)$ denotes the divergence operator with respect to the x variable. It can be shown that if $\rho_t^\star$ is the optimal curve for (7.7), then the joint distribution of the end marginals $(\rho_0^\star, \rho_1^\star)$ coincides with (7.6), 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 (7.7) 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 of this part of the thesis, i.e., 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 overall stochastic processes $\mathbb{P}_t$ (Léonard, 2013; Chen et al., 2021b):

$$\min_{\mathbb{P}_0=\nu,\ \mathbb{P}_1=\nu'} D_{\mathrm{KL}}(\mathbb{P}_t \| \mathbb{Q}_t). \tag{7.9}$$

In practice, ν and ν' typically arise as the (empirical) *marginal* distributions of 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 (7.9) is thus interpreted as the best dynamics that conforms to the prior belief $\mathbb{Q}_t$ while respecting the data marginals ($\mathbb{P}_0^\star = \nu, \mathbb{P}_1^\star = \nu'$).
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):

$$\mathrm{d}Y_t = (c_t Y_t + \alpha_t)\,\mathrm{d}t + g_t\,\mathrm{d}\mathbb{W}_t := f_t\,\mathrm{d}t + g_t\,\mathrm{d}\mathbb{W}_t. \tag{7.10}$$

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

$$\min_{\mathbb{P}_0=\nu,\,\mathbb{P}_1=\nu'} D_{\mathrm{KL}}(\mathbb{P}_t \| Y_t). \tag{7.11}$$

Then (7.11) is equivalent to

$$\inf_{(\rho_t, v_t)} \mathbb{E}\left[\int_0^1 \frac{\|v_t\|^2}{2g_t^2} + \frac{g_t^2}{8}\|\nabla \log \rho_t\|^2 - \frac{1}{2}\langle f_t, \nabla \log \rho_t\rangle\,\mathrm{d}t\right] \tag{7.12}$$

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

$$\partial_t \rho_t = -\nabla_x \cdot (\rho_t(f_t + v_t)). \tag{7.13}$$

The proof of Theorem 3, which we defer to Appendix A.4, 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.[5]

7.2.2.2 *The Gaussian Schrödinger Bridge Problem*

The central goal of this section is to derive the closed-form solution of SBs when the marginal constraints are Gaussians $\xi \sim \mathcal{N}(\mu, \Sigma)$, $\xi' \sim \mathcal{N}(\mu', \Sigma')$.

5 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.

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 AND RELATED WORK. In order to analyze (GSB), we first notice that the objective in (7.12) 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 $\sigma\mathbb{W}_t$-GSBs. Up to a constant factor, this is simply (7.7), so Theorem 3 reduces to the well-known fact that $\sigma\mathbb{W}_t$-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 (7.12). 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 Karush-Kuhn-Tucker (KKT) or fixed-point arguments. This is the case for static Gaussian OT (7.4), 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 (7.7) can be recovered from the static one via a simple linear interpolation (McCann, 1997).
- **Ad hoc computations:** When $\sigma > 0$ in (7.7), 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 (Mallasto et al., 2021), which manages to find a closed form for (7.7) (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 of 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 3, GSBs are more general than (7.7) 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 (7.12) present a serious obstruction for generalizing the approach of Mallasto et al. (2021) to Y_t-GSBs when $f_t \neq 0$ or g_t is not constant; this is exemplified by the convoluted expressions in our Theorem 5, 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., (7.7). 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 (7.17)) on covariance matrices, thus essentially reducing the solutions of GSBs to solving a matrix equation. This route is enabled via yet another equivalent formulation of (7.7), 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 5). 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*

Following the connections established in Section 3.2.4, let us consider the following *action minimization* problem with fixed endpoints $x, x' \in \mathbb{R}^d$:

$$\min_{x(0)=x, x(1)=x'} \int_0^1 \frac{1}{2}\|\dot{x}(t)\|^2 - U(x(t))\,\mathrm{d}t, \tag{7.14}$$

where the minimum is taken over all piecewise smooth curves. A celebrated result in physics asserts that the optimal curve for (7.14) satisfies the *Euler-Lagrange* equation:

$$\ddot{x}(t) = -\nabla U(x(t)), \quad x(0) = x, \quad x(1) = x'. \tag{7.15}$$

In particular, when $U \equiv 0$, (7.15) reduces to $\ddot{x} \equiv 0$, i.e., $x(t)$ is a straight line connecting x and x'.
More generally, one can consider (7.14) on any *Riemannian manifold*, provided that the Euclidean norm $\|\cdot\|_2$ in (7.14) is replaced by the corresponding Riemannian norm. In this case, the Euler-Lagrange equation (7.15) 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 (7.14) 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 4 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:

$$\mathcal{T}_\Sigma \mathbb{S}^d_{++} := \{U \in \mathbb{R}^{d\times d} : U^T = U\}. \tag{7.16}$$

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

$$A: \quad \Sigma A + A\Sigma = U. \tag{7.17}$$

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

$$\langle U, V \rangle_\Sigma := \frac{1}{2} \operatorname{tr} \mathcal{L}_\Sigma[U] V \tag{7.18}$$

satisfies all the axioms of the Riemannian metric; additional background on the Bures-Wasserstein geometry can be found in Appendix A.5.1.

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):

$$\min_{\mathbb{P}_0 = \mathcal{N}(0,\Sigma),\, \mathbb{P}_1 = \mathcal{N}(0,\Sigma')} D_{\mathrm{KL}}(\mathbb{P}_t \| \sigma \mathbb{W}_t). \tag{7.19}$$

Then we have:

Theorem 4. *The minimizer of* (7.19) *(and hence* (7.7)*) coincides with the solution of the action minimization problem:*

$$\min_{\Sigma_0 = \Sigma, \Sigma_1 = \Sigma'} \int_0^1 \frac{1}{2} \|\dot{\Sigma}_t\|^2_{\Sigma_t} - \mathcal{U}_\sigma(\Sigma_t)\, \mathrm{d}t \tag{7.20}$$

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* (7.19) *solves the Euler-Lagrange equation in the Bures-Wasserstein geometry:*

$$\nabla_{\dot{\Sigma}_t} \dot{\Sigma}_t = -\operatorname{grad} \mathcal{U}_\sigma(\Sigma_t), \quad \Sigma_0 = \Sigma, \quad \Sigma_1 = \Sigma', \tag{7.21}$$

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 (7.7) or (7.19) is not new; it is derived in Mallasto et al. (2021) via a strenuous and rather unenlightening calculation:

$$\Sigma_t := \bar{t}^2 \Sigma + t^2 \Sigma' + t \cdot \bar{t} \Big(C_\sigma + C_\sigma^T + \sigma^2 I \Big). \tag{7.22}$$

Here, $\bar{t} := 1 - t$ and C_σ is defined in (7.5). However, the interpretation of (7.22) as the minimizer of (7.20) is new and suggests a principled avenue towards the closed-form solution of $\sigma\mathbb{W}_t$-GSBs: solve the Euler-Lagrange equation (7.21). Inspecting the formulas for $\nabla_{\dot{\Sigma}_t}\dot{\Sigma}_t$ and $\operatorname{grad}\mathcal{U}_\sigma(\Sigma_t)$ (see (A.5.5) and (A.5.6)), one can further reduce (7.21) to computing the Lyapunov operator $\mathcal{L}_{\Sigma_t}[\dot{\Sigma}_t]$, which presents the bottleneck in the proof of Theorem 4 as there is, in general, no closed form for the matrix equation (7.17). To this end, our main contribution is the following technical Lemma:

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

$$\tilde{S}_t := t\Sigma' + \bar{t}C_\sigma - \bar{t}\Sigma - tC_\sigma^T + \frac{\sigma^2}{2}(\bar{t} - t)I. \tag{7.23}$$

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

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

$$\tilde{S}_t^T\Sigma_t^{-1}\cdot\Sigma_t^{-1} + \Sigma_t^{-1}\cdot\Sigma_t^{-1}\tilde{S}_t = \tilde{S}_t^T + \tilde{S}_t = \dot{\Sigma}_t \tag{7.24}$$

which is more or less equivalent to the original Euler-Lagrange equation (7.21); we defer the details to Appendix A.5.2.
To conclude, in contrast to the purely technical approach of Mallasto et al. (2021), our Theorem 4 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 (7.23) is itself *not* symmetric. Other consequences of Theorem 4 that might be of independent interest can be found in Appendix A.5.3. We also note that when $\sigma = 0$, the solution to (7.20) 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 4 and of (7.22) 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 (7.10) is (Platen and Bruti-Liberati, 2010):

$$Y_t = \tau_t \left(Y_0 + \int_0^t \tau_s^{-1} \alpha_s \, \mathrm{d}s + \int_0^t \tau_s^{-1} g_s \, \mathrm{d}\mathbb{W}_s \right). \tag{7.25}$$

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 $\mathbb{W}_t$ and Itô's isometry (Protter, 2005), we compute:

$$\mathbb{E}[Y_t \mid Y_0] = \tau_t \left(Y_0 + \int_0^t \tau_s^{-1} \alpha_s \, \mathrm{d}s \right) =: \eta(t) \tag{7.26}$$

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

$$\begin{aligned} &\mathbb{E}\left[(Y_t - \eta(t)) \left(Y_{t'} - \eta\left(t'\right)\right)^T \,\middle|\, Y_0 \right] \\ &\qquad = \left(\tau_t \tau_{t'} \int_0^t \tau_s^{-2} g_s^2 \, \mathrm{d}s \right) I =: \kappa\left(t, t'\right) I. \end{aligned} \tag{7.27}$$

7.2.4.2 *Main Result*

We now present the main result of this section. With the important application of diffusion-based models in mind, we will not only derive solution curves as in (7.22) 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 (7.5).

Theorem 5. *Denote by $\mathbb{P}_t$ the solution to Gaussian Schrödinger bridges (GSB). Set*

$$
\begin{aligned}
&r_t := \frac{\kappa(t,1)}{\kappa(1,1)}, \quad \bar{r}_t := \tau_t - r_t\tau_1, \quad \sigma_\star := \sqrt{\tau_1^{-1}\kappa(1,1)},\\
&\zeta(t) := \tau_t \int_0^t \tau_s^{-1}\alpha_s \,\mathrm{d}s, \rho_t := \frac{\int_0^t \tau_s^{-2} g_s^2 \,\mathrm{d}s}{\int_0^1 \tau_s^{-2} g_s^2 \,\mathrm{d}s},\\
&P_t := \dot{r}_t(r_t\Sigma_1 + \bar{r}_t C_{\sigma_\star}), \quad Q_t := -\dot{\bar{r}}_t(\bar{r}_t\Sigma_0 + r_t C_{\sigma_\star}),\\
&S_t := P_t - Q_t^T + \Big[c_t\kappa(t,t)(1-\rho_t) - g_t^2\rho_t\Big] I. \qquad (7.28)
\end{aligned}
$$

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*

$$\mu_t := \bar{r}_t\mu_0 + r_t\mu_1 + \zeta(t) - r_t\zeta(1), \qquad (7.29)$$

$$\Sigma_t := \bar{r}_t^2\Sigma_0 + r_t^2\Sigma_1 + r_t\bar{r}_t\Big(C_{\sigma_\star} + C_{\sigma_\star}^T\Big) + \kappa(t,t)(1-\rho_t)I. \qquad (7.30)$$

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

$$\mathrm{d}X_t = f_{\mathcal{N}}(t, X_t)\,\mathrm{d}t + g_t\,\mathrm{d}\mathbb{W}_t \qquad (7.31)$$

where

$$f_{\mathcal{N}}(t,x) := S_t^T\Sigma_t^{-1}(x-\mu_t) + \dot{\mu}_t. \qquad (7.32)$$

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

As in Theorem 4, the key step in the proof of Theorem 5 is to recognize the symmetry of the matrix $S_t^T\Sigma_t^{-1}$ where S_t, defined in (7.28), simply becomes the $\tilde{S}_t$ in Lemma 1 (up to an additive factor of $\frac{\sigma^2\bar{I}}{2}I$) 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 5 will follow a slightly different route. In any case, given the symmetry of $S_t^T\Sigma_t^{-1}$, the proof simply boils down to a series of straightforward calculations; see Appendix A.6.

CLOSED FORMS FOR CONDITIONAL DISTRIBUTIONS. In many practical applications such as generative modeling, a requirement to employ the SDE representation of GSBs in (7.31) is that its *conditional distributions* given the initial points can be computed efficiently. As an immediate corollary of Theorem 5, we obtain the following closed-form expressions for these conditional distributions.

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 (7.28).

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	$c_t \equiv 0$ $g_t = \sqrt{\dot{\mathrm{q}}(t)}$	$\mathrm{q}(t)$	$\mathrm{q}(1)$	$\frac{\mathrm{q}(t)}{\mathrm{q}(1)}$	$1-\frac{\mathrm{q}(t)}{\mathrm{q}(1)}$	$\frac{\mathrm{q}(t)}{\mathrm{q}(1)}$	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	$\int_0^t \alpha_s\,\mathrm{d}s$

Corollary 2. *Let $X_t \sim \mathbb{P}_t$ be the solution to (GSB). Then the conditional distribution of X_t given endpoints has a simple solution: $X_t|X_0 = x_0 \sim \mathcal{N}\left(\mu_{t|0}, \Sigma_{t|0}\right)$, where*

$$\mu_{t|0} = \bar{r}_t x_0 + r_t\left(\mu_1 + C_{\sigma_*}^T \Sigma_0^{-1}(x_0 - \mu_0)\right) + \zeta(t) - r_t\zeta(1), \tag{7.33}$$

$$\Sigma_{t|0} = r_t^2\left(\Sigma_1 - C_\sigma^T \Sigma_0^{-1} C_\sigma\right) + \kappa(t,t)(1-\rho_t)I. \tag{7.34}$$

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

$$\mu_{t|1} = r_t x_1 + \bar{r}_t\left(\mu_0 + C_{\sigma_*} \Sigma_1^{-1}(x_1 - \mu_1)\right) + \zeta(t) - r_t\zeta(1), \tag{7.35}$$

$$\Sigma_{t|1} = \bar{r}_t^2\left(\Sigma_0 - C_\sigma \Sigma_1^{-1} C_\sigma^T\right) + \kappa(t,t)(1-\rho_t)I. \tag{7.36}$$

Examples of GSBs. Our framework captures the 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 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 (7.30) is known for BM (Mallasto et al., 2021), what is crucial in these applications is the SDE presentation (7.31), 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: *Recovering Stochastic Dynamics via Gaussian Schrödinger Bridges*

Building on the closed-form solutions in Section 7.2.4, we present an end-to-end learning paradigm that takes two marginal distributions μ_0, μ_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 (7.32) and (7.33)-(7.36). 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 $f_{\mathcal{N}}\,\mathrm{d}t + g_t\,\mathrm{d}\mathbb{W}_t$ in (7.32), we propose to minimize the objective

$$\min_{\mathbb{P}_0=\mu_0,\,\mathbb{P}_1=\mu_1} D_{\mathrm{KL}}(\mathbb{P}_t \| \mathbb{Q}_t). \tag{7.37}$$

Following the framework of Chen et al. (2022a), we see that the optimal solution to (7.37) is given by two SDEs of the form:

$$\mathrm{d}X_t = (f_{\mathcal{N}} + g_t Z_t)\,\mathrm{d}t + g_t\,\mathrm{d}\mathbb{W}_t, \quad X_0 \sim \mu_0, \tag{7.38a}$$
$$\mathrm{d}X_t = \big(f_{\mathcal{N}} - g_t \hat{Z}_t\big)\,\mathrm{d}t + g_t\,\mathrm{d}\mathbb{W}_t, \quad X_1 \sim \mu_1, \tag{7.38b}$$

where (7.38b) 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 becomes

$$\ell(x_0;\phi) = \int_0^1 \mathbb{E}_{(7.38a)}\Big[\frac{1}{2}\|\hat{Z}_t^\phi\|^2 + g\nabla_x \cdot \hat{Z}_t^\phi + \langle Z_t^\theta, \hat{Z}_t^\phi\rangle\,\mathrm{d}t \Big| X_0 = x_0\Big], \tag{7.39a}$$
$$\ell(x_1;\theta) = \int_0^1 \mathbb{E}_{(7.38b)}\Big[\frac{1}{2}\|Z_t^\theta\|^2 + g\nabla_x \cdot Z_t^\theta + \langle \hat{Z}_t^\phi, Z_t^\theta\rangle\,\mathrm{d}t \Big| X_1 = x_1\Big], \tag{7.39b}$$

i.e., the scheme introduced in Section 7.1 but based on an underlying GSB. 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 (7.39a)-(7.39b) reduces to simulating (7.32) *conditioned* on the given start or end data points. Thanks to our

Algorithm 2 Forward and Backward Pretraining

Input: Marginal distributions μ_0, μ_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 (7.33)-(7.34) with $x_0 \sim \mu_0$
  Compute $\ell(x_0; \phi)$ via (7.39a)
  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 (7.35)-(7.36) with $x_1 \sim \mu_1$
  Compute $\ell(x_1; \theta)$ via (7.39b)
  Update $\theta_0 \leftarrow \theta_0 - \gamma_\theta \nabla \ell(x_1; \theta_0)$

closed-form expressions, this can be easily achieved by drawing Gaussian variables with mean and covariance prescribed in (7.38a)-(7.38b). The pretraining procedure is summarized in Algorithm 2.

Step 3: Alternating minimization. After the pretraining phase, we switch to minimizing (7.39a)-(7.39b) with general drifts in (7.38a)-(7.38b). 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. 7.2 for an illustration.

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 where μ_0 is a simple Gaussian. Indeed, while performing competitively in the generative setting ($\mathcal{N}_0 \to \mu_1$), our method *outperforms* when modeling the evolution of two complex distributions ($\mu_0 \to \mu_1$), i.e., the most general and ambitious setting to estimate a bridge. More concretely, our goal is two-fold:

1. To solve the **generative modeling** problem, i.e., to generate μ_0 or μ_1 from a standard Gaussian noise, and

Algorithm 3 GSBFLOW

Input: Marginal distributions μ_0, μ_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 (7.37)

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 (7.38a) with $x_0 \sim \mu_0$
    Compute $\ell(x_0; \phi)$ via (7.39a)
    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 (7.38b) with $x_1 \sim \mu_1$
    Compute $\ell(x_1; \theta)$ via (7.39b)
    Update $\theta \leftarrow \theta - \gamma_\theta \nabla \ell(x_1; \theta)$

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 $\mathbb{P}_0 = \mu_0, \mathbb{P}_1 = \mu_1$ (interpolation).

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.
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

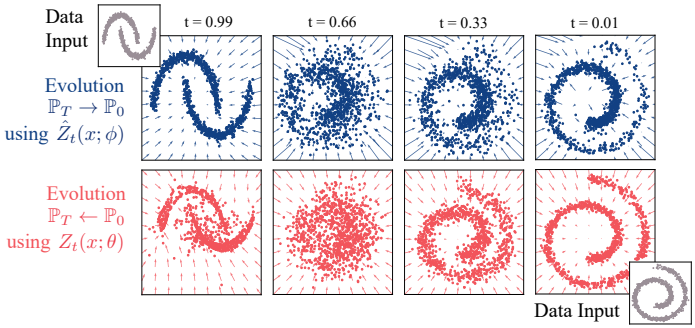


Figure 7.3: Illustration of the time-dependent drifts learned by GSBFLOW with VE SDE for two toy marginal distributions. *Top.* Evolution of μ_1 (moons) $\to$ μ_0 (spiral) via backward policy $\hat{Z}_t^{\phi}(x)$. *Bottom.* Evolution of μ_0 (spiral) $\to$ μ_1 (moons) via forward policy $Z_t^{\theta}(x)$.

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).

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	$\mathbf{27.85 \pm 0.68}$

half circles (μ_1) into a spiral (μ_0). Fig. 7.3 shows the trajectories learned by GSBFLOW based on the VESDE (see Table 7.1).

While it is sufficient to parameterize only a single policy ($\hat{Z}_t^{\phi}(x)$) in generative modeling, the task of learning to evolve μ_0 into μ_1 requires one to recover *both* vector fields $\hat{Z}_t^{\phi}(x)$ and $Z_t^{\theta}(x)$. As demonstrated in Fig. 7.3, 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.

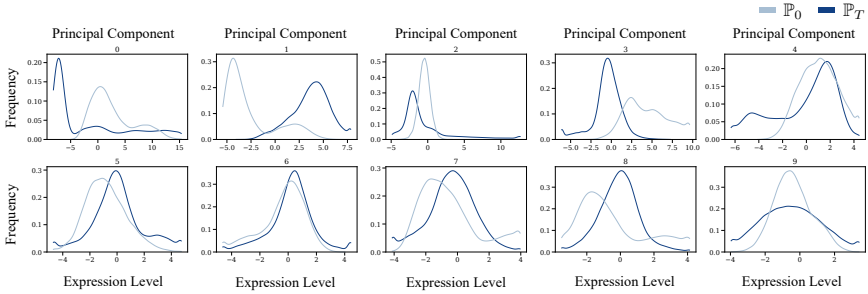


Figure 7.4: The expression levels of the first 10 principal components from the dataset by Schiebinger et al. (2019).

7.2.6.2 *Single-Cell Dynamics*

Let us consider the evolution of single-cell populations, for which we can collect the empirical distributions μ_0, μ_1 of its expression levels at the times $t = 0, 1$ (Schiebinger et al., 2019; Moon et al., 2019).

Our choice of Y_t; the GSBFLOW.

Instead of directly diving into the numerical solution of SBs, we first empirically verify that the distributions μ_0, μ_1 in single-cell genomics are typically close to *non-standard* Gaussian distributions (see Fig. 7.4 for the canonical dataset by Schiebinger et al. (2019)).
Since the solutions of SBs are Lipschitz in terms of μ_0, μ_1 (Carlier et al., 2022), a reasonable approximation to the original SB objective is to replace μ_0, μ_1 by Gaussians with matching moments. This results in a GSB problem which can be solved in closed form by our Theorem 5. 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 μ_0 and μ_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 $\mu_0 \to \mu_1$ and $\mu_1 \to \mu_0$, respectively. The full algorithm is presented in Algorithm 3 for completeness.

Single-cell genomics via GSBs.

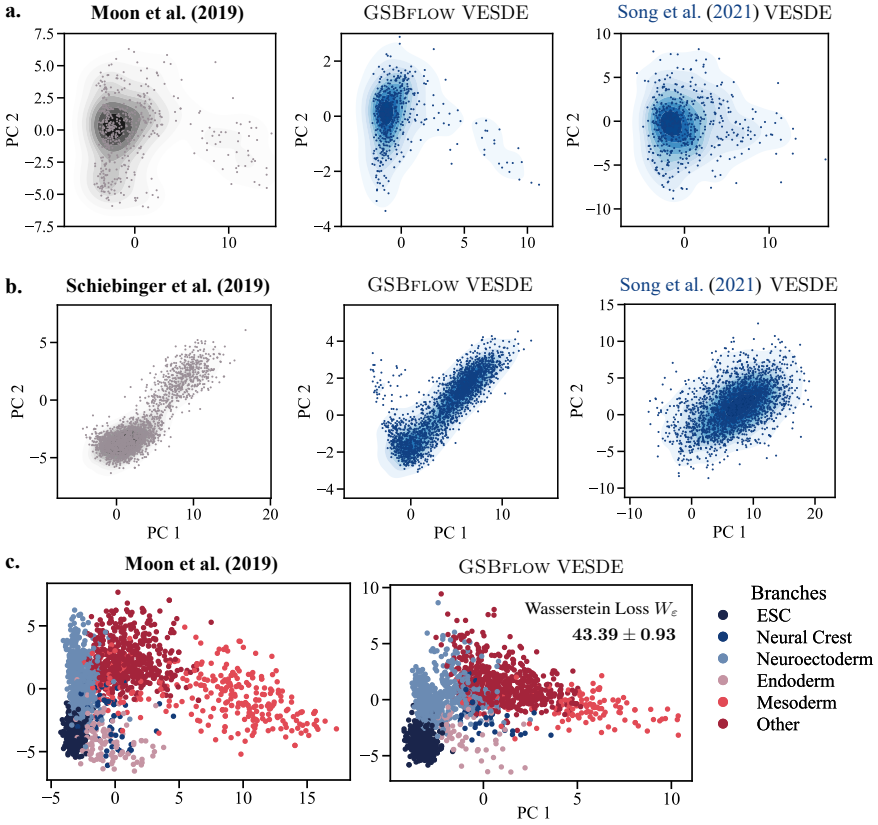


Figure 7.5: a.-b. Visual evaluation of the ability of our method to model the **generation** of data from **a.** Moon et al. (2019) and **b.** Schiebinger et al. (2019). Density plots are visualized in 2D PCA space and show generated data points using either GSBflow (our method) or the procedure in Song 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.

GENERATIVE MODELING SETTING. We investigate the ability of GSBflow to generate cell populations μ_1 from noise $\mathcal{N}_0$ ($\mathcal{N}_0 \to \mu_1$, Fig. 7.5a, b) on the canonical datasets (Moon et al., 2019; Schiebinger et al., 2019); as well as to predict the dynamics of single-cell genomics ($\mu_0 \to \mu_1$, Fig. 7.5c) (Moon et al., 2019), i.e., the inference of cell populations μ_1 resulting from

the developmental process of an initial cell population μ_0, with the goal of learning individual 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. 7.5a, b. GSBFLOW performs competitively on reconstructing embryoid body differentiation landscapes (Moon et al., 2019), and outperforms score-based generative models baselines on the induced pluripotent stem cell (iPSC) reprogramming task (Schiebinger et al., 2019) as quantified by W_ε between data and predictions. For more details on datasets and evaluation metrics, see Appendix A.2 and Appendix A.3, respectively.

INTERPOLATION SETTING. 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 μ_0, while μ_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. 7.5c, upper panel) is captured. Fig. 7.5c (lower panel) closely resembles the data (see W_ε in Fig. 7.5c) 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.

LIMITATIONS 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.

7.3 DIFFUSION SCHRÖDINGER BRIDGES FROM SPARSE TRAJECTORIES

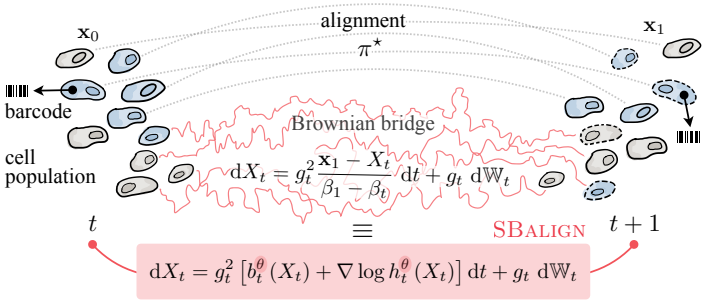


Figure 7.6: Overview of SBalign. In the setting of cell differentiation we aim at learning the evolutionary process that morphs a population from its stat at t to $t+1$. Through genetic tagging (i.e., barcodes) we are able to trace progenitor cells at time point t into their descendants $t+1$. This provides us with an alignment between populations at consecutive time steps. Our goal is then to recover a stochastic trajectory from x_0 to x_1. To achieve this, we connect the characterization of an SDE conditioned on x_0 and x_1 (utilizing the Doob's *h-transform*) with that of a Brownian bridge between x_0 and x_1 (classical Schrödinger bridge theory), leading to a simpler training procedure with lower variance and strong empirical results.

As introduced in the previous chapter, diffusion Schrödinger bridges (De Bortoli et al., 2021b; Chen et al., 2022a; Vargas et al., 2021; Liu et al., 2022a) 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 μ_0, μ_1 are two (empirical) distributions between which we wish to interpolate, then a tacit assumption in the literature is that the dependence of μ_0 and μ_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 μ_0 and μ_1 naturally come in pairs $(x_0^i, x_1^i)_i^N$, which is common in many biological phenomena. 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 Fig. 7.6).
Motivated by these observations, the goal of this section 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 based on barcoded measurements of cell differentiation in hematopoiesis. 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., 2022a). 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. (2023b) to solve the problem of constrained diffusion. However, their fundamental motivation is different from ours. Liu et al. (2023b) 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 $(x^i_{0,\text{test}})_{i=1}^m$ 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 $(x^i_{1,\text{test}})_{i=1}^m$ is correct, whereas ours is capable of predicting that $x^i_{1,\text{test}}$ is precisely the destination of $x^i_{0,\text{test}}$ for each i.

PROBLEM FORMULATION. Suppose that we are given access to i.i.d. *aligned* data $(x_0^i, x_1^i)_{i=1}^N$, where the marginal distribution of x_0^i's is μ_0 and of x_1^i's is μ_1. Typically, we view μ_0 as the empirical marginal distribution of a stochastic process observed at time $t = 0$, and likewise μ_1 the empirical marginal observed at $t = 1$. The goal is to reconstruct the stochastic process $\mathbb{P}_t$ based on $(x_0^i, x_1^i)_{i=1}^N$, i.e., to *interpolate* between μ_0 and μ_1.

Such a task is ubiquitous in biological applications. For instance, in molecular dynamics simulations, we have access to trajectories $\left(x_t^i\right)_{t\in[0,1]}$, where x_0^i and 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 (7.9) 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):

$$\pi^\star := \underset{\mathbb{P}_0=\mu_0,\, \mathbb{P}_1=\mu_1}{\arg\min} \; D_{\mathrm{KL}}(\mathbb{P}_{0,1} \| \mathbb{Q}_{0,1}), \tag{7.40}$$

where the minimization is over all *couplings* of μ_0 and μ_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 (7.9), then the joint distribution $\mathbb{P}_{0,1}$ necessarily coincides with the $\pi^\star$ in (7.40). Moreover, since in DSBs the data is always assumed to arise from $\mathbb{P}_t$, we see that:

> The *aligned* data $(x_0^i, x_1^i)_{i=1}^N$ 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 endpoints, 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:

$$\mathrm{d}\mathbb{Q}_t = g_t \, \mathrm{d}\mathbb{W}_s. \tag{7.41}$$

In this case, it is well-known that $\mathbb{Q}_t$ *conditioned* to start at x_0 and end at x_1 can be written in another SDE (Mansuy and Yor, 2008; Liu et al., 2023b):

$$\mathrm{d}X_t = g_t^2 \frac{x_1 - X_t}{\beta_1 - \beta_t} \, \mathrm{d}t + g_t \, \mathrm{d}\mathbb{W}_s \tag{7.42}$$

where $X_0 = x_0$ and

$$\beta_t := \int_0^t g_s^2 \, \mathrm{d}s. \tag{7.43}$$

We call the processes in (7.42) 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 of the following classical result in Schrödinger bridge theory: Draw a sample $(x_0, x_1) \sim \pi^\star$ and connect them via (7.42). 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)

$$\mathrm{d}X_t = g_t^2 b_t(X_t)\,\mathrm{d}t + g_t\,\mathrm{d}W_s. \tag{7.44}$$

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 (7.44) *conditioned* to start at x_0 and end at x_1 is given by another SDE (Doob, 1984; Rogers and Williams, 2000):

$$\mathrm{d}X_t = g_t^2[b_t(X_t) + \nabla \log h_t(X_t)]\,\mathrm{d}t + g_t\,\mathrm{d}W_s \tag{7.45}$$

where $h_t(x) := \mathbb{P}(X_1 = x_1 | X_t = x)$ is the *Doob's h function*. Notice that we have suppressed the dependence of h_t on x_0 and x_1 for notational simplicity.

LOSS FUNCTION. Since both (7.42) and (7.45) 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 (7.42) and (7.45), we can learn the optimal parameter $\theta^\star$ via optimization of the loss function

$$\ell(\theta) := \mathbb{E}\left[\int_0^1 \left\| \frac{x_1 - X_t}{\beta_1 - \beta_t} - \nabla \log h_t^\theta(X_t) \right\|^2 \mathrm{d}t\right] \tag{7.46}$$

where h_t^θ is determined by b_t^θ as well as the drawn samples (x_0, x_1). In short, assuming that, for each θ, we can compute h_t^θ *based only on* b_t^θ, we can then backpropagate through (7.46) and optimize it using any off-the-shelf algorithm.

A SLIGHTLY MODIFIED (7.46). Even with infinite data and a neural network with sufficient capacity, the loss function defined in (7.46) does converge to 0. For the purpose of numerical stability, we instead propose to modify (7.46) to:

$$\ell(\theta) := \mathbb{E}\left[\int_0^1 \left\| \frac{x_1 - X_t}{\beta_1 - \beta_t} - \left(b_t^\theta + \nabla \log h_t^\theta(X_t)\right) \right\|^2 \mathrm{d}t\right] \tag{7.47}$$

Algorithm 4 SBALIGN

Input: Aligned data $(x_0^i, x_1^i)_{i=1}^N$, 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 $h_{t,\eta}$

Initialize $\theta \leftarrow \theta_0$, $\phi \leftarrow \phi_0$
for $k = 1$ **to** K **do**
  Draw a mini-batch of samples from $(x_0^i, x_1^i)_{i=1}^N$
  Compute empirical average of loss ℓ (7.51) with mini-batch
  Update $\phi \leftarrow \phi - \gamma_\phi \nabla \ell(\theta, \phi)$
  Update $\theta \leftarrow \theta - \gamma_\theta \nabla \ell(\theta, \phi)$

which is clearly equivalent to (7.46) at the true solution of b_t. Notice that (7.47) bears a similar form as the popular score-matching objective employed in previous works (Song and Ermon, 2019; Song et al., 2021):

$$\ell(\theta) := \mathbb{E}\left[\int_0^1 \left\| \nabla \log p(x_t|x_0) - s^\theta(X_t, t) \right\|^2 \mathrm{d}t\right], \tag{7.48}$$

where the term $\frac{x_1 - X_t}{\beta_1 - \beta_t}$ is akin to $\nabla \log p(x_t|x_0)$, 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 (7.45), we see that, given (x_0, x_1), it can be written as the conditional expectation of an indicator function:

$$h_t(x) = \mathbb{P}(X_1 = x_1 | X_t = x) = \mathbb{E}\left[\mathbb{1}_{\{x_1\}} | X_t = x\right] \tag{7.49}$$

where the expectation is over (7.44). Functions of the form (7.49) lend themselves 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 $\mathbb{1}_{\{x_1\}}$ by its *smoothed* version (Zhang and Chen, 2022; Holdijk et al., 2022):

$$h_{t,\eta}(x) := \mathbb{E}\left[\exp\left(-\frac{1}{2\eta}\|X_1 - x_1\|^2\right) | X_t = x\right]. \tag{7.50}$$

Here, η is a regularization parameter that controls how much we "soften" the constraint, and we have $\lim_{\eta \to 0} h_{t,\eta} = h_t$.
Although the computation of (7.50) can be done via a standard application of the Feynman-Kac formula (Rogers and Williams, 2000), an altogether

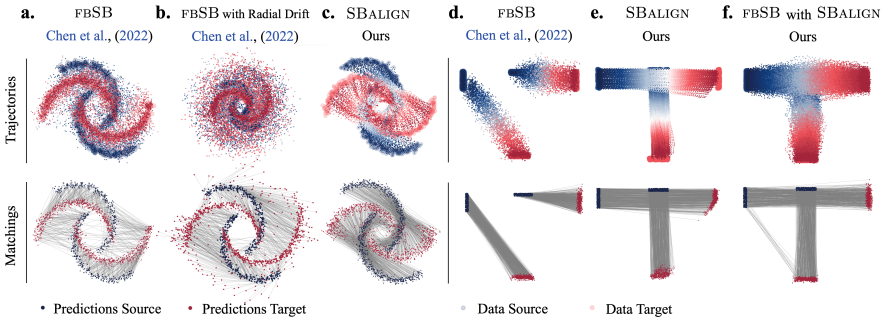


Figure 7.7: Experimental results on the Moon dataset (**a-c**) and T-dataset (**d-f**). The top row shows the trajectory sampled using the learned drift, and the bottom row shows the matching based on the learned drift. Compared to other baselines, SBALIGN is able to learn an appropriate drift respecting the true alignment. (**f**) further showcases the utility of SBALIGN's learned drift as a suitable reference process to improve other training methods.

easier approach is to parametrize $h_{t,\eta}$ 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 (7.44), 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., 2023b), 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 (x_0, x_1) is an input pair, then multiple runs of the SDE (7.44) starting from x_0 should, on average, produce samples that are in the proximity of 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:

$$\ell(\theta,\phi) := \mathbb{E}\left[\int_0^1 \left\| \frac{x_1 - X_t}{\beta_1 - \beta_t} - \left(b_t^\theta + m^\phi(X_t)\right)\right\|^2 + \lambda_t \|m^\phi(x_t)\|^2 \,\mathrm{d}t\right] \tag{7.51}$$

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*

Our algorithm finds solutions to SBs on aligned data by relying on samples drawn from the (optimal) coupling $\pi^\star$. This is what differentiates it from classical SBs –which instead only consider samples from $\hat{\mathbb{P}}_0$ and $\hat{\mathbb{P}}_1$– and plays a critical role in avoiding IPF-like iterates. However, SBALIGN reliance on samples from $\pi^\star$ may become a limitation, when the available information on alignments is insufficient.

If the number of pairings is limited, it is unrealistic to hope for an accurate solution to the aligned SB problem. However, the interpolation between $\hat{\mathbb{P}}_0$ and $\hat{\mathbb{P}}_1$ learned by SBALIGN can potentially be leveraged as a starting point to obtain a better reference process, which can then be used when solving a classical SB on the same marginals. In other words, the drift $b_t^{\text{aligned}}(X_t)$ learned through SBALIGN can be used *as is* to construct a data-informed alternative $\tilde{\mathbb{Q}}$ to the standard Brownian motion, defined by paths:

$$\tilde{X}_t = b_t^{\text{aligned}}(\tilde{X}_t)dt + g_t dW_t$$

Intuitively, solving a standard SB problem with $\tilde{\mathbb{Q}}$ as reference is beneficial because the (imperfect) coupling of marginals learned by SBALIGN ($\tilde{\mathbb{Q}}_{01}$) is, in general, closer to the truth than $\mathbb{Q}_{01}$.

Improving reference processes through pre-training or data-dependent initialization has 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-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 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 $\pi^\star_{01}$.

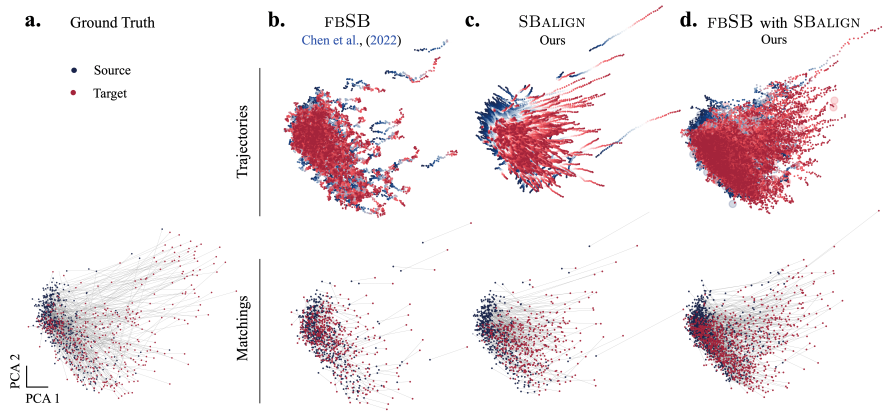


Figure 7.8: Cell differentiation trajectories based on (**a**) the ground truth and (**b-d**) learned drifts. SBALIGN is able to learn an appropriate drift underlying the true differentiation process while respecting the alignment. (**d**) Using the learned drift from SBALIGN as a reference process helps improve the drift learned by other training methods.

7.3.3 *Empirical Evaluation*

In this section, we evaluate SBALIGN in different settings involving 2-dimensional synthetic datasets, and the task of reconstructing cellular differentiation processes.

7.3.3.1 *Synthetic Dynamics*

We run our algorithm on two synthetic datasets and compare the results with classic diffusion 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. 7.7a-c) consists of two distributions, each supported on two semi-circles (μ_0 drawn in *blue* and μ_1 in *red*). μ_1 was obtained from μ_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. 7.7a), which tend to map points to their

closest neighbor in μ_1 (e.g., some points in the upper arm of μ_0 are brought towards the left rather than towards the right). While being a minimizer of (7.9), 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. 7.7c).
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. 7.7b), 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 (7.9). This is already visible in simple examples like (Fig. 7.7d-f), in which the value of good candidate prior drifts at a specific location needs to vary wildly in time. In this dataset, μ_0 and μ_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. 7.7e) but unsurprisingly fail when using the baseline FBSB (Fig. 7.7d) 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. 7.7e), 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 initially 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 Section 7.3.2, in the 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. 7.7f).

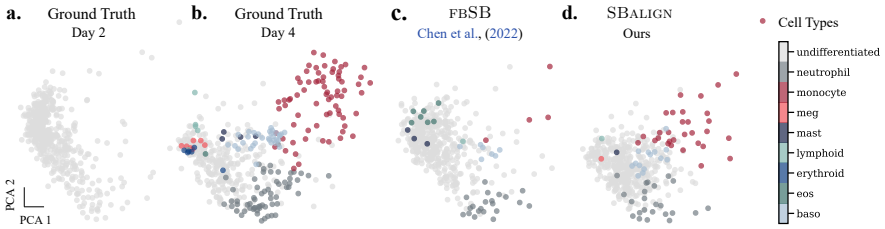


Figure 7.9: Cell type prediction on the differentiation dataset. All distributions are plotted on the first two principal components. **a-b:** Ground truth cell types on day 2 and day 4 respectively. **c-d:** FBSB and SBALIGN cell type predictions on day 4. SBALIGN is able to better model the underlying differentiation processes and capture the diversity in cell types.

7.3.3.2 *Single-Cell Dynamics*

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).
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 time. The dataset consists of two snapshots: the first, recorded on day 2 when most cells are still undifferentiated (see Fig. 7.9a), and a second, on day 4, comprising many different mature cell types (see Fig. 7.9b). For details on the dataset, see Appendix A.2.5. 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.
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. To assess the performance of SBALIGN and the baselines, we monitor several metrics, which include distributional distances, i.e., MMD (Gretton

Table 7.3: Cell differentiation prediction results. Means and standard deviations (in parentheses) of distributional metrics (MMD, W_ε), alignment-based metrics (ℓ_2, RMSD), and cell type classification accuracy.

	Cell Differentiation				
Methods	MMD ↓	W_ε ↓	ℓ_2(PS) ↓	RMSD ↓	Class. Acc. ↑
FBSB	1.55e-2 (0.03e-2)	12.50 (0.04)	4.08 (0.04)	9.64e-1 (0.02e-1)	56.2% (0.7%)
FBSB with SBALIGN	5.31e-3 (0.25e-3)	10.54 (0.08)	0.99 (0.12)	9.85e-1 (0.07e-1)	47.0% (1.5%)
SBALIGN	1.07e-2 (0.01e-2)	11.11 (0.02)	1.24 (0.02)	9.21e-1 (0.01e-1)	56.3% (0.7%)

et al., 2012) and W_ε (Cuturi, 2013), as well as average perturbation signatures (PS), i.e., ℓ_2(PS), and root-mean-square deviation (RMSD). Details on the considered evaluation metrics can be found in Appendix A.3. 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. actual cell type for all the models.

SBALIGN finds matching between cell states on days 2 and 4 (Fig. 7.8c, bottom) which resemble the observed ones (Fig. 7.8a) but also reconstructs the entire evolution path of transcriptomic profiles (Fig. 7.8c, top). It outperforms the baseline FBSB (Table 7.3) in all metrics: Remarkably, our method exceeds the performances of the baseline also on distributional metrics and not uniquely on alignment-based ones. We also leverage SBALIGN predictions to recover the type of cells at the end of the differentiation process (Fig. 7.9d). We do that by training a classifier on differentiated cells observed on day 4 and subsequently classify our predictions. 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).

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. 7.8c) 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. 7.9d 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 Table 7.3).

7.3.4 *Discussion*

In this section, 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)

Optimal transport, both through its theory and computation, has enabled breakthroughs using a multi-pronged approach, blending elements from convex optimization, e.g., linear and quadratic assignment problems, the Sinkhorn algorithm; analysis, e.g., partial differential equations (Jordan et al., 1998; Bunne et al., 2022b) with links to Monge-Ampère equation (Caffarelli, 2003); stochastic calculus, e.g., diffusion models (Song et al., 2021) and Schrödinger bridges (De Bortoli et al., 2021b; Chen et al., 2022a; Bunne et al., 2023a); statistics, e.g., analysis of sampling algorithms (Weed and Bach, 2019), generalized quantiles (Carlier et al., 2016; Cuturi et al., 2019), and generative model fitting (Salimans et al., 2018; Genevay et al., 2018; Bunne et al., 2019, 2023b); as well as deep architectures (De Bie et al., 2019). As such, it provides a unifying framework for modeling population dynamics through maps (Section 3.1.1), as well as ordinary (Section 3.2.2), partial (Section 3.2.3), or stochastic differential equations (Section 3.2.5).
This thesis introduced (●) neural network-based parameterization of these different flavors of OT and demonstrated their far-reaching applications in the field of biomedicine. Leveraging prior successes of OT in single-cell biology (Schiebinger et al., 2019; Lavenant et al., 2021), the neural OT algorithms presented herein (●) incorporate OT as an inductive bias within deep learning frameworks. This approach is motivated by OT's unique capability to realign (●) distributions and model their evolution over time. Consequently, it enables us to (●) reconstruct single-cell dynamics from disparate and unpaired measurements. These methods are now an integral part of open-source libraries (Cuturi et al., 2022; Klein et al., 2023) and show strong quantitative improvements and an increase in robustness to noise, making them a go-to method to determine how perturbations affect cellular properties, to reconstruct the most likely trajectory single cells take upon perturbation, and subsequently to assist in a better understanding of driving factors of cell fate decision and cellular evasion mechanisms.

CONTRIBUTIONS AND SUMMARY

In the following, we summarize the contributions of this thesis, discuss current limitations and open questions, and provide an outlook on an exciting avenue of future work. For this, let us revisit the following core questions we addressed:

How to learn dynamic treatment responses at the single cell level and make predictions to unseen patients?

Learning perturbation responses of an existing patient cohort enables inference of treatment responses for new, previously unseen patients, assuming that we capture heterogeneous drug reactions of patients during training. We thus seek to learn a perturbation model that robustly describes the cellular dynamics upon intervention while still accounting for underlying variability across samples. In Chapter 4 we proposed to learn a neural Monge map (3.1) that aligns the control and perturbed cell population by parameterizing the OT semi-dual (3.6) using input convex neural networks (Bunne et al., 2023b; Makkuva et al., 2020; Amos et al., 2017). By inducing an important theory-motivated inductive bias essential to model stability, these approaches outperforms prior methods in predicting heterogeneous tumor cell responses to a diverse set of cancer drugs and are currently employed to predict treatment outcomes of a large clinical study (Irmisch et al., 2021).

How to predict the outcome of combination therapies and adapt our tools to scalable high-resolution methods?

While these models can predict cell perturbation responses to single drugs, a key challenge in the treatment of many diseases is to predict the effect of combination therapies. Besides the algorithmic challenge, the space of possible combinations is much vaster than the number of cells one can measure, resulting in highly under-sampled experiments. To scale up the experimental capacity, recent studies have resorted to random and composite experiments (Norman et al., 2019; Cleary and Regev, 2020). Chapter 5 has thus focused on proposing a general framework that allows to not only predict the outcomes of combination treatments but also to train models on random, composite experiments. We achieve this by learning a parametric family of context-aware transport maps instead of individual maps for each condition (Bunne et al., 2022a). The approach trains a single data-efficient

model for all perturbations and respective combinations, where the considered setting is passed as a context variable to the model. This also enables us to predict the outcomes of combination therapies —in line with the setup of random, composite large-scale experiments. Beyond, the learned map can generalize to unseen combination therapies once trained on similar contexts.

How can we model heterogeneous continuous-time dynamics and trajectories from discrete-time measurements?

Beyond mappings, OT provides a mathematical link to geometric variational frameworks that allow studying flows of distributions on metric spaces (see Section 3.2). This enables us to model cellular dynamics as control problems described through systems of stochastic (Bunne et al., 2023a; Somnath et al., 2023, Chapter 7) or partial differential equations (Bunne et al., 2022b, Chapter 6). In particular, Chapter 6 proposed a causal model for population dynamics relying on the Jordan-Kinderlehrer-Otto (JKO) flow, widely regarded as one of the most influential mathematical breakthroughs in recent history. By modeling dynamics as a gradient flow, cells decrease collectively an energy which one seeks to learn, e.g., the Waddington potential (Bunne et al., 2022b).
Further, cell fate decisions are of stochastic nature and cellular dynamics intrinsically noisy (Wilkinson, 2009). Approaches treating cellular behavior 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 through diffusion Schrödinger bridges, Chapter 7 introduces several approaches that account for such biological heteroscedasticity (Bunne et al., 2023a; Somnath et al., 2023).

LIMITATIONS

Single-cell expression profiling technoologies provide a detailed look into the molecular states of individual cells. Due to their destructive nature, however, they do not allow continuous measurements of molecular properties over time. While numerous method aim to uncover trajectories of single cells from population data, they all face the same challenge: Sequentially observed distribution of cell states can be potentially produced by multiple dynamics and mechanisms of gene regulation. The *ill-defined nature*

of the problem thus makes it necessary to pose certain assumptions on the underlying cellular dynamics (Weinreb et al., 2018):
The mathematical foundation of this work builds on the biological intuition that perturbations incrementally alter the molecular profiles of cells. If this principle is violated, however, and perturbations strongly disrupt the population to an unidentifiable level, the performance of optimal transport-based algorithms decreases. As baseline methods are similarly effected, these instances call for a more complicated mathematical machinery. Such tools, however, are so far unable to scale to settings with more than a few genes (Heydari et al., 2022) and ultimately, a fine granularity of measurements throughout the time course is required to successfully recover large cell state changes between consecutive time points (Tritschler et al., 2019).

We also observe that the predictive performance of neural optimal transport methods drops when perturbation effects are too strong, e.g., a drug strongly disturbs the cell states; a similar decrease in performance is observed for the considered baselines. The principle underlying the optimal transport theory is ideally suited for acute cellular perturbations during which single cells do not redistribute entirely and undergo arbitrary changes in multidimensional measurement space, but typically only in a few dimensions, such that the overall correlation structure of both distributions 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, *mechanistic* models of the underlying biology or models that integrate observations of multiple smaller time steps (Raue et al., 2015; Busch et al., 2015).
Furthermore, if a system exhibits rotations and oscillations within two consecutive snapshots that are not captured by measurements, data-driven models without additional knowledge (including OT-based methods) will not be able to recover such complex dynamics (Weinreb et al., 2018). 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. We leave it to future work, to investigate choices of alternative cost functions.
Beyond, the current system is not able to recover perturbations (other than cell flux) that have effects on the cell counts, for example, through proliferation and death events (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; Pariset et al., 2023; Chen et al., 2022c; Baradat and Lavenant, 2021).
Lastly, despite having provided a proof-of-concept of the capacity of neural OT methods 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, the generalization capacity of the proposed methods 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.

VISION FOR THE FUTURE

The challenges posed by the inherently complex and constantly changing patterns of interactions in biological systems call for innovative computational solutions. Through the development of static and dynamic neural optimal transport schemes, this thesis has laid a solid foundation for addressing these challenges. As we look ahead, several key questions emerge that shape the future of this research:

Does knowledge of chemical and genetic perturbation responses help us to identify new drug targets?

Which genetic perturbation is connected to a drug's mode of action? Given a genetic cause of disease, which drug should be selected for the successful treatment of a patient? Chapter 4 and Chapter 5 contributed to our understanding of heterogeneous single-cell responses to chemical and genetic perturbations (Bunne et al., 2019, 2022a). Understanding similarities and detecting correspondences between the genetic roots of the disease and the mode of action of chemical drugs are key to the discovery of targets for

which new drugs need to be designed. While optimal transport can be used to describe the evolution of measures over time, one can also rely on its ability to return a matching based on the structural similarity of objects (Mémoli, 2011; Bunne et al., 2019), making it a centerpiece for solving this task. Building on intuitions acquired when solving matching problems of unaligned cell populations, future research will develop OT-based machine learning algorithms to identify novel targets for drug development.

How to develop methods embracing the nature of massively parallelized high-throughput methods?

For decades, experimental biology required us to specifically select the targets we want to observe. Modern massively parallelized high-throughput methods have made it possible to measure thousands of such targets (possibly in combination) in one experiment and collect rich information about the gene expression profiles of each cell. Current computational methodology, however, is often overwhelmed with this depth of data resolution. In Chapter 5, we have provided methods that can deal with the randomized and composite nature of approaches, such as compressed Perturb-Seq (Dixit et al., 2016; Cleary et al., 2017; Cleary and Regev, 2020; Roohani et al., 2022). But our work does not end there: With the rise of new methodologies such as Live-seq (Chen et al., 2022b, Section 7.3) or added levels of dimensions through spatial transcriptomics (Marx, 2021), we need to drive the design of machine learning algorithms even further. Disease can not be solely modeled from the single-cell perspective, but emerges in tissues. Modeling tissue structures and interactions by utilizing principles of geometric deep learning (Fischer et al., 2023) holds the promise to further unveil mechanisms of human disease and thus makes for an exciting avenue of future work.

What are objective milestones toward end-to-end machine learning-guided treatment discovery and planning?

This thesis has evolved around achieving the trifecta of (i.) theory-inspired and (ii.) modular deep learning architectures (iii.) with appropriate inductive biases. Integrating these paradigms into machine learning methods that operate on various challenges of treatment discovery and planning has allowed novel biological insights and created a new state-of-the-art: OT lets us predict meaningful perturbation responses

of a patient on the level of single cells (Bunne et al., 2023b, 2022a), and thus improves our understanding of disease mechanisms by deciphering the underlying molecular processes (Bunne et al., 2022b, 2023a). Further, predicting geometrically valid conformations of a drug-target pair increases the accuracy of current drug design pipelines (Somnath et al., 2021; Ganea et al., 2022; Somnath et al., 2023).

This provides fertile grounds to start asking bigger questions: How can we incorporate insights made from predicting a patient's drug response into target discovery (Huang et al., 2022)? Or, inversely, how to use knowledge of drugs and their connected targets when designing personalized treatment plans (Nilforoshan et al., 2023)? Building on recent developments in reinforcement learning and control theory, how do we innovate the design of personalized, sequential combination therapies for patients? These questions reveal that we cannot address the different challenges that arise throughout the drug design process in isolation. Taking this thesis as a starting point, future work will focus on developing mathematical and computational principles for end-to-end machine learning-guided treatment discovery and planning: By designing feedback loops that allow information sharing across different stages of drug design.

BIBLIOGRAPHY

Britt Adamson, Thomas M Norman, Marco Jost, Min Y Cho, James K Nuñez, Yuwen Chen, Jacqueline E Villalta, Luke A Gilbert, Max A Horlbeck, Marco Y Hein, Ryan A Pak, Andrew N Gray, Carol A Gross, Atray Dixit, Oren Parnas, Aviv Regev, and Jonathan S Weissman. A multiplexed single-cell CRISPR screening platform enables systematic dissection of the unfolded protein response. *Cell*, 167(7), 2016.

Michael S Albergo, Nicholas M Boffi, and Eric Vanden-Eijnden. Stochastic interpolants: A unifying framework for flows and diffusions. *arXiv Preprint arXiv:2303.08797*, 2023.

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. In *International Conference on Machine Learning (ICML)*, 2023.

André-Marie Ampère. *Mémoire contenant l'application de la théorie exposée dans le XVII. e Cahier du Journal de l'École polytechnique, à l'intégration des équations aux différentielles partielles du premier et du second ordre*. De l'Imprimerie royale, 1819.

Brian DO Anderson. Reverse-time diffusion equation models. *Stochastic Processes and their Applications*, 12(3), 1982.

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

Aymeric Baradat and Hugo Lavenant. Regularized unbalanced optimal transport as entropy minimization with respect to branching Brownian motion. *arXiv Preprint arXiv:2111.01666*, 2021.

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.

Alev Baysoy, Zhiliang Bai, Rahul Satija, and Rong Fan. The technological landscape and applications of single-cell multi-omics. *Nature Reviews Molecular Cell Biology*, 2023.

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.

Yoshua Bengio, Réjean Ducharme, and Pascal Vincent. A Neural Probabilistic Language Model. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 13, 2000.

Doris Berchtold, Nico Battich, and Lucas Pelkmans. A systems-level study reveals regulators of membrane-less organelles in human cells. *Molecular Cell*, 72(6), 2018.

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. *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, Michiel Van De Panne, Sylvain Paris, and Wolfgang Heidrich. Displacement interpolation using Lagrangian mass transport. In *Proceedings of the 2011 SIGGRAPH Asia conference*, 2011.

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.

Johannes Brandstetter, Daniel Worrall, and Max Welling. Message Passing Neural PDE Solvers. In *International Conference on Learning Representations (ICLR)*, 2022.

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.

Eric Brouzes, Martina Medkova, Neal Savenelli, Dave Marran, Mariusz Twardowski, J Brian Hutchison, Jonathan M Rothberg, Darren R Link, Norbert Perrimon, and Michael L Samuels. Droplet microfluidic technology for single-cell high-throughput screening. *Proceedings of the National Academy of Sciences*, 106(34), 2009.

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, Marco 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 Rätsch. 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.

Katrin Busch, Kay Klapproth, Melania Barile, Michael Flossdorf, Tim Holland-Letz, Susan M Schlenner, Michael Reth, Thomas Höfer, and Hans-Reimer Rodewald. Fundamental properties of unperturbed haematopoiesis from stem cells in vivo. *Nature*, 518(7540), 2015.

Luis A Caffarelli. Interior $W^{2,p}$ estimates for solutions of the Monge-Ampère equation. *Annals of Mathematics*, 1990.

Luis A Caffarelli. Allocation Maps with General Cost Functions. In *Partial Differential Equations and Applications*, volume 177 of Lecture Notes in Pure and Appl. Math. Dekker, 1996.

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

Luis A Caffarelli. The Monge-Ampère equation and optimal transportation, an elementary review. *Lecture Notes in Mathematics: Optimal Transportation and Applications*, 1813, 2003.

Kenneth F Caluya and Abhishek Halder. Wasserstein Proximal Algorithms for the Schrödinger Bridge Problem: Density Control with Nonlinear Drift. *IEEE Transactions on Automatic Control*, 67(3), 2021.

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

Guillaume Carlier, Victor Chernozhukov, and Alfred Galichon. Vector quantile regression: an optimal transport approach. *Ann. Statist.*, 44(3), 2016.

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

Anne E Carpenter, Thouis R Jones, Michael R Lamprecht, Colin Clarke, In Han Kang, Ola Friman, David A Guertin, Joo Han Chang, Robert A Lindquist, Jason Moffat, et al. CellProfiler: image analysis software for identifying and quantifying cell phenotypes. *Genome Biology*, 7(10), 2006.

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.

Ricky TQ Chen, Yulia Rubanova, Jesse Bettencourt, and David Duvenaud. Neural Ordinary Differential Equations. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

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.

Yongxin Chen, Tryphon T Georgiou, and Michele Pavon. The most likely evolution of diffusing and vanishing particles: Schrödinger Bridges with unbalanced marginals. *SIAM Journal on Control and Optimization*, 60(4), 2022c.

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, 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.

Sumit Chopra, Raia Hadsell, and Yann LeCun. Learning a similarity metric discriminatively, with application to face verification. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1. IEEE, 2005.

Brian Cleary and Aviv Regev. The necessity and power of random, under-sampled experiments in biology. *arXiv Preprint 2012.12961*, 2020.

Brian Cleary, Le Cong, Anthea Cheung, Eric S Lander, and Aviv Regev. Efficient generation of transcriptomic profiles by random composite measurements. *Cell*, 171(6), 2017.

Julian D Cole. On a quasi-linear parabolic equation occurring in aerodynamics. *Quarterly of Applied Mathematics*, 9(3), 1951.

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.

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

Marco Cuturi, Olivier Teboul, and Jean-Philippe Vert. Differentiable Ranks and Sorting using Optimal Transport. *Advances in Neural Information Processing Systems (NeurIPS)*, 32, 2019.

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.

Ibiayi Dagogo-Jack and Alice T Shaw. Tumour heterogeneity and resistance to cancer therapies. *Nature Reviews Clinical Oncology*, 15(2), 1991.

Paolo Dai Pra. A stochastic control approach to reciprocal diffusion processes. *Applied mathematics and Optimization*, 23(1), 1991.

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.

Jan De Leeuw and Patrick Mair. Multidimensional Scaling Using Majorization: SMACOF in R. *Journal of Statistical Software*, 31, 2009.

Bachir El Debs, Ramesh Utharala, Irina V Balyasnikova, Andrew D Griffiths, and Christoph A Merten. Functional single-cell hybridoma screening using droplet-based microfluidics. *Proceedings of the National Academy of Sciences*, 109(29), 2012.

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.

Ioannis Exarchos and Evangelos A Theodorou. Stochastic optimal control via forward and backward stochastic differential equations and importance sampling. *Automatica*, 87, 2018.

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.

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.

Chris Finlay, Augusto Gerolin, Adam M Oberman, and Aram-Alexandre Pooladian. Learning normalizing flows from Entropy-Kantorovich potentials. *arXiv Preprint arXiv:2006.06033*, 2020.

David S Fischer, Anna K Fiedler, Eric M Kernfeld, Ryan MJ Genga, Aimée Bastidas-Ponce, Mostafa Bakhti, Heiko Lickert, Jan Hasenauer, Rene Maehr, and Fabian J Theis. Inferring population dynamics from single-cell RNA-sequencing time series data. *Nature Biotechnology*, 37(4), 2019.

David S Fischer, Anna C Schaar, and Fabian J Theis. Modeling intercellular communication in tissues using spatial graphs of cells. *Nature Biotechnology*, 41(3), 2023.

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

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.

Fabian Fröhlich, Thomas Kessler, Daniel Weindl, Alexey Shadrin, Leonard Schmiester, Hendrik Hache, Artur Muradyan, Moritz Sch"utte, Ji-Hyun Lim, Matthias Heinig, et al. Efficient parameter estimation enables the prediction of drug response using a mechanistic pan-cancer pathway model. *Cell Systems*, 7(6), 2018.

Octavian-Eugen Ganea, Lagnajit Pattanaik, Connor W Coley, Regina Barzilay, Klavs F Jensen, William H Green, and Tommi S Jaakkola. GeoMol: Torsional Geometric Generation of Molecular 3D Conformer Ensembles. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

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.

Wilfrid Gangbo and Robert J McCann. Optimal maps in Monge's mass transport problem. *Comptes Rendus de l'Academie des Sciences-Serie I-Mathematique*, 321(12), 1995.

Wilfrid Gangbo and Robert J McCann. The geometry of optimal transportation. *Acta Mathematica*, 177(2), 1996.

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.

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 Matemática 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.

Will Grathwohl, Ricky TQ Chen, Jesse Bettencourt, Ilya Sutskever, and David Duvenaud. FFJORD: Free-Form Continuous Dynamics for Scalable Reversible Generative Models. In *International Conference on Learning Representations (ICLR)*, 2019.

Victoria A Green and Lucas Pelkmans. A systems survey of progressive host-cell reorganization during rotavirus infection. *Cell Host & Microbe*, 20(1), 2016.

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

Manuel Guizar-Sicairos, Samuel T Thurman, and James R Fienup. Efficient subpixel image registration algorithms. *Optics Letters*, 33(2), 2008.

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, Tomás 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.

Eberhard Hopf. The Partial Differential Equation $u_t + uu_x = \mu^*_{xx}$. *Communications on Pure and Applied Mathematics*, 3(3), 1950.

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.

Kexin Huang, Tianfan Fu, Wenhao Gao, Yue Zhao, Yusuf Roohani, Jure Leskovec, Connor W Coley, Cao Xiao, Jimeng Sun, and Marinka Zitnik. Artificial intelligence foundation for therapeutic science. *Nature chemical biology*, 18(10), 2022.

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.

Aapo Hyvärinen and Peter Dayan. Estimation of Non-Normalized Statistical Models by Score Matching. *Journal of Machine Learning Research (JMLR)*, 6(4), 2005.

Anja Irmisch, Ximena Bonilla, Stéphane Chevrier, Kjong-Van Lehmann, Franziska Singer, Nora C Toussaint, Cinzia Esposito, Julien Mena, Emanuela S Milani, Ruben Casanova, et al. The Tumor Profiler Study: Integrated, multi-omic, functional tumor profiling for clinical decision support. *Cancer Cell*, 39(3), 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.

Junteng Jia and Austin R Benson. Neural Jump Stochastic Differential Equations. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019.

Qingzhu Jia, Han Chu, Zheng Jin, Haixia Long, and Bo Zhu. High-throughput single-cell sequencing in cancer research. *Signal Transduction and Targeted Therapy*, 7(1), 2022.

Wengong Jin, Jeremy Wohlwend, Regina Barzilay, and Tommi Jaakkola. Iterative Refinement Graph Neural Network for Antibody Sequence-Structure Co-design. In *International Conference on Learning Representations (ICLR)*, 2022.

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

MH Julius, T Masuda, and LA Herzenberg. Demonstration that antigen-binding cells are precursors of antibody-producing cells after purification with a fluorescence-activated cell sorter. *Proceedings of the National Academy of Sciences*, 69(7), 1972.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Zídek, Anna Potapenko, et al. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873), 2021.

Kenji Kamimoto, Christy M Hoffmann, and Samantha A Morris. Dissecting cell identity via network inference and in silico gene perturbation. *Nature*, 2023.

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.

Thomas N Kipf and Max Welling. Semi-Supervised Classification with Graph Convolutional Networks. In *International Conference on Learning Representations (ICLR)*, 2017.

Dominik Klein, Giovanni Palla, Marius Lange, Michal Klein, Zoe Piran, Manuel Gander, Laetitia Meng-Papaxanthos, Michael Sterr, Aimee Bastidas-Ponce, Marta Tarquis-Medina, et al. Mapping cells through time and space with moscot. *bioRxiv*, 2023.

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.

Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. DiffWave: A Versatile Diffusion Model for Audio Synthesis. In *International Conference on Learning Representations (ICLR)*, 2021.

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.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. ImageNet Classification with Deep Convolutional Neural Networks. *Advances in Neural Information Processing Systems (NeurIPS)*, 25, 2012.

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.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-Based Learning Applied to Document Recognition. *Proceedings of the IEEE*, 86(11), 1998.

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.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow Matching for Generative Modeling. In *International Conference on Learning Representations (ICLR)*, 2023.

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

Guan-Horng Liu, Arash Vahdat, De-An Huang, Evangelos A Theodorou, Weili Nie, and Anima Anandkumar. I^2 SB: Image-to-Image Schrödinger Bridge. In *International Conference on Machine Learning (ICML)*, 2023a.

Xingchao Liu, Lemeng Wu, Mao Ye, and Qiang Liu. Flow Straight and Fast: Learning to Generate and Transfer Data with Rectified Flow. *arXiv Preprint arXiv:2209.03003*, 2022b.

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

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.

Malte D Luecken and Fabian J Theis. Current best practices in single-cell RNA-seqanalysis: a tutorial. *Molecular Systems Biology*, 15(6), 2019.

Jin Ma and Jiongmin Yong. *Forward-Backward Stochastic Differential Equations and their Applications*. Springer Science & Business Media, 1999.

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.

Gail R Martin and Martin J Evans. Differentiation of Clonal Lines of Teratocarcinoma Cells: Formation of Embryoid Bodies In Vitro. *Proceedings of the National Academy of Sciences (PNAS)*, 72(4), 1975.

Vivien Marx. Method of the Year: spatially resolved transcriptomics. *Nature Methods*, 18(1), 2021.

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

Linas Mazutis, John Gilbert, W Lloyd Ung, David A Weitz, Andrew D Griffiths, and John A Heyman. Single-cell analysis and sorting using droplet-based microfluidics. *Nature Protocols*, 8(5), 2013.

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.

Al Mead. Review of the Development of Multidimensional Scaling Methods. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 41(1), 1992.

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.

Toshio Mikami. Dynamical Systems in the Variational Formulation of the Fokker-Planck Equation by the Wasserstein Metric. *Applied Mathematics and Optimization*, 42, 2000.

Toshio Mikami. Optimal control for absolutely continuous stochastic processes and the mass transportation problem. *Electronic Communications in Probability*, 2002.

Toshio Mikami and Michéle Thieullen. Optimal transportation problem by stochastic optimal control. *SIAM Journal on Control and Optimization*, 47, 2008.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient Estimation of Word Representations in Vector Space. *International Conference on Learning Representations (ICLR), Workshop Track*, 2013a.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed Representations of Words and Phrases and their Compositionality. *Advances in Neural Information Processing Systems (NeurIPS)*, 26, 2013b.

Tomávs Mikolov, Wen-tau Yih, and Geoffrey Zweig. Linguistic Regularities in Continuous Space Word Representations. In *Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, 2013c.

Tom M Mitchell. The Need for Biases in Learning Generalizations. Technical Report CBM-TR-117, Rutgers University, 1980.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540), 2015.

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.

Quan Hoang Nhan Dam, Trung Le, Tu Dinh Nguyen, Hung Bui, and Dinh Phung. Three-Player Wasserstein GAN via Amortised Duality. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2019.

Hamed Nilforoshan, Michael Moor, Yusuf Roohani, Yining Chen, Anja Šurina, Michihiro Yasunaga, Sara Oblak, and Jure Leskovec. Zero-shot causal learning. *arXiv Preprint arXiv:2301.12292*, 2023.

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.

Matteo Pariset, Ya-Ping Hsieh, Charlotte Bunne, Andreas Krause, and Valentin De Bortoli. Unbalanced Diffusion Schrödinger Bridge. *arXiv preprint arXiv:2306.09099*, 2023.

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.

F. Pedregosa, G. Varoquaux, A. Gramfort, V Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. VanderPlas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research (JMLR)*, 12, 2011.

Stefan Peidli, Tessa Durakis Green, Ciyue Shen, Torsten Gross, Joseph Min, Jake Taylor-King, Debora Marks, Augustin Luna, Nils Bluthgen, and Chris Sander. scPerturb: Information resource for harmonized single-cell perturbation data. *bioRxiv*, 2022.

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, Heli Ben-Hamu, Carles Domingo-Enrich, Brandon Amos, Yaron Lipman, and Ricky Chen. Multisample Flow Matching: Straightening Flows with Minibatch Couplings. In *International Conference on Machine Learning (ICML)*, 2023a.

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)*, 2023b.

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.

Maziar Raissi, Paris Perdikaris, and George E Karniadakis. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics*, 378, 2019.

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.

Andreas Raue, Bernhard Steiert, Max Schelker, Clemens Kreutz, Tim Maiwald, Helge Hass, Joep Vanlier, Christian Tönsing, Lorenz Adlung, Raphael Engesser, et al. Data2dynamics: a modeling environment tailored to parameter estimation in dynamical systems. *Bioinformatics*, 31(21), 2015.

Danilo Rezende and Shakir Mohamed. Variational Inference with Normalizing Flows. In *International Conference on Machine Learning (ICML)*, volume 37, 2015.

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.

Hannes Risken. *The Fokker-Planck Equation*. Springer, 1996.

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.

Yusuf Roohani, Kexin Huang, and Jure Leskovec. GEARS: Predicting transcriptional outcomes of novel multi-gene perturbations. *BioRxiv*, 2022.

Litu Rout, Alexander Korotin, and Evgeny Burnaev. Generative Modeling with Optimal Transport Maps. In *International Conference on Learning Representations (ICLR)*, 2021.

Ludger Rüschendorf. Bounds for Distributions with Multivariate Marginals. *Lecture Notes-Monograph Series*, 1991a.

Ludger Rüschendorf. Fréchet-bounds and their applications. In *Advances in Probability Distributions with Given Marginals*, volume 67 of Mathematics and Its Applications. Springer, 1991b.

Sergei Rybakov, Mohammad Lotfollahi, Fabian J Theis, and F Alexander Wolf. Learning interpretable latent autoencoder representations with annotations of feature sets. *Machine Learning in Computational Biology*, 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.

Geoffrey Schiebinger. Reconstructing developmental landscapes and trajectories from single-cell data. *Current Opinion in Systems Biology*, 27, 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.

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

Erwin Schrödinger. Sur la théorie relativiste de l'électron et l'interprétation de la mécanique quantique. In *Annales de l'institut Henri Poincaré*, volume 2, 1932.

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.

Vivien Seguy, Bharath Bhushan Damodaran, Rémi Flamary, Nicolas Courty, Antoine Rolet, and Mathieu Blondel. Large-Scale Optimal Transport and Mapping Estimation. In *International Conference on Learning Representations (ICLR)*, volume 30, 2018.

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.

Michael J Shamblott, Candace L Kerr, Joyce Axelman, John W Littlefield, Gregory O Clark, Ethan S Patterson, Russell C Addis, Jennifer N Kraszewski, Kathleen C Kent, and John D Gearhart. Derivation and Differentiation of Human Embryonic Germ Cells. In *Essentials of Stem Cell Biology*. Elsevier, 2009.

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

David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *Nature*, 529(7587), 2016.

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.

Berend Snijder, Raphael Sacher, Pauli Rämö, Prisca Liberali, Karin Mench, Nina Wolfrum, Laura Burleigh, Cameron C Scott, Monique H Verheije, Jason Mercer, et al. Single-cell analysis of population context advances RNAi screening at multiple levels. *Molecular Systems Biology*, 8(1):579, 2012.

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

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

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.

David G Spiller, Christopher D Wood, David A Rand, and Michael RH White. Measurement of single-cell dynamics. *Nature*, 465(7299), 2010.

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.

Vasileios Stathias, Anna M Jermakowicz, Marie E Maloof, Michele Forlin, Winston Walters, Robert K Suter, Michael A Durante, Sion L Williams, J William Harbour, Claude-Henry Volmar, et al. Drug and disease signature integration identifies synergistic combinations in glioblastoma. *Nature Communications*, 9(1), 2018.

Thomas Stoeger, Nico Battich, Markus D Herrmann, Yauhen Yakimovich, and Lucas Pelkmans. Computer vision for image-based transcriptomics. *Methods*, 85, 2015.

Amirhossein Taghvaei and Amin Jalali. 2-Wasserstein Approximation via Restricted Convex Potentials with Application to Improved Training for GANs. *arXiv Preprint arXiv:1902.07197*, 2019.

Kazutoshi Takahashi and Shinya Yamanaka. Induction of Pluripotent Stem Cells from Mouse Embryonic and Adult Fibroblast Cultures by Defined Factors. *Cell*, 126(4), 2006.

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.

Andrew E Teschendorff and Andrew P Feinberg. Statistical mechanics meets single-cell biology. *Nature Reviews Genetics*, 22(7), 2018.

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.

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.

W Tvarusko, M Bentele, Tom Misteli, R Rudolf, C Kaether, David L Spector, HH Gerdes, and R Eils. Time-resolved analysis and visualization of dynamic processes in living cells. *Proceedings of the National Academy of Sciences*, 96(14), 1999.

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.

Stefan Van der Walt, Johannes L Schönberger, Juan Nunez-Iglesias, Franccois Boulogne, Joshua D Warner, Neil Yager, Emmanuelle Gouillart, and Tony Yu. scikit-image: image processing in Python. *PeerJ*, 2:e453, 2014.

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.

Conrad Hall Waddington. *The Strategy of the Genes, a Discussion of Some Aspects of Theoretical Biology*. G. Allen and Unwin, 1957.

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.

Jonathan Weed and Francis Bach. Sharp asymptotic and finite-sample rates of convergence of empirical measures in wasserstein distance. *Bernoulli*, 25(4A), 2019.

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.

F Alexander Wolf, Philipp Angerer, and Fabian J Theis. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biology*, 19(1), 2018.

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 D 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 D 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.

Bo Yuan, Ciyue Shen, Augustin Luna, Anil Korkut, Debora S Marks, John Ingraham, and Chris Sander. CellBox: Interpretable Machine Learning forPerturbation Biology with Application to the Designof Cancer Combination Therapy. *Cell Systems*, 12(2), 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.

Grace XY Zheng, Jessica M Terry, Phillip Belgrader, Paul Ryvkin, Zachary W Bent, Ryan Wilson, Solongo B Ziraldo, Tobias D Wheeler, Geoff P McDermott, Junjie Zhu, et al. Massively parallel digital transcriptional profiling of single cells. *Nature Communications*, 8(1), 2017.

CURRICULUM VITAE

PERSONAL DATA

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

EDUCATION

2019 – 2023	Eidgenössische Technische Hochschule (ETH), Zürich, Switzerland *Final Degree:* Doctor of Science
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, Marco 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 Rätsch. 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.

Matteo Manica, Charlotte Bunne, Roland Mathis, Joris Cadow, Mehmet Eren Ahsen, Gustavo A Stolovitzky, and Maria Rodriguez Martinez. COSIFER: a Python package for the consensus inference of molecular interaction networks. *Bioinformatics*, 37(14), 2020.

Matteo Pariset, Ya-Ping Hsieh, Charlotte Bunne, Andreas Krause, and Valentin De Bortoli. Unbalanced Diffusion Schrödinger Bridge. *arXiv Preprint arXiv:2306.09099*, 2023.

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

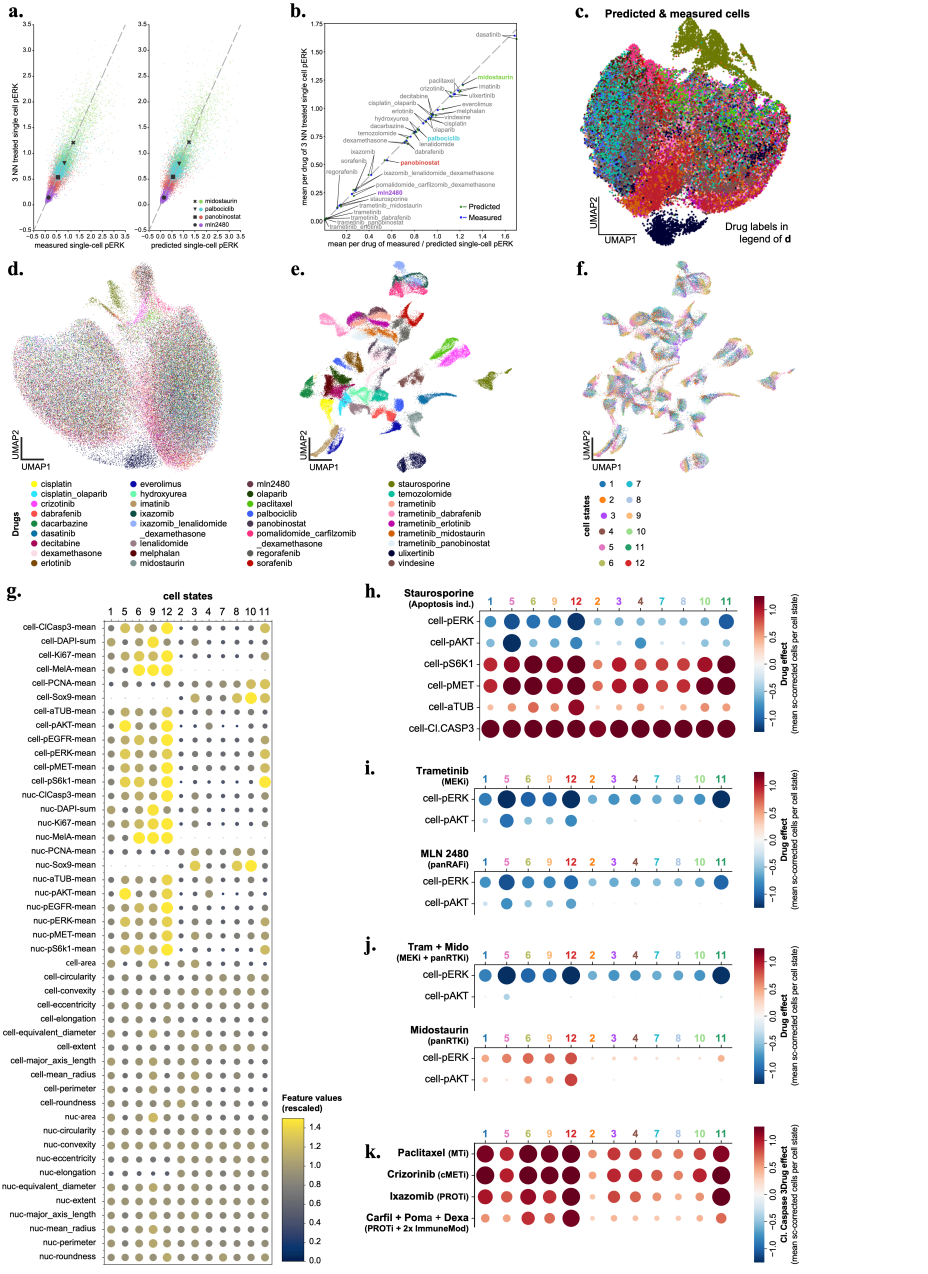


Figure A.1: **a.** High similarity of measured and CellOT-predicted single-cell pERK (phosphor ERK1/2) values at the single-cell level. Scatter plots compare the relationship between measured pERK values of cells (left) treated with Midostaurin (green dots), Palbociclib (blue dots), Panobinostat (red 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. A 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, and 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, and 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, and 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, and each row a drug treatment.

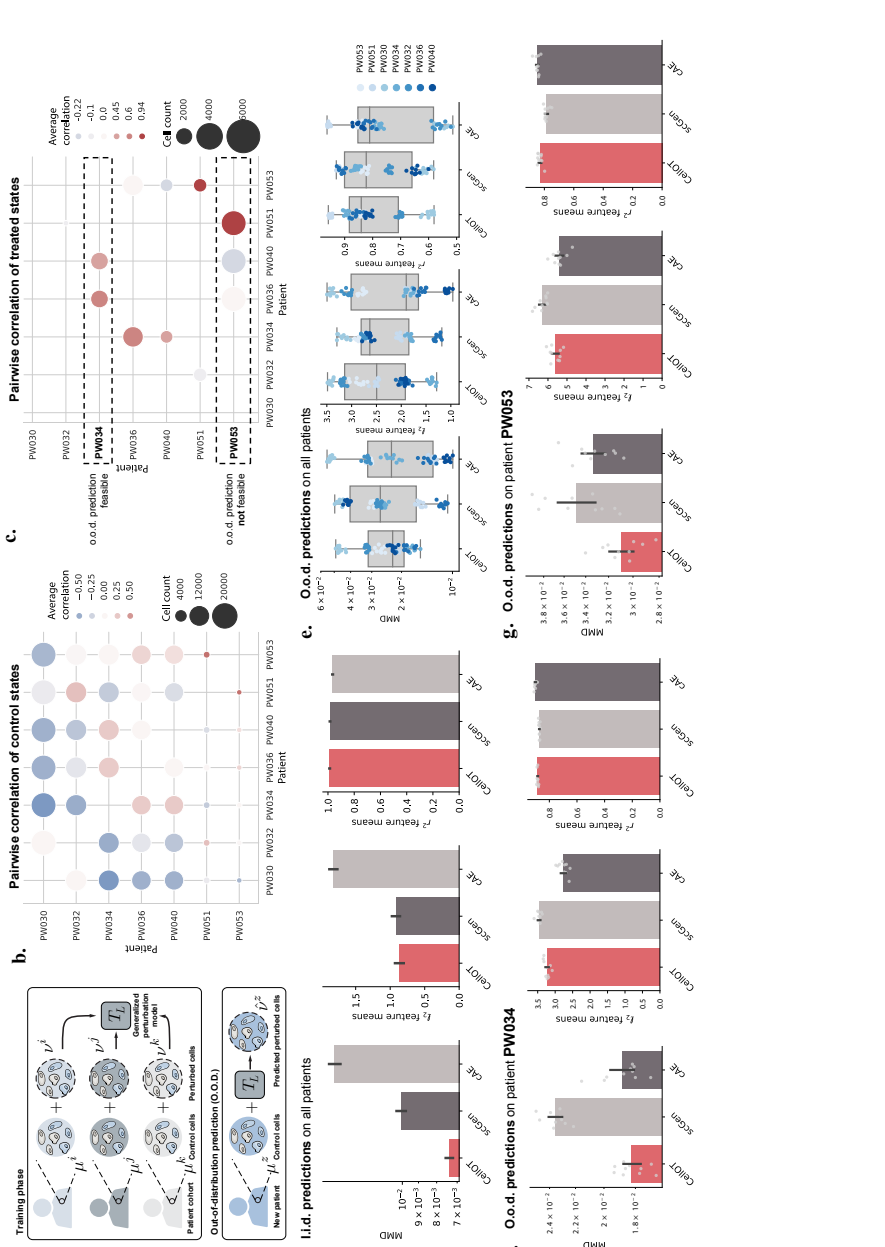

a.
Training phase
Patient cohort
Control cells
Perturbed cells
Generalized perturbation model
Out-of-distribution prediction (O.O.D.)
New patient
Control cells
Predicted perturbed cells
b.
Pairwise correlation of control states
Average correlation
Cell count
Patient
c.
Pairwise correlation of treated states
o.o.d. prediction feasible
o.o.d. prediction not feasible
d.
i.i.d. predictions on all patients
MMD
e.
O.o.d. predictions on all patients
f.
O.o.d. predictions on patient PW034
g.
O.o.d. predictions on patient PW053
CellOT
scGen
CAE

Figure A.2: 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.

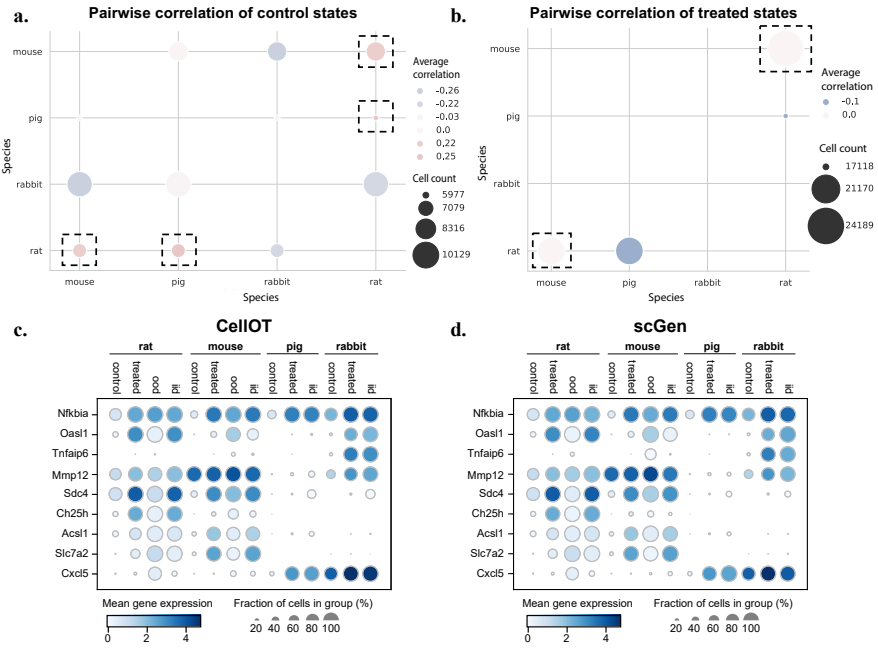


Figure A.3: Analysis and further results of the cross-species dataset. **a.** Pairwise average correlation of the PCA embeddings of the control states between species. **b.** Pairwise average correlation of the PCA embeddings of the treated states between patients, masked to only those patient pairs 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 a 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.

A.2 DATASETS

We evaluate methods introduced in this thesis on different tasks, each consisting of a pair of source μ and target measure ν. In particular, we consider single-cell datasets in which populations of single cells have been monitored with modern high-throughput methods such as single-cell RNA sequencing or optical phenotyping technologies (see Section 2.1). In the following, we introduce each dataset, describe preprocessing steps, feature selection, and data splits.

A.2.1 *Srivatsan et al. (2020)*

Cancer drugs reduce uncontrolled cell growth and proliferation by inhibiting DNA replication and RNA transcription as well as targeting proteins crucial for cancer progression. In doing so, they modulate downstream signaling cascades, affect cell growth and morphology, and alter gene expression profiles of single cells. Srivatsan et al. (2020) conduct a scRNA-seq-based phenotyping screen of transcriptional responses to thousands of independent perturbations at single-cell resolution. The measured cell population contains three well-characterized cancer cell lines, including A549, a human lung adenocarcinoma, K562, a chronic myelogenous leukemia, and MCF7, a mammary adenocarcinoma cell line. Due to the different transcriptional profiles of each cancer cell line, drug compounds might cause divergent cellular responses in each subpopulation. The dataset contains 17,565 control cells as well as a varying number of cells perturbed by different drugs with different dosages, i.e., 10 nM, 100 nM, 1,000 nM, 10,000 nM.

DATA PREPROCESSING. The dataset is available for download in the gene expression omnibus (GEO) database under accession number GSM4150378. For data quality control and preprocessing, we follow the analysis of Lotfollahi et al. (2023). The count matrix obtained from GEO consists of 581,777 cells. The data were subset to half its size, with 290,888 cells remaining after quality control for all 188 different compounds. We proceeded with log-transformation and the selection of 1,000 highly variable genes (HVG) using `scanpy` (Wolf et al., 2018).

FEATURE SELECTION. Single-cell RNA sequencing data is very high-dimensional, even after selecting 1,000 highly-variable genes. For the downstream analysis of how well the overall perturbation effect has been captured, we thus select the top 50 marker genes, i.e., those genes which show strong differences between perturbed and unperturbed states. This analysis is conducted based on the `scanpy`'s function `rank_genes_groups`, setting unperturbed cells as reference (Wolf et al., 2018). The influence of the number of considered marker genes on different evaluation metrics is further analyzed in Bunne et al. (2023b).

A.2.2 *Bunne et al. (2023b)*

Besides scRNA-seq, optical phenotypic screens, e.g., multiplexing tools such as 4i (Gut et al., 2018), are able to capture meaningful features related to both the treatment response heterogeneity (e.g., the phosphorylation or dephosphorylation of a kinase in a signaling pathway) and the pre-existing cell-to-cell variability (e.g., protein levels related to different cellular states or cell cycle phases) which my determine treatment response. In order to derive a proof-of-concept study and test if the proposed methods are able to capture heterogeneous single-cell responses, we consider two co-cultured primary melanoma cell lines (M130219 and M130429), which were derived from the same melanoma patient from different body sites Bunne et al. (2023b). M130219 originates from a subcutaneous biopsy taken during treatment with Bimetinib (MEKi), whereas M130429 was derived from a bone autopsy one month after stopping said targeted therapy (Raaijmakers et al., 2015). Both cell lines share the same driver mutation (NRAS Q61R) but are phenotypically diverse. The cell lines were screened with 34 different drugs, partially applied as combination therapies.

DATA PREPROCESSING. The cells were seeded in a 384-well plate, and allowed to settle and adhere overnight. Drugs and dimethyl sulfoxide as the vehicle control was added to the cells the next morning and incubated for 8 and 24 hours, respectively, after which the cells were fixed with paraformaldehyde. Subsequently, 6 cycles of 4i were performed, for which the images were acquired with an automated high-content microscope. We utilized a mixture of two melanoma tumor cell lines (ratio 1:1) in order to image a total of 97,748.
Consequently, the cell lines are also classed as two different melanoma subtypes due to, amongst others, differences in marker expression (Raaijmakers et al., 2015): the former a mesenchymal subtype ($SOX9^{+}$, $MelA^{-}$), the latter a melanocytic subtype ($Sox9^{-}$, $MelA^{+}$). 10,995 cells were imaged in the DMSO-treated control state and the rest are treated with one of the 34 cancer therapies. Between 2,000 and 3,000 cells are profiled per treatment. All image analysis steps were performed by our in-house platform called `TissueMAPS`. The steps included illumination correction (Snijder et al., 2012), alignment of images from different acquisition cycles using fast Fourier transform (Guizar-Sicairos et al., 2008), segmentation of nuclei and cell outlines (Stoeger et al., 2015), as well cellular and nuclear measurements of

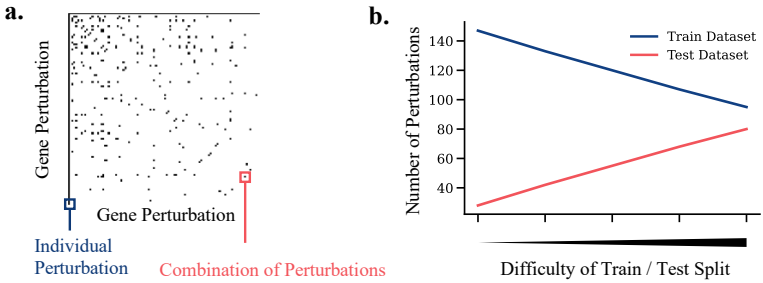


Figure A.4: a. The indicator matrix of all individual perturbations as well as those perturbation pairs available in combination (black) in the dataset by Norman et al. (2019). **b.** Size of the different train/test splits of the dataset by Norman et al. (2019). The train set contains all single perturbations as well as a decreasing number of combinations with increasing difficulty of the data split. For more details, see Appendix A.2.3.

intensity and morphology features using the `scikit-image` library (Van der Walt et al., 2014).
The extracted marker intensities and morphological features are then re-normalized to the same numerical scale by dividing each feature with its 75th percentile computed on control cells. Values are then transformed with a log1p ($x \leftarrow \log(x+1)$) function[6].

FEATURE SELECTION. A total of 47 features are reported, 21 morphological features and 26 protein intensities.

A.2.3 *Norman et al. (2019)*

Genetic interactions and their joint expression give rise to an inconceivable organismal complexity and uncountable many diverse phenotypes and behaviors. Constructing a systematic genetic interaction map is crucial for a better understanding of cellular mechanisms in health and disease. Thus, Norman et al. (2019) conducted single-cell, pooled transcriptional profiling of CRISPR-mediated perturbations to link genetic perturbation to its transcriptional consequences using the Perturb-Seq technology (Dixit et al., 2016; Adamson et al., 2016). The dataset consists of individual perturbations as well as joint overexpression of different genes, allowing us to study the phenotypic consequences of perturbing a pair of genes alone

6 The dataset can be downloaded via `https://doi.org/10.3929/ethz-b-000609681`.

or in combination. The indicator matrix of all individual perturbations as well as those pairs available in combination can be found in Fig. A.4a.

DATA PREPROCESSING. The dataset is available for download in the GEO database under accession number GSE133344. For data quality control and preprocessing, we follow the analysis of Lotfollahi et al. (2023). We discarded those genetic perturbations with less than 250 cells, resulting in a dataset with 92 individual perturbations and 84 perturbations in combination. This further included, the exclusion of particular subsets of control cells with in total of 98,419 remaining, data normalization, log-transformation, and selection of 1,500 highly-variant genes using `scanpy` (Wolf et al., 2018).

FEATURE SELECTION. Similar to Appendix A.2.1, for evaluation we select the top 50 marker genes, i.e., those genes most strongly affected by the particular genetic perturbation.

DATA SPLITS. For the evaluation conducted in Chapter 5, we create different train/test dataset splits of increasing difficulty by following Lotfollahi et al. (2023). The train splits hereby always contain all 92 individual perturbations as well as varying numbers of combinations. The easiest train split contains 55 perturbations, while the test set only carries 28 combinations that are unknown in the evaluation. Consecutive splits get increasingly harder, comprising 42, 29, 16, and 4 combinations in the train set (besides all single perturbations) and 41, 54, 67, and 79 combinations in the test set, respectively (see Fig. A.4b).

A.2.4 *Moon et al. (2019)*

Developmental processes in biology involve tissue and organ development, body axis formation, cell division, and cell differentiation, e.g., the development of stem cells into functional cell types. An example of such a process is the differentiation of ESC into hematopoietic, cardiac, neural, pancreatic, hepatocytic, and germ lineages. This development can be approximated *in vitro* using embryoid bodies (EB) (Martin and Evans, 1975), three-dimensional aggregates of pluripotent stem cells, including ESCs (Shamblott et al., 2009). Recently, Moon et al. (2019) conducted a scRNA-seq analysis to unveil the developmental trajectories, as well as cellular and molecular identities through which early lineage precursors emerge from

human ESCs. The dataset is available via Mendeley Data (V6N743H5NG)[7]. In the following, we describe the preprocessing of the raw scRNA-seq data as well as the lineage branch analysis extracting the functional cell types emerging in this developmental process.

DATA PREPROCESSING. To preprocess the data, we follow the analysis of Moon et al. (2019) as well as Luecken and Theis (2019). Moon et al. (2019) originally measure approximately 31,000 cells over a 27 days differentiation time course, comprising gene expression matrices and barcodes, i.e., DNA tags used to identify reads originating from the same cell. The measured cells are then filtered in a quality control stage, their gene expression levels normalized and further processed in a feature selection step, where only highly-differentiated genes are selected. After quality control, the dataset consists of 15,150 cells and 17,945 genes.

FEATURE SELECTION. We extract 4,000 HVG using the 10X genomics preprocessing software `Cell Ranger` (Zheng et al., 2017) to further reduce the dimensionality of the dataset and include only the most informative genes. Given the resulting data matrix with 15,150 cells and 4,000 genes across 5 different time points, we compute a corresponding low-dimensional embedding using PCA. We use the first 20 or 30 PCs for predicting population dynamics in Chapter 6 and 7. This is in alignment with previous analysis of developmental trajectories, which use 5 (Tong et al., 2020) and 30 PCs (Schiebinger et al., 2019), respectively.

LINEAGE BRANCH ANALYSIS. Besides evaluating how well methods presented in this thesis resemble the spatiotemporal dynamics, we analyze their ability to capture biological heterogeneity. Serving as an *in vitro* model of early embryogenesis, embryoid bodies differentiation captures the development of ESCs into the mesoderm, endoderm, neuroectoderm, neural crest, and others. Using an initial k-means clustering ($k = 30$) and following Moon et al. (2019, Fig. 6, Suppl. Note 4), we compute lineage branch classes for all cells in a 10-dimensional embedding space using PHATE, a non-linear dimensionality reduction method capturing a denoised representation of both the local and global structure of a dataset. Details of the annotation can be found in Bunne et al. (2022b, 2023a).
We then train a k-NN classifier ($k = 5$) to infer the lineage branch class based on a 30-dimensional PCA embedding of a cell (ESC: 0, neural crest:

7 Dataset available via `https://data.mendeley.com/datasets/v6n743h5ng`.

1, neuroectoderm: 2, endoderm: 3, mesoderm: 4, other: 5). Section 6.3.2 and 7.2.6.2 contain an analysis of the captured lineage branch heterogeneity of the predictions by computing the lineage branch class of each cell using the k-NN classifier.

A.2.5 *Weinreb et al. (2020)*

Weinreb et al. (2020) study lineage tracing on transcriptional landscapes in hematopoiesis, the process of blood regeneration in bone marrow, in which multipotent progenitors give rise to red cells of the blood, as well as myeloid and lymphoid immune cell types. In order to dissect how molecular differences among progenitor cells determine their ability to generate mature cell types, it is crucial to understand the hierarchy of fate decisions connecting stem and progenitor cells through time. Directly linking whole-transcriptome descriptions of cells to their future fate is challenging, however, as single-cell RNA-seq technologies are destructive (see Section 2.1). Weinreb et al. (2020) therefore clonally tag cells with DNA barcodes that can be read using scRNA-seq and enable tracing cellular identities through time. The resulting dataset consists of three snapshots taken on days 2, 4, and 6, respectively. While at day 2 most cells are undifferentiated, at later time points cells have developed into neutrophils, monocytes, megakaryocytes, mast cells, lymphoid precursors, erythrocytes, basophils, eosinophils, etc.

DATA PREPROCESSING. The dataset used in Chapter 4 and Section 7.3 is available for download in the GEO database under accession number GSE140802. For data quality control and preprocessing, we follow the analysis of Lotfollahi et al. (2023), resulting in a dataset containing 130,861 cells. After processing, each observation records the level of expression of 1,622 different highly-variable genes as well as the following metainformation per cell:

- a `timestamp`, expressed in days and taking values in {2, 4, 6},
- a `barcode`, which is a short DNA sequence that allows tracing the identity of cells and their lineage by means of single-cell sequencing readouts,
- an additional `annotation`, which describes the current differentiation fate of the cell.

For experiments conducted in Section 7.3.3.2, we only retain cells with barcodes that appear both on days 2 and 4, taking care of excluding cells that are already differentiated on day 2. We construct matchings by pairing cells measured at two different times but which share the barcode. Additionally, we filter cells to make sure that no one appears in more than one pair.

FEATURE SELECTION. To reduce the dimensionality of these data points, we perform a PCA projection down to 50 components.

A.2.6 *Other Datasets*

Preprocessing for the lupus patients (Kang et al., 2018) and cross-species dataset (Hagai et al., 2018) were inherited from Rybakov et al. (2020) and Lotfollahi et al. (2019), and we would like to thank the authors for hosting this dataset. Lastly, the preprocessing of the glioblastoma patient dataset (Zhao et al., 2021) was adapted from Peidli et al. (2022). The preprocessed datasets are available for download[6].

A.3 EVALUATION METRICS

Since we lack access to the ground truth pair of perturbed and unperturbed observations on the single cell level, we consider evaluation metrics on the level of the distribution of real and predicted perturbation states to analyze the effectiveness of the approaches presented in this thesis.

AVERAGE- AND CORRELATION-BASED DISTANCE. Common evaluation metrics in single-cell biology rely on averages and correlation analysis. ℓ_2 feature means thereby refers to the ℓ_2-distance between means of the observed and predicted distributions. Similarly, r_2 feature means refers to the correlation of the means of the observed and predicted distributions. However, metrics based only on feature means can be insensitive in settings where crucial heterogeneity is not captured. Consider, for example, a target distribution with multiple modes. These metrics will favorably evaluate a uni-modal predicted distribution that simply models the mean of this multi-modal distribution. To this end, we include a distributional distance sensitive to this type of behavior by measuring differences in the properties of higher moments, i.e., the maximum mean discrepancy. We thus also report results based on several distributional metrics:

WASSERSTEIN DISTANCE. We measure accuracy of the predicted target population $\hat{\nu}$ to the observed target population ν using the entropy-regularized Wasserstein distance (Cuturi, 2013) provided in the `OTT` library (Cuturi et al., 2022) introduced in Section 3.1.2 and defined in (3.4).

MAXIMUM MEAN DISCREPANCY. Kernel maximum mean discrepancy (Gretton et al., 2012) is another metric to measure distances between distributions, i.e., for our purpose between the predicted target population $\hat{\nu}$ to the observed target population ν. Given two random variables x and y with distributions $\hat{\nu}$ and ν, and a kernel function ω, Gretton et al. (2012) define the squared MMD as:

$$\mathrm{MMD}(\hat{\nu}, \nu; \omega) = \mathbb{E}_{x,x'}[\omega(x, x')] + \mathbb{E}_{y,y'}[\omega(y, y')] - 2\mathbb{E}_{x,y}[\omega(x, y)].$$

We report an unbiased estimate of $\mathrm{MMD}(\hat{\nu}, \nu)$, in which the expectations are evaluated by averages over the population particles in each set. We utilize the RBF kernel, and as is usually done, report the MMD as an average over several length scales, i.e., `np.logspace(1, -3)`.

PERTURBATION SIGNATURES. A common method to quantify the effect of a perturbation on a population is to compute its perturbation signature (Stathias et al., 2018, (PS)), computed via the difference in means between the distribution of perturbed states and control states of each feature, e.g., here individual genes. ℓ_2(PS) then refers to the ℓ_2-distance between the perturbation signatures computed on the observed and predicted distributions, ν and $\hat{\nu}$. As before, let μ be the set of observed unperturbed population particles, ν the set of observed perturbed particles, as well as $\hat{\nu}$ the predicted perturbed state of population μ. The ℓ_2(PS) is then defined as

$$\mathrm{PS}(\nu, \mu) = \frac{1}{m} \sum_{y_i \in \nu} y_i - \frac{1}{n} \sum_{x_i \in \mu} x_i,$$

where n is the size of the unperturbed and m of the perturbed population. We report the ℓ_2 distance between the observed signature $\mathrm{PS}(\nu, \mu)$ and the predicted signature $\mathrm{PS}(\hat{\nu}, \mu)$, which is equivalent to simply computing the difference in the means between the observed and predicted distributions.

A.4 PROOF OF THEOREM 3

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

$$X_t \sim \mathbb{P}_t : \quad \mathrm{d}X_t = (f_t(X_t) + w_t(X_t))\,\mathrm{d}t + g_t\,\mathrm{d}\mathbb{W}_t. \tag{A.4.1}$$

The Fokker-Planck equation for the SDE (A.4.1) is

$$\partial_t \rho_t = -\nabla_x \cdot (\rho_t(f_t + w_t)) + \frac{g_t^2}{2}\Delta\rho_t. \tag{A.4.2}$$

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

$$D_{\mathrm{KL}}(\mathbb{P}_t \| Y_t) = \mathbb{E}\left[\int_0^1 \frac{\|w_t\|^2}{2g_t^2}\,\mathrm{d}t\right]. \tag{A.4.3}$$

Using a change of variable $v_t = w_t - \frac{g_t^2}{2}\nabla \log \rho_t$, we see that (A.4.2) is equivalent to

$$\partial_t \rho_t = -\nabla_x \cdot (\rho_t(f_t + v_t)). \tag{A.4.4}$$

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.4.3) becomes

$$\mathbb{E}\left[\int_0^1 \frac{\|v_t\|^2}{2g_t^2} + \frac{g_t^2}{8}\|\nabla \log \rho_t\|^2 + \frac{1}{2}\langle v_t, \nabla \log \rho_t\rangle\,\mathrm{d}t\right]. \tag{A.4.5}$$

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

$$\begin{aligned}
H(\rho_1) - H(\rho_0) &= \int_0^1 \partial_t H(\rho_t)\,\mathrm{d}t \\
&= \int_0^1 \int (1 + \log \rho_t)\partial_t \rho_t \mathrm{d}x\,\mathrm{d}t \\
&= \int_0^1 \int (1 + \log \rho_t) \cdot (-\nabla_x \cdot (\rho_t(f_t + v_t)))\mathrm{d}x\,\mathrm{d}t \qquad \text{by (A.4.2)} \\
&= \int_0^1 \int \rho_t \langle \nabla \log \rho_t, f_t + v_t\rangle \mathrm{d}x\,\mathrm{d}t
\end{aligned}$$

by integration by parts for the divergence operator. Therefore,

$$\mathbb{E}\left[\int_0^1 \langle \nabla \log \rho_t, v_t\rangle\,\mathrm{d}t\right] = H(\rho_1) - H(\rho_0) - \mathbb{E}\left[\int_0^1 \langle \nabla \log \rho_t, f_t\rangle\,\mathrm{d}t\right] \tag{A.4.6}$$

which concludes the proof. □

A.5 THE BURES-WASSERSTEIN GEOMETRY OF GAUSSIAN SBS

A.5.1 *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:

$$\forall\, U, V \in \mathcal{T}_\Sigma \mathbb{S}^d_{++}, \quad \langle U, V \rangle_\Sigma := \operatorname{tr} \mathcal{L}_\Sigma[U] \Sigma \mathcal{L}_\Sigma[V] = \frac{1}{2} \operatorname{tr} \mathcal{L}_\Sigma[U] V. \tag{A.5.1}$$

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 $F \colon \mathbb{S}^d_{++} \to \mathbb{R}$,

$$\mathcal{T}_\Sigma \mathbb{S}^d_{++} \ni \operatorname{grad} F(\Sigma) := 2\Big(\nabla F(\Sigma) \Sigma + \Sigma \nabla F(\Sigma)^T \Big) \tag{A.5.2}$$

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

$$\begin{aligned} \mathcal{L}_\Sigma[\operatorname{grad} F(\Sigma)] &= 2 \mathcal{L}_\Sigma[\nabla F(\Sigma) \Sigma + \Sigma \nabla F(\Sigma)] && \text{(A.5.3)} \\ &= 2 \nabla F(\Sigma) && \text{(A.5.4)} \end{aligned}$$

by definition of the Lyapunov operator. In other words,

$$\operatorname{grad} F(\Sigma) = \mathcal{L}_\Sigma^{-1}[2 \nabla F]. \tag{A.5.5}$$

Lastly, we recall the Bures-Wasserstein *acceleration* of a curve $\Sigma_t : [0, 1] \to \mathbb{S}^d_{++}$, which we denote by $\nabla_{\dot\Sigma_t} \dot\Sigma_t$:[8]

$$\nabla_{\dot\Sigma_t} \dot\Sigma_t = \ddot\Sigma_t - \Big(\mathcal{L}_{\Sigma_t}[\dot\Sigma_t] \dot\Sigma_t + \dot\Sigma_t \mathcal{L}_{\Sigma_t}[\dot\Sigma_t] \Big) + \Big(\Sigma_t \Big(\mathcal{L}_{\Sigma_t}[\dot\Sigma_t] \Big)^2 + \Big(\mathcal{L}_{\Sigma_t}[\dot\Sigma_t] \Big)^2 \Sigma_t \Big). \tag{A.5.6}$$

A.5.2 *Proof of Theorem 4*

For convenience, we restate Theorem 4 in full below:

Theorem 4. *The minimizer of* (7.19) *(and hence* (7.7)*) coincides with the solution of the action minimization problem:*

$$\min_{\Sigma_0 = \Sigma, \Sigma_1 = \Sigma'} \int_0^1 \frac{1}{2} \|\dot\Sigma_t\|^2_{\Sigma_t} - \mathcal{U}_\sigma(\Sigma_t)\, \mathrm{d}t \tag{7.20}$$

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

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* (7.19) *solves the Euler-Lagrange equation in the Bures-Wasserstein geometry:*

$$\nabla_{\dot\Sigma_t}\dot\Sigma_t = -\operatorname{grad}\mathcal{U}_\sigma(\Sigma_t), \quad \Sigma_0 = \Sigma, \quad \Sigma_1 = \Sigma', \tag{7.21}$$

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 (7.21) for the curve (7.22).

A.5.2.1 *Verifying the Euler-Lagrange Equation* (7.21)

We begin by noting that the boundary conditions in (7.21) hold for the curve in (7.22).
We now compute the two sides of (7.21) separately:

THE RIGHT-HAND SIDE OF (7.21): $-\operatorname{grad}\mathcal{U}_\sigma(\Sigma_t)$. 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.5.2) that the negative Bures-Wasserstein gradient of $\mathcal{U}_\sigma(\Sigma_t)$ is

$$\begin{aligned} -\operatorname{grad}\mathcal{U}_\sigma(\Sigma_t) &= -2\left(\frac{\sigma^4}{8}\Sigma_t^{-1}\cdot\Sigma_t^{-1}\cdot\Sigma_t + \Sigma_t\cdot\frac{\sigma^4}{8}\Sigma_t^{-1}\cdot\Sigma_t^{-1}\right) \\ &= -\frac{\sigma^4}{2}\Sigma_t^{-1}. \end{aligned} \tag{A.5.7}$$

THE LEFT-HAND SIDE OF (7.21): $\nabla_{\dot\Sigma_t}\dot\Sigma_t$. 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.5.1. *Define the matrix* $\tilde S_t$ *to be:*

$$\tilde S_t := t\Sigma\prime + \bar t C_\sigma - \bar t\Sigma - tC_\sigma^T + \frac{\sigma^2}{2}(\bar t - t)I. \tag{A.5.8}$$

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

$$A: \quad A\Sigma_t + \Sigma_t A = \dot\Sigma_t. \tag{A.5.9}$$

Moreover, $\tilde S_t$ *satisfies the following identity:*

$$\dot{\tilde S}_t - \Sigma_t^{-1}\tilde S_t^2 = -\frac{\sigma^4}{4}\Sigma_t^{-1}. \tag{A.5.10}$$

Before commencing the proof of Lemma A.5.1, let us show how it readily leads us to (7.21).
Recall the definition of $\nabla_{\Sigma_t}\dot{\Sigma}_t$ in (A.5.6). First, note that, by (7.22) and (A.5.8),

$$\frac{1}{2}\ddot{\Sigma}_t = \Sigma + \Sigma\prime - \left(C_\sigma + C_\sigma^T + \sigma^2 I\right) \tag{A.5.11}$$

$$= \dot{\tilde{S}}_t. \tag{A.5.12}$$

On the other hand, Lemma A.5.1 entails that

$$\begin{aligned}\Sigma_t\left(\mathcal{L}_{\Sigma_t}[\dot{\Sigma}_t]\right)^2 + \left(\mathcal{L}_{\Sigma_t}[\dot{\Sigma}_t]\right)^2\Sigma_t &= \Sigma_t\mathcal{L}_{\Sigma_t}[\dot{\Sigma}_t]\cdot\mathcal{L}_{\Sigma_t}[\dot{\Sigma}_t] + \mathcal{L}_{\Sigma_t}[\dot{\Sigma}_t]\cdot\mathcal{L}_{\Sigma_t}[\dot{\Sigma}_t]\Sigma_t \\ &= \Sigma_t\Sigma_t^{-1}\tilde{S}_t\cdot\mathcal{L}_{\Sigma_t}[\dot{\Sigma}_t] + \mathcal{L}_{\Sigma_t}[\dot{\Sigma}_t]\cdot\tilde{S}_t^T\Sigma_t^{-1}\Sigma_t \\ &= \tilde{S}_t\mathcal{L}_{\Sigma_t}[\dot{\Sigma}_t] + \mathcal{L}_{\Sigma_t}[\dot{\Sigma}_t]\tilde{S}_t^T. \end{aligned} \tag{A.5.13}$$

By noting, again from Lemma A.5.1,

$$\begin{aligned}\dot{\Sigma}_t &= \tilde{S}_t^T\Sigma_t^{-1}\cdot\Sigma_t + \Sigma_t\cdot\Sigma_t^{-1}\tilde{S}_t \\ &= \tilde{S}_t + \tilde{S}_t^T, \end{aligned} \tag{A.5.14}$$

we thus get

$$\begin{aligned}&\Sigma_t\left(\mathcal{L}_{\Sigma_t}[\dot{\Sigma}_t]\right)^2 + \left(\mathcal{L}_{\Sigma_t}[\dot{\Sigma}_t]\right)^2\Sigma_t - \left(\mathcal{L}_{\Sigma_t}[\dot{\Sigma}_t]\dot{\Sigma}_t + \dot{\Sigma}_t\mathcal{L}_{\Sigma_t}[\dot{\Sigma}_t]\right) \\ &\quad= \left(\tilde{S}_t - \dot{\Sigma}_t\right)\mathcal{L}_{\Sigma_t}[\dot{\Sigma}_t] + \mathcal{L}_{\Sigma_t}[\dot{\Sigma}_t]\left(\tilde{S}_t^T - \dot{\Sigma}_t\right) \\ &\quad= -\left(\tilde{S}_t^T\mathcal{L}_{\Sigma_t}[\dot{\Sigma}_t] + \mathcal{L}_{\Sigma_t}[\dot{\Sigma}_t]\tilde{S}_t\right) \end{aligned} \tag{A.5.15}$$

by (A.5.14). But $\tilde{S}_t^T\mathcal{L}_{\Sigma_t}[\dot{\Sigma}_t] = \tilde{S}_t^T\cdot\Sigma_t^{-1}\tilde{S}_t = \Sigma_t^{-1}\tilde{S}_t^2$ by symmetry of $\tilde{S}_t^T\Sigma_t^{-1}$ 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.5.6) reduces to

$$\nabla_{\dot{\Sigma}_t}\dot{\Sigma}_t = 2\dot{\tilde{S}}_t - 2\Sigma_t^{-1}\tilde{S}_t^2. \tag{A.5.16}$$

In lieu of (7.21), (A.5.7), and (A.5.16), the proof of (7.20) can thus be reduced to showing

$$2\dot{\tilde{S}}_t - 2\Sigma_t^{-1}\tilde{S}_t^2 = -\frac{\sigma^4}{2}\Sigma_t^{-1} \tag{A.5.17}$$

which is exactly (A.5.10).

Proof of Lemma A.5.1. We now prove Lemma A.5.1. 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 $C_\sigma \Sigma = \Sigma C_\sigma^T$. In addition, we have

$$\begin{aligned} C_\sigma^{-1}\Sigma &= 2\left(\Sigma^{\frac{1}{2}} D_\sigma \Sigma^{-\frac{1}{2}} - \sigma^2 I\right)^{-1}\Sigma \\ &= 2\left(\Sigma^{-\frac{1}{2}} D_\sigma \Sigma^{-\frac{1}{2}} - \sigma^2 \Sigma^{-1}\right)^{-1} \\ &= \Sigma C_\sigma^{-T}. \end{aligned} \tag{A.5.18}$$

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

$$C_\sigma^2 + \sigma^2 C_\sigma = \Sigma\Sigma\prime. \tag{A.5.19}$$

We therefore have

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

which, together with (A.5.18), implies

$$\begin{aligned} C_\sigma^T \Sigma\prime &= \Sigma\prime\Sigma C_\sigma^{-T}\Sigma\prime - \sigma^2\Sigma\prime \\ &= \Sigma\prime C_\sigma^{-1}\Sigma\Sigma\prime - \sigma^2\Sigma\prime \\ &= \Sigma\prime C_\sigma. \end{aligned} \tag{A.5.20}$$

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

$$P_t := t\Sigma\prime + \bar{t}C_\sigma, \quad Q_t := \bar{t}\Sigma + tC_\sigma. \tag{A.5.21}$$

Note that, by (A.5.20),

$$\begin{aligned} \Sigma\prime P_t^{-1} &= \left(P_t \Sigma\prime^{-1}\right)^{-1} \\ &= \left(tI + \bar{t}C_\sigma\Sigma\prime^{-1}\right)^{-1} \\ &= \left(tI + \bar{t}\Sigma\prime^{-1}C_\sigma^T\right)^{-1} \\ &= \left(\Sigma\prime^{-1}P_t^T\right)^{-1} \\ &= P_t^{-T}\Sigma\prime. \end{aligned} \tag{A.5.22}$$

A similar calculation leading to (A.5.22) shows

$$Q_t^{-1}\Sigma = \Sigma Q_t^{-T}. \tag{A.5.23}$$

PROOF OF SYMMETRY OF $\tilde{S}_t^T \Sigma_t^{-1}$. We get, by (A.5.19) and (A.5.20),

$$\begin{aligned} P_t^2 + \sigma^2 \bar{t} P_t &= t^2 \Sigma\prime^2 + \bar{t}^2 C_\sigma^2 + t\bar{t}(\Sigma\prime C_\sigma + C_\sigma \Sigma\prime) + \sigma^2 t\bar{t}\Sigma\prime + \sigma^2 \bar{t}^2 C_\sigma \\ &= t^2 \Sigma\prime^2 + \bar{t}^2 \Big(C_\sigma^2 + \sigma^2 C_\sigma \Big) + t\bar{t} \Big(C_\sigma^T \Sigma\prime + C_\sigma \Sigma\prime \Big) + \sigma^2 t\bar{t}\Sigma\prime \\ &= t^2 \Sigma\prime^2 + \bar{t}^2 \Sigma\Sigma\prime + t\bar{t} \Big(C_\sigma^T + C_\sigma + \sigma^2 I \Big) \Sigma\prime = \Sigma_t \Sigma\prime. \end{aligned} \tag{A.5.24}$$

It then follows from (A.5.24) that

$$P_t = \Sigma_t \Sigma\prime P_t^{-1} - \sigma^2 \bar{t} I, \tag{A.5.25}$$

$$P_t^T = P_t^{-T} \Sigma\prime \Sigma_t - \sigma^2 \bar{t} I. \tag{A.5.26}$$

As a result, we get, by (A.5.22) and (A.5.25)-(A.5.26),

$$\begin{aligned} \Sigma_t^{-1} P_t &= \Sigma\prime P_t^{-1} - \sigma^2 \bar{t} \Sigma_t^{-1} \\ &= P_t^{-T} \Sigma\prime - \sigma^2 \bar{t} \Sigma_t^{-1} \\ &= P_t^T \Sigma_t^{-1}. \end{aligned} \tag{A.5.27}$$

In exactly the same vein, we have

$$Q_t^2 + \sigma^2 t Q_t = \Sigma\Sigma_t \tag{A.5.28}$$

as well as

$$\Sigma_t^{-1} Q_t^T = Q_t \Sigma_t^{-1}. \tag{A.5.29}$$

The symmetry of $\tilde{S}_t^T \Sigma_t^{-1}$ is then an immediate consequence of (A.5.27) and (A.5.29). In addition, we have

$$\begin{aligned} \dot{\Sigma}_t &= 2t\Sigma\prime - 2\bar{t}\Sigma + (\bar{t} - t) \Big(C_\sigma + C_\sigma^T + \sigma^2 I \Big) \\ &= \tilde{S}_t + \tilde{S}_t^T. \end{aligned} \tag{A.5.30}$$

Combining the symmetry of $\tilde{S}_t^T \Sigma_t^{-1}$ and (A.5.30), we see that

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

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

PROOF OF (A.5.10). We next compute

$$\begin{aligned} P_t Q_t^T &= (t\Sigma\prime + \bar{t}C_\sigma)\left(\bar{t}\Sigma + tC_\sigma^T\right) \\ &= t\bar{t}\Sigma\prime\Sigma + t^2\Sigma\prime C_\sigma^T + \bar{t}^2 C_\sigma\Sigma + t\bar{t}C_\sigma C_\sigma^T \\ &= \bar{t}^2\Sigma C_\sigma^T + t^2\Sigma\prime C_\sigma^T + t\bar{t}\left(C_\sigma^{T2} + \sigma^2 C_\sigma^T\right) + t\bar{t}C_\sigma C_\sigma^T \\ &= \Sigma_t C_\sigma^T \end{aligned} \tag{A.5.31}$$

where we have used (A.5.18) in the third equality of (A.5.31). A similar computation further shows

$$Q_t^T P_t = \Sigma_t C_\sigma. \tag{A.5.32}$$

We thus get, by combining (A.5.24) (A.5.28)

$$\begin{aligned} \tilde{S}_t^2 &= P_t^2 - P_t Q_t^T + \frac{\sigma^2}{2}(\bar{t}-t)P_t - Q_t^T P_t + Q_t^{T2} - \frac{\sigma^2}{2}(\bar{t}-t)Q_t^T + \frac{\sigma^2}{2}(\bar{t}-t)P_t \\ &\quad - \frac{\sigma^2}{2}(\bar{t}-t)Q_t^T + \frac{\sigma^4}{4}(\bar{t}-t)^2 I \qquad \text{(A.5.33)} \\ &= P_t^2 + \sigma^2(\bar{t}-t)P_t + Q_t^{T2} - \sigma^2(\bar{t}-t)Q_t^T - \left(P_t Q_t^T + Q_t^T P_t\right) + \frac{\sigma^4}{4}(\bar{t}-t)^2 I \\ &= \Sigma_t\Sigma\prime - \sigma^2 t P_t + \Sigma_t\Sigma - \sigma^2\bar{t}Q_t^T - \left(\Sigma_t C_\sigma^T + \Sigma_t C_\sigma\right) + \frac{\sigma^4}{4}(\bar{t}-t)^2 I - \sigma^2\Sigma_t + \sigma^2\Sigma_t \\ &= \Sigma_t\left(\Sigma + \Sigma\prime - \left(C_\sigma + C_\sigma^T + \sigma^2 I\right)\right) + \sigma^2\left(\Sigma_t - tP_t - \bar{t}Q_t^T\right) + \frac{\sigma^4}{4}(\bar{t}-t)^2 I \\ &= \Sigma_t\dot{\tilde{S}}_t + \sigma^2\cdot t\bar{t}\sigma^2 I + \frac{\sigma^4}{4}(\bar{t}-t)^2 I \\ &= \Sigma_t\dot{\tilde{S}}_t + \frac{\sigma^4}{4} I \end{aligned} \tag{A.5.34}$$

where the third equality follows from (A.5.24), (A.5.28), and (A.5.31)-(A.5.32), and the fifth equality follows from (A.5.8). Multiplying both sides of (A.5.34) by Σ_t^{-1} from the right yields the desired (A.5.10). □

A.5.2.2 *Equivalence between (7.19) and (7.20)*

We first note that, by (A.5.1) and Lemma A.5.1,

$$\begin{aligned}\frac{1}{2}\|\dot{\Sigma}_t\|^2_{\Sigma_t} &= \frac{1}{2}\operatorname{tr}\mathcal{L}_{\Sigma_t}[\dot{\Sigma}_t]\Sigma_t\mathcal{L}_{\Sigma_t}[\dot{\Sigma}_t] \\ &= \frac{1}{2}\operatorname{tr}\tilde{S}_t^T\Sigma_t^{-1}\cdot\Sigma_t\cdot\Sigma_t^{-1}\tilde{S}_t \\ &= \frac{1}{2}\operatorname{tr}\tilde{S}_t^T\Sigma_t^{-1}\tilde{S}_t, \end{aligned} \tag{A.5.35}$$

and therefore the integrand in (7.20) is equal to

$$\operatorname{tr}\left(\frac{1}{2}\tilde{S}_t^T\Sigma_t^{-1}\tilde{S}_t + \frac{\sigma^4}{8}\Sigma_t^{-1}\right). \tag{A.5.36}$$

To proveed, we will need another formulation of (7.19), which is (Chen et al., 2016; Gentil et al., 2017) specialized to our case:

Lemma A.5.2. *Let $\mathcal{N}_0 := \mathcal{N}(0,\Sigma)$ and $\mathcal{N}_1 := \mathcal{N}(0,\Sigma\prime)$. Then (7.19) is equivalent to*

$$\min_{\rho_0=\mathcal{N}_0,\rho_1=\mathcal{N}_1}\int_0^1 \mathbb{E}\left[\frac{1}{2}\|\nabla\Phi_t\|^2 + \frac{\sigma^4}{8}\|\nabla\log\rho_t\|^2\right]\mathrm{d}t \tag{A.5.37}$$

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:

$$\partial_t\rho_t = -\nabla_x\cdot(\rho_t\nabla\Phi_t). \tag{A.5.38}$$

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

$$\frac{\mathrm{d}}{\mathrm{d}t}\det A(t) = \det A(t)\cdot\operatorname{tr}A^{-1}(t)\cdot\frac{\mathrm{d}}{\mathrm{d}t}A(t). \tag{A.5.39}$$

We are now ready to finish the proof of Theorem 4. By Léonard (2013), the optimal curve for (A.5.37) is Gaussian with zero mean. We denote by Σ_t the covariance of the solution at time t. By (A.5.39), we have

$$\begin{aligned}\partial_t \rho_t(x) &= \partial_t\left((2\pi)^{\frac{d}{2}}(\det\Sigma_t)^{-\frac{1}{2}}\exp\left(-\frac{1}{2}x^T\Sigma_t^{-1}x\right)\right)\\ &= (2\pi)^{\frac{d}{2}}\left(-\frac{1}{2}(\det\Sigma_t)^{-\frac{3}{2}}\right)\cdot\det\Sigma_t\cdot\operatorname{tr}\Sigma_t^{-1}\dot\Sigma_t\exp\left(-\frac{1}{2}x^T\Sigma_t^{-1}x\right)\\ &\qquad+(2\pi)^{\frac{d}{2}}(\det\Sigma_t)^{-\frac{1}{2}}\exp\left(-\frac{1}{2}x^T\Sigma_t^{-1}x\right)\cdot\left(\frac{1}{2}x^T\Sigma_t^{-1}\dot\Sigma_t\Sigma_t^{-1}x\right)\\ &= \rho_t(x)\cdot\left(\frac{1}{2}x^T\Sigma_t^{-1}\dot\Sigma_t\Sigma_t^{-1}x-\frac{1}{2}\operatorname{tr}\Sigma_t^{-1}\dot\Sigma_t\right).\end{aligned}\tag{A.5.40}$$

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

$$\nabla_x\cdot(\rho_t\nabla\Phi_t)=\langle\nabla\rho_t,\nabla\Phi_t\rangle+\rho_t\Delta\Phi_t.\tag{A.5.41}$$

Since $\nabla\rho_t=\rho_t\left(-\Sigma_t^{-1}x\right)$, the continuity equation (A.5.38) together with (A.5.40)-(A.5.41) implies that Σ_t must satisfy

$$\Delta\Phi_t=\frac{1}{2}\operatorname{tr}\Sigma_t^{-1}\dot\Sigma_t,\tag{A.5.42}$$

$$\langle\Sigma_t^{-1}x,\nabla\Phi_t(x)\rangle=\frac{1}{2}\langle\Sigma_t^{-1}x,\dot\Sigma_t\Sigma_t^{-1}x\rangle,\quad\forall x\in\mathbb{R}^d.\tag{A.5.43}$$

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

$$\operatorname{tr}\tilde S_t^T\Sigma_t^{-1}=\frac{1}{2}\operatorname{tr}\dot\Sigma_t\Sigma_t^{-1},\tag{A.5.44}$$

$$\operatorname{tr}\Sigma_t^{-1}\tilde S_t^T\Sigma_t^{-1}xx^T=\frac{1}{2}\operatorname{tr}\Sigma_t^{-1}\dot\Sigma_t\Sigma_t^{-1}xx^T,\quad\forall x\in\mathbb{R}^d.\tag{A.5.45}$$

Therefore, we see that

$$\begin{aligned}\mathbb{E}\left[\|\nabla\Phi_t\|^2\right]&=\mathbb{E}\left[\operatorname{tr}\tilde S_t^T\Sigma_t^{-1}xx^T\Sigma_t^{-1}\tilde S_t\right]\\ &=\operatorname{tr}\tilde S_t^T\Sigma_t^{-1}\,\mathbb{E}[xx^T]\Sigma_t^{-1}\tilde S_t\\ &=\operatorname{tr}\tilde S_t^T\Sigma_t^{-1}\tilde S_t.\end{aligned}\tag{A.5.46}$$

Furthermore, we have

$$\begin{aligned}\mathbb{E}\left[\|\nabla\log\rho_t\|^2\right]&=\mathbb{E}\left[\operatorname{tr}\Sigma_t^{-1}xx^T\Sigma_t^{-1}\right]\\ &=\operatorname{tr}\Sigma_t^{-1}.\end{aligned}\tag{A.5.47}$$

Finally, since the optimal vector field $\nabla\Phi_t$ is a gradient field, we must have $\tilde{S}_t^T\Sigma_t^{-1} = \Sigma_t^{-1}\tilde{S}_t$. Combing all the above, we see that (A.5.37) is equivalent to

$$\min_{\substack{\Sigma_0=\Sigma,\Sigma_1=\Sigma' \\ \tilde{S}_t^T\Sigma_t^{-1}=\Sigma_t^{-1}\tilde{S}_t}} \int_0^1 \operatorname{tr}\left(\frac{1}{2}\tilde{S}_t^T\Sigma_t^{-1}\tilde{S}_t + \frac{\sigma^4}{8}\Sigma_t^{-1}\right) \mathrm{d}t \tag{A.5.48}$$

which, in view of (A.5.36), is exactly the same as (7.20).

A.5.3 *Some Interesting Consequences of Theorem 4*

Here, we collect some interesting corollaries of Theorem 4, although they will not be used in the rest of the thesis.

A.5.3.1 *Conservation of Hamiltonian*

The first result concerns the *Hamiltonian formulation* of the action minimization problem (7.20).

Corollary 3 (Conservation of Hamiltonian)**.** *Define the* ***Hamiltonian*** *associated with* (7.20) *to be*

$$\begin{aligned}\mathcal{H}(\Sigma_t) &:= \frac{1}{2}\|\dot{\Sigma}_t\|_{\Sigma_t}^2 + \mathcal{U}_\sigma(\Sigma_t) \\ &= \operatorname{tr}\left(\frac{1}{2}\tilde{S}_t^T\Sigma_t^{-1}\tilde{S}_t - \frac{\sigma^4}{8}\Sigma_t^{-1}\right).\end{aligned} \tag{H}$$

Then the Hamiltonian is conserved along Σ_t:

$$\dot{\mathcal{H}} \equiv 0, \text{ or, equivalently, } \mathcal{H}(\Sigma_t) = \operatorname{tr}(\Sigma + \Sigma' - D_\sigma) \text{ for all } t. \tag{A.5.49}$$

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 4.

A.5.3.2 *Connection to Fisher Information*

The "potential energy" term $\mathcal{U}_\sigma(\Sigma_t)$ in (7.20) has an interesting origin: It is, up to a constant, the *entropy production rate*, i.e., the *Fisher information*.

Lemma A.5.3. *Let $\rho \sim \mathcal{N}(0, \Sigma)$, and let $H(\Sigma)$ be the (negative) Shannon entropy of ρ. Then*

$$\mathcal{U}_\sigma(\Sigma_t) = \frac{1}{2}\mathcal{I}_\sigma(\Sigma) \tag{A.5.50}$$

where

$$\mathcal{I}_\sigma(\Sigma) := \frac{\sigma^4}{4}\|\text{grad}\, H(\Sigma)\|_\Sigma^2. \tag{A.5.51}$$

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.5.1) and (A.5.5),

$$\begin{aligned}
\mathcal{I}_\sigma(\Sigma) &= \frac{\sigma^4}{4}\|\text{grad}\, H(\Sigma)\|_\Sigma^2 \\
&= \frac{\sigma^4}{4}\langle \text{grad}\, H(\Sigma), \text{grad}\, H(\Sigma)\rangle_\Sigma \\
&= \frac{\sigma^4}{4}\,\text{tr}\,\mathcal{L}_\Sigma[\text{grad}\, H(\Sigma)]\Sigma\mathcal{L}_\Sigma[\text{grad}\, H(\Sigma)] \\
&= \frac{\sigma^4}{4}\,\text{tr}\,\mathcal{L}_\Sigma[\mathcal{L}_\Sigma^{-1}[2\nabla H(\Sigma)]]\Sigma\mathcal{L}_\Sigma[\mathcal{L}_\Sigma^{-1}[2\nabla H(\Sigma)]] \\
&= \frac{\sigma^4}{4}(-\Sigma^{-1})\Sigma(-\Sigma^{-1}) = \frac{\sigma^4}{4}\Sigma^{-1}.
\end{aligned}$$

□

An infinite-dimensional version of Lemma A.5.3 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.5.51) has been shown to be intimately related to the *gradient flow* of entropy:

$$\dot{\Sigma}_t = -\,\text{grad}\, H(\Sigma). \tag{A.5.52}$$

We refer the interested readers to (Gentil et al., 2020) for details.

A.5.3.3 *Solution of the Schrödinger Systems*

For convenience, we restate the solutions of the Schrödinger system introduced in Section 3.2.5, i.e., Equations (3.25) and (3.26). Another way of solving a system of the form (A.5.37) is via the so-called forward *Schrödinger system* (Chen et al., 2021b; Léonard, 2013):

$$\begin{cases} \partial_t \mu_t + \nabla_x \cdot (\mu_t \nabla \Phi_t) = \frac{\sigma^2}{2}\Delta\mu_t \\ \partial_t \Phi_t + \frac{\|\nabla\Phi_t\|^2}{2} + \frac{\sigma^2}{2}\Delta\Phi_t = 0 \end{cases}. \tag{A.5.53}$$

By the various identities we prove in Appendix A.5.2.1, one can easily show that the solution to (A.5.53) is given by

$$\Phi_t(x) = -\frac{\sigma^2}{4}\log\det\Sigma_t + \frac{\sigma^4}{4}\int_0^t \operatorname{tr}\Sigma_t^{-1}\,\mathrm{d}t + \frac{1}{2}\langle x, \left(\tilde{S}_t^T - \frac{\sigma^2}{2}I\right)\Sigma_t^{-1}x\rangle + \text{const.} \tag{A.5.54}$$

This is in fact the same solution of the *fluid mechanical* problem

$$\min_{\substack{\rho_0=\mathcal{N}_0,\rho_1=\mathcal{N}_1\\ \partial_t\rho_t+\nabla_x\cdot(\rho_t\nabla\Phi_t)=\Delta\rho_t}} \int_0^1 \mathbb{E}\left[\frac{1}{2}\|\nabla\Phi_t\|^2\right]\mathrm{d}t \tag{A.5.55}$$

which is yet another equivalent formulation of (7.19).
There is also a backward Schrödinger system:

$$\begin{cases} -\partial_t\mu_t + \nabla_x\cdot\left(\mu_t\nabla\hat{\Phi}_t\right) = \frac{\sigma^2}{2}\Delta\mu_t \\ -\partial_t\hat{\Phi}_t + \frac{\|\nabla\hat{\Phi}_t\|^2}{2} + \frac{\sigma^2}{2}\Delta\hat{\Phi}_t = 0 \end{cases}, \tag{A.5.56}$$

whose solution is given by

$$\hat{\Phi}_t(x) = -\frac{\sigma^2}{4}\log\det\Sigma_t - \frac{\sigma^4}{4}\int_0^t \operatorname{tr}\Sigma_t^{-1}\,\mathrm{d}t - \frac{1}{2}\langle x, \left(\tilde{S}_t^T + \frac{\sigma^2}{2}I\right)\Sigma_t^{-1}x\rangle + \text{const.} \tag{A.5.57}$$

Notice that

$$\Phi_t + \hat{\Phi}_t = \sigma^2\log\rho_t \tag{A.5.58}$$

which is a well-known feature of the solutions to the forward and backward Schrödinger systems (Chen et al., 2021b; Léonard, 2013).

A.6 PROOF OF THE CLOSED-FORM SOLUTIONS FOR GAUSSIAN SBS

A.6.1 *Preliminaries for the Proof of Theorem 5*

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.6.1 (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):*

$$(2\pi)^{-\frac{d}{2}}(\det M)^{\frac{1}{2}} \int_{\mathbb{R}^d} v(x) \exp\left(-\frac{1}{2}x^T M x\right) \mathrm{d}x = \exp\left(\frac{1}{2}\partial_x^T M^{-1}\partial_x\right) v(x)\Big|_{x=0} \tag{A.6.1}$$

where $\exp\left(\frac{1}{2}\partial_x^T M^{-1}\partial_x\right)$ is understood as a power series in the differential operators.

Lastly, we recall the elementary

Lemma A.6.2 (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

$$\begin{aligned}\check{\mu} &= \mu_0 + \Sigma_{01}\Sigma_{11}^{-1}(y - \mu_1),\\ \check{\Sigma} &= \Sigma_{00} - \Sigma_{01}\Sigma_{11}^{-1}\Sigma_{10}.\end{aligned} \tag{A.6.2}$$

A.6.2 *The Proof*

We are now ready for the proof. For convenience, we restate Theorem 5 below:

Theorem 5. *Denote by $\mathbb{P}_t$ the solution to Gaussian Schrödinger bridges (GSB). Set*

$$
\begin{aligned}
&r_t := \frac{\kappa(t,1)}{\kappa(1,1)}, \quad \bar{r}_t := \tau_t - r_t\tau_1, \quad \sigma_\star := \sqrt{\tau_1^{-1}\kappa(1,1)},\\
&\zeta(t) := \tau_t \int_0^t \tau_s^{-1}\alpha_s \,\mathrm{d}s, \rho_t := \frac{\int_0^t \tau_s^{-2} g_s^2 \,\mathrm{d}s}{\int_0^1 \tau_s^{-2} g_s^2 \,\mathrm{d}s},\\
&P_t := \dot{r}_t(r_t\Sigma_1 + \bar{r}_t C_{\sigma_\star}), \quad Q_t := -\dot{\bar{r}}_t(\bar{r}_t\Sigma_0 + r_t C_{\sigma_\star}),\\
&S_t := P_t - Q_t^T + \Big[c_t\kappa(t,t)(1-\rho_t) - g_t^2\rho_t\Big] I. \qquad (7.28)
\end{aligned}
$$

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*

$$
\begin{aligned}
\mu_t &:= \bar{r}_t\mu_0 + r_t\mu_1 + \zeta(t) - r_t\zeta(1), \qquad (7.29)\\
\Sigma_t &:= \bar{r}_t^2\Sigma_0 + r_t^2\Sigma_1 + r_t\bar{r}_t\left(C_{\sigma_\star} + C_{\sigma_\star}^T\right) + \kappa(t,t)(1-\rho_t)I. \qquad (7.30)
\end{aligned}
$$

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

$$
\mathrm{d}X_t = f_{\mathcal{N}}(t, X_t)\,\mathrm{d}t + g_t\,\mathrm{d}\mathbb{W}_t \qquad (7.31)
$$

where

$$
f_{\mathcal{N}}(t,x) := S_t^T\Sigma_t^{-1}(x - \mu_t) + \dot{\mu}_t. \qquad (7.32)
$$

Moreover, the matrix $S_t^T\Sigma_t^{-1}$ 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 (7.6), and properties of (7.25) that $\mathbb{P}_t$ is a Markov Gaussian process with mean (7.29) and covariance (7.30).
3. Invoking the *generator theory* (Protter, 2005), to prove (7.31), it suffices to show that X_t satisfies, for any sufficiently regular test function $u : \mathbb{R}^+ \times \mathbb{R}^d \to \mathbb{R}$,

$$
\lim_{h\to 0} \frac{\mathbb{E}\left[u\left(t+h, X_{t+h}\right) \,\middle|\, X_t = x\right]}{h} = \mathcal{L}_t u(t,x), \qquad (A.6.3)
$$

where

$$\mathcal{L}_t u(t,x) := \frac{\partial}{\partial t} u(t,x) + \frac{g_t^2}{2} \Delta u(t,x) + \langle \nabla u(t,x), f_{\mathcal{N}}(t,x) \rangle \tag{A.6.4}$$

is the generator for the process (7.31).

4. Since the marginal/joint/conditional distributions of a Gaussian process are still Gaussian, the expectation in (A.6.3) 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.6.3) and (A.6.4). □

Proof of Theorem 5. From now on, we will invoke the notations in (7.28) without explicit mentions.

THE STATIC GAUSSIAN SB. We begin by solving the *static* Gaussian SB

$$\min_{\mathbb{P}_{01}} D_{\mathrm{KL}}(\mathbb{P}_{01} \| \mathbb{Q}_{01}) \tag{A.6.5}$$

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 (7.26) and covariance (7.27). Thus, if we only consider the endpoint marginal distributions (Y_0, Y_1), it is easy to derive the transition probability:

$$\begin{aligned} \mathbb{Q}(Y_1 = y_1 | Y_0 = y_0) &= (2\pi)^{\frac{d}{2}} \det(\kappa(1,1)I)^{-\frac{1}{2}} \exp(&& \text{(A.6.6)} \\ &\quad - \frac{1}{2}(y_1 - \eta(1))^T (\kappa(1,1)I)^{-1} (y_1 - \eta(1))) \\ &= (2\pi)^{\frac{d}{2}} \det(\kappa(1,1)I)^{-\frac{1}{2}} \exp(&& \text{(A.6.7)} \\ &\quad - \frac{1}{2\kappa(1,1)} \|y_1 - \tau_1 y_0 - \zeta(1)\|^2). \end{aligned}$$

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

$$D_{\mathrm{KL}}(\mathbb{P}_{01}\|\mathbb{Q}_{01}) = \int_{\mathbb{R}^d\times\mathbb{R}^d} \log \frac{\mathrm{d}\mathbb{P}_{01}}{\mathrm{d}\mathbb{Q}_{01}}\,\mathrm{d}\mathbb{P}_{01} \tag{A.6.8}$$

$$= \text{const.} + \frac{1}{2\kappa(1,1)} \int_{\mathbb{R}^d\times\mathbb{R}^d} \|y\prime - \tau_1 y - \tau_1\zeta(1)\|^2\,\mathrm{d}\mathbb{P}_{01}(y,y\prime) \tag{A.6.9}$$

$$+ \int_{\mathbb{R}^d\times\mathbb{R}^d} \log \mathbb{P}_{01}\,\mathrm{d}\mathbb{P}_{01}.$$

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

$$\tilde{\mu}_0 = \tau_1\mu_0 + \zeta(1), \tag{A.6.10}$$

$$\tilde{\Sigma}_0 = \tau_1^2\Sigma_0. \tag{A.6.11}$$

Obviously, there is a one-to-one correspondence between $\mathbb{P}_{01}$ and $\tilde{\mathbb{P}}_{01}$. The first integral in (A.6.9) 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.6.8) over $\mathbb{P}_{01}$ is equivalent to

$$\min_{\tilde{\mathbb{P}}_{01}} D_{\mathrm{KL}}(\tilde{\mathbb{P}}_{01}\|\mathbb{Q}_{01}) \equiv \min_{\tilde{\mathbb{P}}_{01}} \int_{\mathbb{R}^d\times\mathbb{R}^d} \frac{\|y-y\prime\|_2^2}{2}\,\mathrm{d}\tilde{\mathbb{P}}_{01}(y,y\prime) \tag{A.6.12}$$

$$+ \kappa(1,1)\int_{\mathbb{R}^d\times\mathbb{R}^d} \log \tilde{\mathbb{P}}_{01}\,\mathrm{d}\tilde{\mathbb{P}}_{01}. \tag{A.6.13}$$

By (7.6), the solution to (A.6.12) is given by the joint Gaussian

$$\tilde{\mathbb{P}}_{01}^{\star} \sim \mathcal{N}\left(\begin{bmatrix}\tilde{\mu}_0\\ \mu_1\end{bmatrix}, \begin{bmatrix}\tilde{\Sigma}_0 & \tilde{C}_{\tilde{\sigma}}\\ \tilde{C}_{\tilde{\sigma}}^T & \Sigma_1\end{bmatrix}\right) \tag{A.6.14}$$

where $\tilde{\sigma} = \sqrt{\kappa(1,1)}$ and

$$\tilde{C}_{\tilde{\sigma}} = \frac{1}{2}\left(\tilde{\Sigma}_0^{\frac{1}{2}}\tilde{D}_{\tilde{\sigma}}\tilde{\Sigma}_0^{-\frac{1}{2}} - \tilde{\sigma}^2 I\right), \tag{A.6.15}$$

$$\tilde{D}_{\tilde{\sigma}} = \left(4\tilde{\Sigma}_0^{\frac{1}{2}}\Sigma_1\tilde{\Sigma}_0^{\frac{1}{2}} + \tilde{\sigma}^4 I\right)^{\frac{1}{2}}. \tag{A.6.16}$$

The optimal static Gaussian SB $\mathbb{P}^\star_{01}$ is then given by the inverse transform $Y = \tau_1^{-1}(\tilde{Y} - \zeta(1))$, i.e.,

$$\mathbb{P}^\star_{01} \sim \mathcal{N}\left(\begin{bmatrix}\mu_0\\ \mu_1\end{bmatrix}, \begin{bmatrix}\Sigma_0 & \tau_1^{-1}\tilde{C}_{\tilde{\sigma}}\\ \tau_1^{-1}\tilde{C}_{\tilde{\sigma}}^T & \Sigma_1\end{bmatrix}\right). \tag{A.6.17}$$

Rearranging terms and using (A.6.15) and (A.6.16), we get

$$\tau_1^{-1}\tilde{C}_{\tilde{\sigma}} = C_{\sigma_\star} \tag{A.6.18}$$

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 (7.26) and (7.27), one can infer that, given Y_0, the joint distribution of (Y_t, Y_1) is

$$Y_t, Y_1 | Y_0 \sim \mathcal{N}\left(\begin{bmatrix}\eta(t)\\ \eta(1)\end{bmatrix}, \begin{bmatrix}\kappa(t,t)I & \kappa(t,1)I\\ \kappa(t,1)I & \kappa(1,1)I\end{bmatrix}\right). \tag{A.6.19}$$

Therefore, Lemma A.6.2 applied implies that, conditioned on Y_0 and Y_1, Y_t is Gaussian with mean

$$\begin{aligned}
\mathbb{E}[Y_t \,|\, Y_0, Y_1] &= \eta(t) + \frac{\kappa(t,1)}{\kappa(1,1)}(Y_1 - \eta(1))\\
&= \tau_t Y_0 + \zeta(t) + \frac{\kappa(t,1)}{\kappa(1,1)}(Y_1 - \tau_1 Y_0 - \zeta(1))\\
&= \left(\tau_t - \frac{\kappa(t,1)}{\kappa(1,1)}\tau_1\right)Y_0 + \frac{\kappa(t,1)}{\kappa(1,1)}Y_1 + \zeta(t) - \frac{\kappa(t,1)}{\kappa(1,1)}\zeta(1)\\
&= \bar{r}_t Y_0 + r_t Y_1 + \zeta(t) - \tau_t\zeta(1)
\end{aligned} \tag{A.6.20}$$

and covariance process (for any $t\prime \geq t$)

$$\begin{aligned}\mathbb{E}\Big[(Y_t - \mathbb{E}[Y_t \,|\, Y_0, Y_1])(Y_{t\prime} - \mathbb{E}[Y_{t\prime} \,|\, Y_0, Y_1])^T \,\Big|\, Y_0, Y_1\Big] \\ = \left(\kappa(t,t') - \frac{\kappa(t,1)\kappa(t\prime,1)}{\kappa(1,1)}\right) I.\end{aligned} \tag{A.6.21}$$

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.6.21),

$$Y_t | Y_0, Y_1 \overset{\text{law}}{=} \bar{r}_t Y_0 + r_t Y_1 + \zeta(t) - \tau_t \zeta(1) + \xi_t. \tag{A.6.22}$$

FROM $\mathbb{Q}$–BRIDGES TO μ_t AND Σ_t. The disintegration formula of $D_{\mathrm{KL}}(\cdot\|\cdot)$ (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.6.17), and then connecting X_0 and X_1 using the $\mathbb{Q}$–bridges (A.6.22). Namely,

$$X_t \overset{\text{law}}{=} \bar{r}_t X_0 + r_t X_1 + \zeta(t) - r_t \zeta(t) + \xi_t \tag{A.6.23}$$

from which (7.29) and (7.30) follow by a straightforward calculation. Furthermore, in view of (A.6.17) and (A.6.23), 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 5.

THE SDE REPRESENTATION OF X_t. The main idea of proving (7.32) is to compute

$$\lim_{h\to 0} \frac{\mathbb{E}\big[u\big(t+h, X_{t+h}\big) \,\big|\, X_t = x\big] - u(t,x)}{h} \tag{A.6.24}$$

and equate (A.6.24) with the *generator* of (7.31), which is (Protter, 2005)

$$\mathcal{L}_t u(t,x) := \frac{\partial}{\partial t} u(t,x) + \frac{g_t^2}{2} \Delta u(t,x) + \langle \nabla u(t,x), f_{\mathcal{N}}(t,x) \rangle. \tag{A.6.25}$$

Since X_t is a Gaussian process, we may derive the conditional expectation in (A.6.24) using Lemma A.6.2. 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 (7.30), 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

$$\begin{aligned}
\dot{\Sigma}_t &= 2\dot{\bar{r}}_t\bar{r}_t\Sigma_0 + 2\dot{r}_t r_t\Sigma_1 + (\dot{r}_t\bar{r}_t + r_t\dot{\bar{r}}_t)\Big(C_{\sigma_\star} + C_{\sigma_\star}^T\Big) \\
&\qquad + \Big(\frac{\partial}{\partial t}\kappa(t,t) - \dot{r}_t\kappa(t,1) - r_t\frac{\partial}{\partial t}\kappa(t,1)\Big)I \\
&= \dot{r}_t\Big(r_t\Sigma_1 + \bar{r}_t C_{\sigma_\star} + r_t\Sigma_1 + \bar{r}_t C_{\sigma_\star}^T\Big) + \dot{\bar{r}}_t\Big(\bar{r}_t\Sigma_0 + r_t C_{\sigma_\star} + \bar{r}_t\Sigma_0 + r_t C_{\sigma_\star}^T\Big) \\
&\qquad + \Big(\frac{\partial}{\partial t}\kappa(t,t) - 2\dot{r}_t\kappa(t,1)\Big)I \\
&= \Big(P_t + P_t^T\Big) - \Big(Q_t + Q_t^T\Big) + \Big(\frac{\partial}{\partial t}\kappa(t,t) - 2\dot{r}_t\kappa(t,1)\Big)I. \qquad \text{(A.6.26)}
\end{aligned}$$

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:

$$\begin{aligned}
K_{t,t+h} &:= \mathbb{E}\Big[(X_t - \mu_t)\big(X_{t+h} - \mu_{t+h}\big)^T\Big] \\
&= \bar{r}_t\bar{r}_{t+h}\Sigma_0 + r_t r_{t+h}\Sigma_1 + \bar{r}_t r_{t+h}C_{\sigma_\star} + r_t\bar{r}_{t+h}C_{\sigma_\star}^T + \Big(\kappa(t,t+h) \\
&\qquad - r_{t+h}\kappa(t,1)\Big)I \\
&= \Sigma_t + \bar{r}_t(\bar{r}_{t+h} - \bar{r}_t)\Sigma_0 + r_t(r_{t+h} - r_t)\Sigma_1 + \bar{r}_t(r_{t+h} - r_t)C_{\sigma_\star} \\
&\qquad + r_t(\bar{r}_{t+h} - \bar{r}_t)C_{\sigma_\star}^T + (\kappa(t,t+h) - \kappa(t,t) - r_{t+h}\kappa(t,1) + r_t\kappa(t,1))I \\
&= \Sigma_t + \frac{r_{t+h} - r_t}{\dot{r}_t}P_t - \frac{\bar{r}_{t+h} - \bar{r}_t}{\dot{\bar{r}}_t}Q_t^T + \Big(\kappa(t,t+h) \\
&\qquad - \kappa(t,t) - r_{t+h}\kappa(t,1) + r_t\kappa(t,1)\Big)I \\
&= \Sigma_t + h\Big\{P_t - Q_t^T + \Big[\Big(\frac{\partial}{\partial t\prime}\kappa\Big)(t,t) - \dot{r}_t\kappa(t,1)\Big]I\Big\} + o(h), \qquad \text{(A.6.27)}
\end{aligned}$$

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 (7.27) and $\dot{\tau}_t = c_t\tau_t$, we have

$$\begin{aligned}
\Big(\frac{\partial}{\partial t\prime}\kappa\Big)(t,t) &= \frac{\partial}{\partial t\prime}\Big(\tau_t\tau_{t\prime}\int_0^t \tau_s^{-2}g_s^2\,\mathrm{d}s\Big)\Big|_{t\prime=t} \qquad \text{(A.6.28)} \\
&= \dot{\tau}_t\tau_t\int_0^t \tau_s^{-2}g_s^2\,\mathrm{d}s \\
&= c_t\kappa(t,t). \qquad \text{(A.6.29)}
\end{aligned}$$

On the other hand, we have

$$\begin{aligned}\dot{r}_t &= \frac{1}{\kappa(1,1)}\frac{\partial}{\partial t}\left(\tau_t\tau_1\int_0^t \tau_s^{-2}g_s^2\,\mathrm{d}s\right)\\ &= \frac{1}{\kappa(1,1)}\left(c_t\kappa(t,1)+\tau_t^{-1}\tau_1 g_t^2\right)\\ &= c_t r_t + \frac{\tau_1 g_t^2}{\tau_t\kappa(1,1)}. \end{aligned} \tag{A.6.30}$$

Combining (A.6.29) and (A.6.30), 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.6.27) as

$$\begin{aligned}K_{t,t+h} &= \Sigma_t + h\Big\{P_t - Q_t^T + \Big[c_t\kappa(t,t)(1-\rho_t) - g_t^2\rho_t\Big]I\Big\} + o(h)\\ &= \Sigma_t + hS_t + o(h). \end{aligned} \tag{A.6.31}$$

We are now ready to derive (7.31). By Lemma A.6.2, 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.6.31),

$$\begin{aligned}\check{\mu}_{t+h} &= \mu_{t+h} + K_{t,t+h}^T\Sigma_t^{-1}(x-\mu_t)\\ &= \mu_t + h\dot{\mu}_t + \left(I + hS_t^T\Sigma_t^{-1}\right)(x-\mu_t) + o(h)\\ &= x + h\left(S_t^T\Sigma_t^{-1}(x-\mu_t) + \dot{\mu}_t\right) + o(h), \end{aligned} \tag{A.6.32}$$

and, by (A.6.26) and (A.6.27),

$$
\begin{aligned}
\check{\Sigma}_{t+h} &= \Sigma_{t+h} - K_{t,t+h}^T \Sigma_t^{-1} K_{t,t+h} \\
&= \Sigma_t + h\dot{\Sigma}_t - \left(\Sigma_t + hS_t^T + hS_t\right) + o(h) \\
&= h\Bigg[P_t + P_t^T - Q_t - Q_t^T + \left(\frac{\partial}{\partial t}\kappa(t,t) - 2\dot{r}_t\kappa(t,1)\right) I \\
&\qquad - \left(P_t - Q_t^T + \left[\left(\frac{\partial}{\partial t\prime}\kappa\right)(t,t) - \dot{r}_t\kappa(t,1)\right] I\right)^T \\
&\qquad - \left(P_t - Q_t^T + \left[\left(\frac{\partial}{\partial t\prime}\kappa\right)(t,t) - \dot{r}_t\kappa(t,1)\right] I\right)\Bigg] + o(h) \qquad \text{(A.6.33)} \\
&= h\left(\frac{\partial}{\partial t}\kappa(t,t) - 2\left(\frac{\partial}{\partial t\prime}\kappa\right)(t,t)\right) I + o(h). \qquad \text{(A.6.34)}
\end{aligned}
$$

However, by (7.27), we have

$$
\begin{aligned}
\frac{\partial}{\partial t}\kappa(t,t) &= \frac{\partial}{\partial t}\left(\tau_t^2 \int_0^t \tau_s^{-2} g_s^2 \, \mathrm{d}s\right) \\
&= 2\dot{\tau}_t \tau_t \int_0^t \tau_s^{-2} g_s^2 \, \mathrm{d}s + g_t^2, \\
\left(\frac{\partial}{\partial t\prime}\kappa\right)(t,t) &= \frac{\partial}{\partial t\prime}\left(\tau_t \tau_{t\prime} \int_0^t \tau_s^{-2} g_s^2 \, \mathrm{d}s\right)\Bigg|_{t\prime=t} \qquad \text{(A.6.35)} \\
&= \dot{\tau}_t \tau_t \int_0^t \tau_s^{-2} g_s^2 \, \mathrm{d}s, \qquad \text{(A.6.36)}
\end{aligned}
$$

from which (A.6.34) simplifies to

$$
\check{\Sigma}_{t+h} = h g_t^2 I + o(h). \qquad \text{(A.6.37)}
$$

We can now compute $\mathbb{E}\left[u\left(t+h, X_{t+h}\right) \mid X_t = x\right]$ as follows:

$$\mathbb{E}\left[u\left(t+h, X_{t+h}\right) \mid X_t = x\right] = (2\pi)^{\frac{d}{2}} \left(\det \check{\Sigma}_{t+h}\right)^{-\frac{1}{2}} \int_{\mathbb{R}^d} u\left(t+h, x'\right) \exp\left(-\frac{1}{2}\left(x'-\check{\mu}_{t+h}\right)^T \check{\Sigma}_{t+h}^{-1}\left(x'-\check{\mu}_{t+h}\right)\right) \mathrm{d}x' \tag{A.6.38}$$

$$= (2\pi)^{\frac{d}{2}} \left(\det \check{\Sigma}_{t+h}\right)^{-\frac{1}{2}} \int_{\mathbb{R}^d} u\left(t+h, x'+\check{\mu}_{t+h}\right) \exp\left(-\frac{1}{2} x'^T \check{\Sigma}_{t+h}^{-1} x'\right) \mathrm{d}x'. \tag{A.6.39}$$

Invoking Lemma A.6.1, we see that (A.6.39) can be evaluated as

$$\mathbb{E}\left[u\left(t+h, X_{t+h}\right) \mid X_t = x\right] = \exp\left(\frac{1}{2}\partial_{x'}^T \check{\Sigma}_{t+h} \partial_{x'}\right) u\left(t+h, x'+\check{\mu}_{t+h}\right)\Big|_{x'=0}. \tag{A.6.40}$$

Since $\check{\Sigma}_{t+h} = h g_t^2 I + o(h)$ by (A.6.37), expanding the power series $\exp\left(\frac{1}{2}\partial_{x'}^T \check{\Sigma}_{t+h} \partial_{x'}\right)$ and ignoring every $o(h)$ terms, (A.6.40) becomes

$$\mathbb{E}\left[u\left(t+h, X_{t+h}\right) \mid X_t = x\right] = \left(u\left(t+h, x'+\check{\mu}_{t+h}\right) + \frac{h g_t^2}{2} \Delta u\left(t+h, x'+\check{\mu}_{t+h}\right)\right)\Big|_{x'=0} + o(h)$$

$$= u(t+h, \check{\mu}_{t+h}) + \frac{h g_t^2}{2} \Delta u(t+h, \check{\mu}_{t+h}) + o(h). \tag{A.6.41}$$

Recalling from (A.6.32) that $\check{\mu}_{t+h} = x + h\left(S_t^T \Sigma_t^{-1}(x-\mu_t) + \dot{\mu}_t\right) + o(h)$, the Taylor expansion in the x variable for $u(t,x)$ shows that

$$\mathbb{E}\left[u\left(t+h, X_{t+h}\right) \mid X_t = x\right] = u(t+h, x) + h\Big(\frac{g_t^2}{2} \Delta u(t+h, x) + \left\langle \nabla u(t+h, x), S_t^T \Sigma_t^{-1}(x-\mu_t) + \dot{\mu}_t \right\rangle\Big) + o(h) \tag{A.6.42}$$

whence

$$\lim_{h\to 0} \frac{\mathbb{E}\left[u\left(t+h, X_{t+h}\right) \mid X_t = x\right] - u(t,x)}{h} = \frac{\partial}{\partial t} u(t,x) + \frac{g_t^2}{2} \Delta u(t,x) + \left\langle \nabla u(t,x), S_t^T \Sigma_t^{-1} (x - \mu_t) + \dot{\mu}_t \right\rangle. \tag{A.6.43}$$

This is exactly (A.6.25) with $f_{\mathcal{N}}(t,x) \leftarrow S_t^T \Sigma_t^{-1}(x-\mu_t) + \dot{\mu}_t$, which concludes the proof for (7.31) and (7.32).
Finally, by Léonard (2013, (4.2)), the optimal drift $f_{\mathcal{N}}(t,x)$ is a *gradient field*:

$$f_{\mathcal{N}}(t,x) = \nabla \psi(t,x) \tag{A.6.44}$$

for some function $\psi : \mathbb{R}^+ \times \mathbb{R}^d \to \mathbb{R}$, implying that $S_t^T \Sigma_t^{-1}$ must be symmetric. □