Data-Driven Mathematical Analysis with Applications in Dynamical Systems, Biology, and Social Justice

by

Rebecca Santorella

B.A., The College of New Jersey; Ewing, NJ 2017

M.Sc., Brown University; Providence, RI, 2018

A dissertation submitted in partial fulfillment of the
requirements for the degree of Doctor of Philosophy
in The Division of Applied Mathematics at Brown University

PROVIDENCE, RHODE ISLAND

May 2022

© Copyright 2022 by Rebecca Santorella

This dissertation by Rebecca Santorella is accepted in its present form
by The Division of Applied Mathematics as satisfying the
dissertation requirement for the degree of Doctor of Philosophy.

Date_____ _____

Björn Sandstede, Ph.D., Advisor

Recommended to the Graduate Council

Date_____ _____

Ritambhara Singh, Ph.D., Reader

Date_____ _____

Matthew Harrison, Ph.D., Reader

Approved by the Graduate Council

Date_____ _____

Andrew G. Campbell, Dean of the Graduate School

Vita

Education

Brown University, Providence, RI — May 2022
Ph.D. in Applied Mathematics
Adviser: Dr. Bjorn Sandstede

Brown University, Providence, RI — May 2018
Masters of Science, Applied Mathematics

The College of New Jersey, Ewing, NJ — May 2017
Bachelor of Arts in Mathematics — GPA: 3.95/4.0 — *Summa cum laude*

Research

Aug. 2017-*present* **Brown University**
Developed a new way to conduct manifold alignment using Gromov-Wasserstein optimal transport and applied it to biological data.

Extended previous research on using diffusion maps in equation-free modeling. Applied the new methods to conduct bifurcation analysis in an optimal velocity system.

May 2019-*present* **QSIDE Institute**
Compiled a large database of federal criminal convictions in the United States. Studied the sentencing disparities among judges.

Aug. 2015-May 2017 **The College of New Jersey**
Developed a multi-scale model of tumor growth. Used a cellular automaton framework to model the spatio-temporal evolution of cancer cells alongside a probabilistic signaling model within each cell.

Summer 2016 **Research Experience for Undergraduates at ICERM**

Investigated how diffusion maps could be used with equation-free modeling to develop numerical techniques for analyzing multi-scale dynamical systems. Applied the techniques to a well-studied traffic system and reproduced known results.

SUMMER 2015 **Research Experience for Undergraduates at Winthrop University**
Expanded upon an existing model of cancer treatment to account for the cancer stem cell hypothesis. Studied the effects of chemotherapy and periodic radiation on heterogeneous tumors.

Experience

MAY. 2021-*present* **Computational Healthcare Intern, United Health Group**
DUsed SVMs to analyze continuous glucose monitor data from thousands of type 2 diabetics to understand what information can be extracted from these digital signals.

Papers

P. Demetci, **R. Santorella**, B. Sandstede, and R. Singh. "Unsupervised integration of single-cell multi-omics datasets with disproportionate cell type representation." RECOMB, to appear (2022). (preprint)

Chin, T., Ruth, J., Sanford, C., **Santorella, R.** Carter, P., and , Sandstede, B. "Enabling Equation-Free Modeling via Diffusion Maps." JDDE, to appear (2022).

P. Demetci, **R. Santorella**, B. Sandstede, W. Stafford Noble, and R. Singh. "Gromov-Wasserstein optimal transport to align single-cell multi-omics data." RECOMB, to appear (2021).

Ciocanel, M. V., Topaz, C. M., **Santorella, R.**, Sen, S., Smith, C. M., & Hufstetler, A. (2020). *JUSTFAIR: Judicial System Transparency through Federal Archive Inferred Records.* PLOS ONE, 15(10).

Papers in Preparation

Kwegyir-Aggrey, K., **Santorella, R.**, and Brown, S.M. "Everything is Relative: Auditing with Optimal Transport."

C.M. Smith, N. Goldrosen, M.V. Ciocanel, **R. Santorella**, C.M. Topaz, S. Sen. "Racial Disparities in Criminal Sentencing Vary Considerably across Federal Judges."

Teaching Experience

Summer 2020	**Mentor** at EDGE Mentored 14 female incoming graduate students in mathematics. Helped with coursework and facilitated discussions about navigating graduate school.	Providence, RI
Fall 2019	**Mentor** for the Directed Reading Program Supervised an undergraduate who self-studied epidemiology models and their application to measles. Introduced them to numerical simulation and parameter fitting.	Providence, RI
Summer 2019	**Teaching Assistant** at GirlsGetMath@ICERM Introduced high school girls to data science through lectures and computing labs, led a hands-on topology activity, and aided other instructors in their lessons.	Providence, RI
Fall 2018 and Spring 2019	**Teaching Assistant** at Brown University Prepared and held recitations, graded homework and exams, managed seven undergraduate teaching assistants, held office hours, and provided detailed solutions for APMA 1650: Statistical Inference I.	Providence, RI
Aug. 2015-May 2017	**Tutor** at The College of New Jersey Provided group drop-in tutoring for mathematics courses. Proctored exams and substituted for introductory classes. Graded quizzes and homework, held review sessions, and created Mathematica assignments for Calculus B and Real Analysis.	Ewing, NJ

Honors and Awards

NSF Graduate Research Fellow	Aug. 2017
Brown University Presidential Fellow	Aug. 2017
Stella Dafermos Award, Brown University	Mar. 2021
Best Poster Award, ICML	July 2020
Brown Student Chapter Certificate of Recognition	April 2020
Red Sock Award for Best Poster Presentation, SIAM DS19	May 2019
Wendell B. Secor Award for Senior Achievement in Mathematics	May 2017
Robert S. Duncan Award for Junior Achievement in Mathematics	May 2016
Barry M. Goldwater Scholar	Apr. 2016

Presentations

Demetci P., **Santorella R.**, Sandstede B., and Singh, R.. "Unsupervised integration of single-cell multi-omics datasets with disparities in cell-type representation." **MLCB 2021**, November 22, 2021.

Demetci P., **Santorella R.**, Sandstede B., Noble W., and Singh, R.. "Gromov-Wasserstein optimal transport to align single-cell multi-omics data." **RECOMB 2021**, August 30, 2021.

M.V. Ciocanel, C.M. Topaz, **R. Santorella**, S. Sen, C.M. Smith, A. Hufstetler. "JUSTFAIR: Judicial System Transparency through Federal Archive Inferred Records," **MAA-NJ Section Meeting**, April 17, 2021.

Jiang, S., and Santorella R. "Localizing the COMAP Competitions" **Joint Math Meetings**, January 7, 2021.

Demetci P., Santorella R., Sandstede B., Noble W., and Singh, R.. "Gromov-Wasserstein optimal transport to align single-cell multi-omics data." **CASCODA Workshop at ICCABS** , December 10, 2020.

Demetci P., Santorella R., Sandstede B., Noble W., and Singh, R.. "Gromov-Wasserstein optimal transport to align single-cell multi-omics data." **Machine Learning in Computational Biology**, November, 2020.

Demetci P., Santorella R., Sandstede B., Noble W., and Singh, R.. "Gromov-Wasserstein optimal transport to align single-cell multi-omics data." **ICML Workshop on Computational Biology** (ICML WCB), July 2020.

Demetci P., Santorella R., Sandstede B., Noble W., and Singh, R.. "Gromov-Wasserstein optimal transport to align single-cell multi-omics data." **ISMB Machine Learning in Computational and Systems Biology Workshop** (ISMB MLCSB), July 2020.

Carter, P., Chin, T., Ruth, J., Sandstede, B., Sanford, C., and Santorella, R. *Using Diffusion Maps in Equation-Free Modeling* **SIAM Conference on Applications of Dynamical Systems**, Snowbird, UT, Nov, 2019.

Carter, P., Chin, T., Ruth, J., Sandstede, B., Sanford, C., and Santorella, R. *Using Diffusion Maps in Equation-Free Modeling* **Joint Dynamics and PDE Seminar**, Amherst, MA, Nov, 2018.

Gevertz, J., Ochs, M., and Santorella, R. *Multiscale Modeling of Macroscopic-Level Tumor Response to Stochastic Intracellular Signaling* **The College of New Jersey Honors Thesis Presentations**, Ewing, NJ, April, 2017.

Gevertz, J., Ochs, M. and Santorella, R. *Multiscale Modeling of Tumor Response to Stochastic Cell Signaling* **Celebration of Student Achievement**, Ewing, NJ, May, 2016.

Abernathy, Z., Bates, S. and Santorella, R. *A Mathematical Model of Cancer Stem Cell Driven Tumor Growth with Radiation and Chemotherapy Treatment.* **Joint Mathematics Meetings**, Seattle WA, January, 2016.

Outreach and Service

Presenter, Brown University Math Circle	Sept. 2019 - *present*
Organizer, Rose Whelan Society	Sept. 2017 - *present*
Vice President, Brown SIAM Student Chapter	Sept. 2017 - *present*
Mentor, Brown Graduate-Undergraduate Mentoring	Sept. 2017 - *present*
BMCM Organizer, Brown SIAM Student Chapter	Sept. 2017 - *present*
Mentor, Brown Directed Reading Program	Fall 2019

Professional Organizations

Institute for the Quantitative Study of Inclusion, Diversity, and Equity (QSIDE)
Society for Industrial and Applied Mathematics
Association for Women in Mathematics
Phi Beta Kappa

Graduate Coursework

Computational Probability and Statistics
Deep Learning
Functional Analysis
Nonlinear Dynamical Systems
Numerical Solutions for PDEs
Parallel Computing on GPU and CPU Systems
Probability Theory

Skills

Languages:	C++, CUDA, Java, Matlab, MPI, OpenMP, Python, R, Tensorflow, SQL
Methods:	Leep learning, graphical models, hierarchical clustering, k-means clustering, k-nearest neighbors, logistic regression, Monte Carlo methods, optimal transport, support vector machine
Tools:	computing clusters, git, Jupyter, LaTeX, linux, numpy, scikit-learn, pandas, AWS

Dedication

In memory of Sharon Ling and Lindsay Nigro, who shared so many of my experiences as a developing mathematician at the College of New Jersey.

Preface and Acknowledgments

First and foremost, I would like to thank my advisor, Dr. Björn Sandstede for his support over the past five years. I am very grateful to have had an adviser who values his students as people as well as researchers. Second, I would like to thank Dr. Ritambhara Singh for her incredible mentoring, particularly as an early-career faculty member during a pandemic. I would further like to thank Dr. Matthew Harrison for serving on my thesis committee.

Beyond my committee members, I would especially like to thank Dr. Sarah Brown, Dr. Chad Topaz, and Dr. Jana Gevertz for offering me advice and mentorship about mathematics and beyond. Additionally, I would like to thank all of my wonderful collaborators: Pinar Demetci, Kweku Kwegyir-Aggrey, Tracey Chin, Jacob Ruth, Clayton Sanford, Nicholas Goldrosen, Adam Hufstetler, Dr. Paul Carter, Dr. Maria-Veronica Ciocanel, Dr. Shilad Sen, Dr. Christian Michael Smith, and Dr. William Stafford Noble.

Next, I have to thank all of the fantastic graduate students I interacted with at Brown. Without the support of my cohort, my first year would have been even more challenging. I have received an incredible amount of support from the older graduate students in the Sandstede group. I would especially like to thank Karen Larson, Stephanie Dodson, and Ross Parker for answering an unreasonable amount of questions and constantly reassuring me that everything would work out.

Outside of academia, I would like to thank my friends and family for their support, especially the Rhode Island Pokémon Go community and my local Providence friends. Finally, thank you to my cat Daphne for constantly stepping on my keyboard while I was trying to type this thesis.

This thesis is based upon work supported by the National Science Foundation Graduate Research Fellowship under Grant No. 1644760.

Contents

List of Tables

List of Figures

Chapter One

Introduction

1.1 Data-Driven Mathematical Analysis

This work explores how to use data-driven methods in various settings, from conducting traditional analysis in dynamical systems to exposing disparities in federal sentencing. First, we use diffusion maps and equation-free modeling to conduct bifurcation analysis of an optimal velocity system. Second, we construct a new dataset of federal criminal sentencing data that includes demographics on the judge and defendant, critical case details, and sentencing outcomes. Next, we use optimal transport for auditing automated systems. Finally, we use Gromov-Wasserstein optimal transport to align single-cell multi-omics data. We give a brief introduction to each of these projects below and discuss them in full detail in Chapters 2-7.

1.2 Enabling Equation-Free Modeling via Diffusion Maps

Equation-free modeling aims at extracting low-dimensional macroscopic dynamics from complex high-dimensional systems that govern the evolution of microscopic states. This algorithm relies on lifting and restriction operators that map macroscopic states to microscopic states and vice versa Kevrekidis et al. (2003, 2004), Kevrekidis and Samaey (2009). Combined with simulations of the microscopic state, this algorithm can be used to apply Newton solvers to the implicitly defined low-dimensional macroscopic system or solve it more efficiently using direct numerical simulations. The key challenge is the construction of the lifting and restrictions operators that usually require a priori insight into the underlying application.

In this paper, we design an application-independent algorithm that uses diffusion maps to construct these operators from simulation data. Specifically, we use the diffusion map to select the macroscopic variables used in equation-free modeling and then define lifting and restriction operators through the diffusion map embedding. To demonstrate this framework, we apply it to an optimal velocity traffic model of cars driving around a ring road. We are able to conduct bifurcation analysis of traffic jams entirely in the macroscopic variables under two settings. First, we factor out translational symmetry due to the cars driving around the ring road and construct a one-dimensional diffusion map embedding. Second, we keep the symmetry and also solve for the periods of the traffic jam around the road. In both cases we are able to accurately reproduce bifurcation diagrams.

1.3 Judicial System Transparency through Federal Archive Inferred Records

In the United States, the public has a constitutional right to access criminal trial proceedings Ardia (2016). In practice, it can be difficult or impossible for the public to exercise this right. Furthermore, accessing detailed case information after sentencing can be nearly impossible. While sentencing disparity is an active area of research, most results focus on aggregate biases in the system due to the lack of data identifying judges. These works find biases in increased sentence length against male, low-income, and Black individuals Mustard (2001), Rehavi and Starr (2014), Yang (2015), Starr (2015). Additional studies use private databases to assess judge characteristics such as political affiliation and their interactions with sentencing Schanzenbach and Tiller (2008), Yang (2014, 2015), Cohen and Yang (2019). Although some

judge-focused research exists, it all relies on private data agreements that prohibit the naming of individual judges. Publicly released datasets such as those released by the United States Sentencing Commission purposely omit the identity of the sentencing judge Schanzenbach and Tiller (2008).

To make a connection between sentencing information and judge identity, we present JUSTFAIR: Judicial System Transparency through Federal Archive Inferred Records, a database of criminal sentencing decisions made in federal district courts. We have compiled this data set from public sources, including the United States Sentencing Commission, the Federal Judicial Center, the Public Access to Court Electronic Records system, and Wikipedia. With nearly 600,000 records from 2001—2018, JUSTFAIR is the first large-scale, free, public database that links information about defendants and their demographic characteristics with information about their federal crimes, their sentences, and, crucially, the identity of the sentencing judge.

Next, we use this database to analyze disparities in federal criminal sentencing. Studying a subset of 380,000 criminal cases in federal district courts from 2006 to 2019, we replicate previous findings that aggregate, conditional racial disparities in sentence lengths are significant. We further show that estimated racial disparities in sentencing vary considerably across judges. Results suggest that judges assign white defendants sentences that are conditionally 13% shorter than Black defendants' and 19% shorter than Hispanic defendants', on average. A judge who is one standard deviation above average in terms of estimated Black-white disparity gives Black defendants sentences that are conditionally 39% longer than white defendants', compared to the average disparity of 13%. A judge who is one standard deviation above average in terms of estimated Hispanic-white disparity gives Hispanic defendants sentences that are conditionally 49% longer than white defendants', compared

to the average disparity of 19%.

1.4 Optimal Transport

Finding the most efficient way to move something from one point to another is a common problem. Optimal transport generalizes this concept to moving a group of objects from one configuration to another. Specifically, it seeks a map that transforms one probability distribution into another through the least amount of work Peyré et al. (2019). With the development of fast solvers, optimal transport has increasingly been used in applications in a broad range of fields, including computational biology and fairness Xue et al. (2020), Taskesen et al. (2020), Dwork et al. (2012), Silvia et al. (2020), Feldman et al. (2015), Zehlike et al. (2020), Gordaliza et al. (2019), Jiang et al. (2019), Feldman et al. (2015), Johndrow et al. (2019), Chiappa and Pacchiano (2021), Alvarez-Melis and Jaakkola (2018), Schiebinger et al. (2019), Cang and Nie (2020), Yang et al. (2018), Yang and Uhler (2019). The field of optimal transport is extremely active, with many variations and applications introduced recently. In this work, we will focus on Kantorovich optimal transport, which allows mass to be split during transport. Figure 1.1 shows an example transport map between two distributions.

1.4.1 Auditing with Optimal Transport

Our first application of optimal transport uses it to audit automated decision-making systems. To study discrimination in automated decision-making systems, we need auditing tools that can detect and explain the structure of unfairness. Current

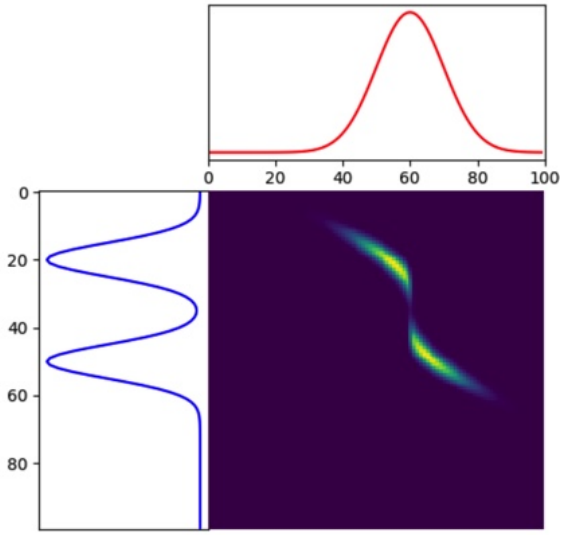

Figure 1.1: Optimal transport seeks to transform one distribution into another through the least amount of work. Shown are two distributions and the map to transform them into each other found by optimal transport.

methods depend on rigid definitions of fairness that measure group-wise disparities in performance Angwin et al. (2016), Buolamwini and Gebru (2018), Obermeyer et al. (2019). While these audits can reveal some disparities, they cannot capture every possible scenario such as intersections of identity and notions of fairness that contradict each other Raji et al. (2020), Abebe et al. (2020). Furthermore, many metrics are difficult to compute in practice Hu and Chen (2020).

To address these issues, we present an optimal transport-based approach to auditing that offers an interpretable and quantifiable exploration of bias and its structure. While many methods have used optimal transport to de-bias training data to train classifiers, we propose a new auditing framework that compares a pair of outcome distributions from classifiers to one another. By comparing the outcome distributions rather than the feature space, we can understand how individuals are treated differently, even from different populations. This work uses the optimal transport map to examine disparate treatment at the individual, subgroup, and group lev-

els. Our framework recovers well-known examples of algorithmic discrimination and explores the structure of bias.

1.4.2 Single-Cell Multi-Omics Alignment Using Optimal Transport

Our second use of optimal transport is aligning single-cell multi-omics datasets. Multi-omics datasets describe a biological object from different modes to gain different perspectives on it. Single-cell multi-omics datasets contain different molecular views of a cell, such as DNA methylation and chromatin accessibility. Integrated analysis of multi-omics data allows the study of how different molecular views in the genome interact to regulate cellular processes; however, with a few exceptions, applying multiple sequencing assays on the same single cell is not possible. Though such data integration of single-cell measurements is critical for understanding cell development and disease, the lack of correspondence between different measurements makes such efforts challenging. Several unsupervised algorithms can align heterogeneous single-cell measurements in a shared space, enabling the creation of mappings between single cells in separate data domains Liu et al. (2019), Cao et al. (2020, 2021). However, these algorithms require hyperparameter tuning for high-quality alignments, which is difficult in an unsupervised setting without correspondence information for validation.

We present Single-Cell alignment using Optimal Transport (SCOT), an unsupervised learning algorithm that uses Gromov Wasserstein-based optimal transport to align single-cell multi-omics datasets. We compare the alignment performance of SCOT with state-of-the-art algorithms on four simulated and two real-world datasets.

SCOT performs on par with state-of-the-art methods but is faster and requires tuning fewer hyperparameters. Furthermore, we provide an algorithm for SCOT to use Gromov Wasserstein distance to guide the parameter selection. Thus, unlike previous methods, SCOT aligns well without using any orthogonal correspondence information to pick the hyperparameters.

Our first iteration of SCOT performs alignment extremely well on balanced datasets, meaning datasets where the proportions of cell types are similar across modes. In particular, SCOT and other recent unsupervised algorithms have been primarily benchmarked on co-assay experiments rather than the more common single-cell experiments taken from separately sampled cell populations. Therefore, most existing methods, including SCOT, perform subpar alignments on such datasets. Thus, we also present an updated version of SCOT that uses unbalanced optimal transport to handle disproportionate cell-type representation and differing sample sizes across single-cell measurements. We show that our proposed method, SCOTv2, consistently yields quality alignments on five real-world single-cell datasets with varying cell-type proportions and is computationally tractable. Additionally, we extend SCOTv2 to integrate multiple ($M > 2$) single-cell measurements and present a self-tuning heuristic process to select hyperparameters in the absence of any orthogonal correspondence information.

1.5 Outline and Overview

The following chapters are mostly taken with minor edits and additions from the cited paper at the head of each chapter. In Chapter 2, we propose a method for equation-free modeling with diffusion maps and apply it to data generated from an optimal

velocity system to conduct bifurcation analysis. This work was a collaboration with Dr. Björn Sandstede, Tracy Chin, Jacob Ruth, Clayton Sanford, and Dr. Paul Carter and has been accepted for publication Chin et al. (2022). I helped develop the theoretical framework, wrote most of the code, produced all of the simulations, and wrote most of the paper. In Chapter 3, we describe how we web-scraped and merged databases together to create a new federal criminal sentencing database, JUSTFAIR. In Chapter 4, we conduct analyses on the JUSTFAIR dataset to show that Black and Hispanic defendants receive at least double the sentence lengths of white defendants. This work was done with Dr. Maria-Veronica Ciocanel, Dr. Chad M. Topaz, Dr. Shilad Sen, Dr. Christian Michael Smith, Nicholas Goldrosen, and Adam Hufstetler. The data paper is published in Ciocanel et al. (2020a), and the analysis is a preprint Smith et al. (2021). I did most of the initial merging of the databases as well as the updated database, helped write the paper, and contributed to the theoretical analysis.

Chapter 5 considers another interrogation of social justice through data analysis by using optimal transport to conduct audits on automated systems. This work was done in collaboration with Kweku Kwegyir-Aggrey and Dr. Sarah M. Brown and is a preprint Kwegyir-Aggrey et al. (2021). I developed the theoretical framework, ran simulations, and wrote much of the paper. Chapter 6 proposes another use of optimal transport, specifically Gromov-Wasserstein optimal transport, to integrate and align single-cell multi-omics data. We call this method SCOT (Single Cell alignment using Optimal Transport). In Chapter 7, we present an updated version of this algorithm to deal with unbalanced datasets and more than two domains. This work was done in collaboration with Pinar Demetci, Dr. Björn Sandstede, Dr. William Stafford Noble, and Dr. Ritambhara Singh. The original method is published in Demetci et al. (2021), and the second iteration has been accepted for publication Demetci

et al. (2022). I developed the theoretical framework, ran simulations, and wrote much of the paper. Finally, Chapter 8 discusses concluding remarks and possible avenues for future work.

Chapter Two

Enabling Equation-Free Modeling via Diffusion Maps

This chapter is based on work presented in Chin et al. (2022).

2.1 Introduction

In many complex dynamical systems, low-dimensional macroscopic behavior emerges from interactions at the high-dimensional microscopic level. For instance, traffic jams are global macroscopic structures that emerge from the interactions of many individual cars that move along a road: traffic jams can be captured meaningfully by a single macroscopic quantity, namely the standard deviation of the distances between consecutive cars from the mean (see below for further details). Mathematically, these macroscopic structures live on low-dimensional invariant manifolds, and our goal is to exploit their existence, and the accompanying reduction in effective dimension, when we conduct bifurcation analyses or carry out direct simulations. In most cases, these invariant manifolds, or the reduced vector field that governs the evolution of the macroscopic variables that parametrize these manifolds, are not known. Equation-free modeling, so named because the macroscopic system is not governed by an explicit ordinary differential equation, estimates macroscopic behavior through a multi-scale approach that exploits the connection between the macro and microlevels Kevrekidis et al. (2004), Kevrekidis and Samaey (2009), Kevrekidis et al. (2003). All equation-free methods depend on the following algorithm that attempts to describe the macroscopic system implicitly Kevrekidis and Samaey (2009):

(1) lift: build the microstate from the macrostate using the lifting operator $\mathcal{L}$;

(2) evolve: simulate the microstate for short bursts using the evolution Φ_t; and

(3) restrict: calculate the macrostate from the evolved microstate using the restric-

tion operator $\mathcal{R}$.

One of the biggest challenges in equation-free modeling is the selection of lifting and restriction operators since the choice of macroscopic observables may not always be obvious Marschler et al. (2014b). One way to pick relevant macroscopic variables is to use a dimension-reduction technique such as diffusion maps. Diffusion maps embed high dimensional data sets into low-dimensional Euclidean spaces. Unlike standard linear methods such as principal component analysis, diffusion maps are able to find useful low-dimensional parametrizations of the original data sets even when the data lie on or near a nonlinear manifold Coifman and Lafon (2006), Coifman et al. (2005), Lafon (2004). In this paper, we show that low-dimensional embeddings given through diffusion maps can be used to identify macroscopic variables in equation-free modeling and define lifting and restriction operators. We note though that this approach will not necessarily provide a physical interpretation of the resulting parametrization.

To illustrate our algorithm and demonstrate its effectiveness, we apply it to the same traffic model Bando et al. (1995) to which traditional equation-free modeling had been applied in earlier work Marschler et al. (2014a). In this model, N cars drive around a ring road of length L. We assume that all drivers follow the same deterministic behavior governed by

$$\tau \frac{d^2 x_n}{dt^2} + \frac{dx_n}{dt} = V(x_{n+1} - x_n), \quad x_n \in \mathbb{R}/L\mathbb{Z}, \quad n = 1, 2, ...N, \tag{2.1.1}$$

where x_n is the position of the nth car, τ reflects the inertia of cars (or, alternatively, reaction times of drivers), and V is the optimal velocity function defined by

$$V(d) = v_0(\tanh(d - h) + \tanh(h)),$$

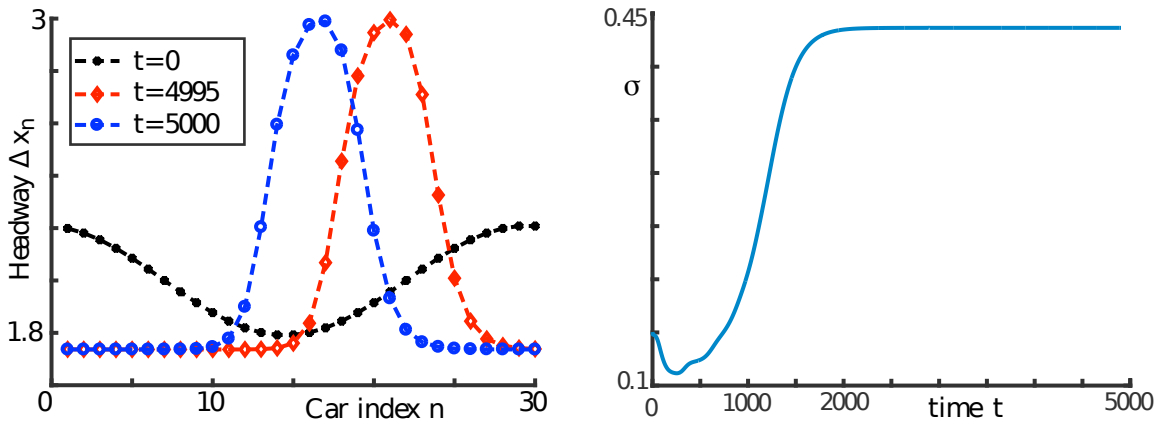

Figure 2.1: Shown are the time evolution of the headway profile (left) and the standard deviation from the mean headway (right) for a solution of (2.1.1) with $v_0 = 1$ (all other parameters are as in Table 2.1). In the left panel, the initial state (black) is compared with the final traveling-wave solution (red to blue): note that the traffic jam corresponds to the region where the headways are small, so that the car density is high. The right panel shows the time evolution of the deviation σ of the headways from the mean headway: the increase and eventual convergence of σ to a larger value indicates the emergence of a stable traffic jam.

where $v_0(1 + \tanh(h))$ is the maximal velocity and h determines the desired safety distance between cars. In this model, each driver adjusts their acceleration to attain an optimal velocity based on the distance $\Delta x_n = x_{n+1} - x_n$ to the car in front, which is also referred to as the headway. Two common traffic patterns can emerge as stable solutions in this model, namely free-flow solutions, where cars are evenly spaced and move with the same speed, and traveling-wave solutions, which correspond to traffic jams Marschler et al. (2014a). To illustrate traffic jam solutions, it is convenient to monitor the *headways* of cars: as indicated in Figure 2.1, localized traveling-wave profiles correspond to traffic jams.

In Marschler et al. (2014a), equation-free methods were used to trace out the bifurcation diagram for the existence and stability of traffic jam solutions as the parameter v_0 varies. For this analysis, the standard deviation σ of the headways,

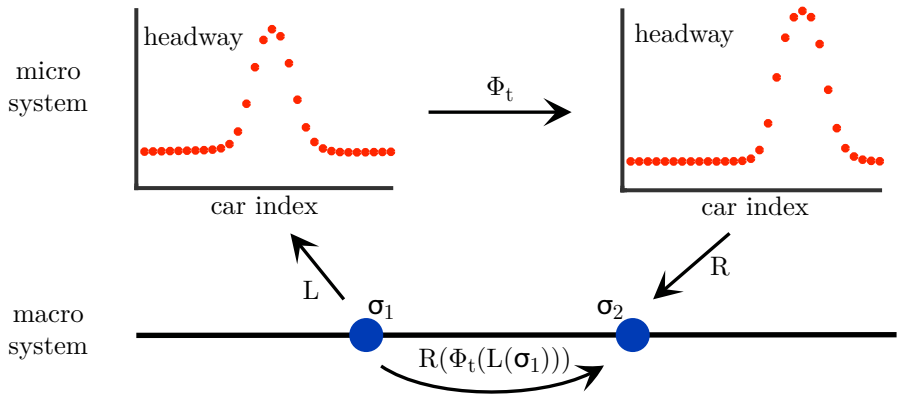

Figure 2.2: Overview of the time stepping scheme in the context of the traffic model Marschler et al. (2014a). The macroscopic state σ_1 is lifted to the microscopic profile $\mathcal{L}(\sigma_1)$. The microsystem is then evolved for time t to the next microstate $\Phi_t(\mathcal{L}(\sigma_1))$. Finally, this profile is restricted to find the evolution of the macroscopic state $\sigma_2 = \mathcal{R}(\Phi_t(\mathcal{L}(\sigma_1)))$.

defined by

$$\sigma = \sqrt{\frac{1}{N-1}\sum_{n=1}^{N}(\Delta x_n - \langle \Delta x_n \rangle)^2},$$

was used as the macroscopic variable, where $\langle \Delta x_n \rangle = \frac{L}{N}$ is the average headway. Low standard deviations correspond to free-flow solutions, while large values of the standard deviations correspond to traffic jams (see Figure 2.1). With σ as the macroscopic variable, the lifting and restriction operators are easy to define in terms of σ: as illustrated in Figure 2.2, lifting is accomplished by changing the locations of cars to match a given value of σ, and restricting is achieved by calculating σ for a given profile. Choosing σ as the macroscopic variable requires knowledge of the underlying structure of the traffic system. The standard deviation is a particularly good choice of macroscopic variable since it is invariant under translation.

We will see that the proposed approach via diffusion maps will, when applied to the same traffic model, generate a parametrization that recovers the standard deviation as one of the macroscopic variables; in addition, it identifies the location of the traveling wave as a second dimension in the parametrization. Using this new embed-

ding as our macroscopic variables, we apply equation-free methods to reproduce the bifurcation diagram in Marschler et al. (2014a), but we do so with restriction and lifting operators that emerge automatically from our diffusion map analysis without using prior knowledge of the system. In particular, we use the Nyström extension for our restriction operator, which gives estimates for new components of an eigenvector of a matrix constructed from data Liu et al. (2014), Marschler et al. (2014a), Sonday et al. (2009). We define a new lifting operator that creates microstates from linear combinations of existing data points. We apply these techniques to trace out the bifurcation diagram of traveling waves in the traffic system.

The remainder of this paper is organized as follows. In §2.2, we present an overview of equation-free methodologies. Then, in §2.3, we introduce the concept of diffusion maps and define diffusion map based operators to be used in equation-free modeling. We then apply these techniques to conduct bifurcation analysis in a traffic flow model in §2.4. Finally, in §2.5, we summarize our conclusions and give an outlook of open problems.

2.2 Overview of equation-free modeling

The equation-free approach is appropriate when working with a dynamical system of large dimension $N \gg 1$ that reflects a known microscopic evolution law with N variables and an attracting, low-dimensional, transversely stable manifold of dimension D. The N-dimensional system is referred to as the *microsystem*, and the D-dimensional manifold is the *macrosystem*. We assume that the system exhibits a sufficiently prominent time-scale separation: more precisely, we assume that the dynamics on the D-dimensional attracting manifold is slow compared to the fast

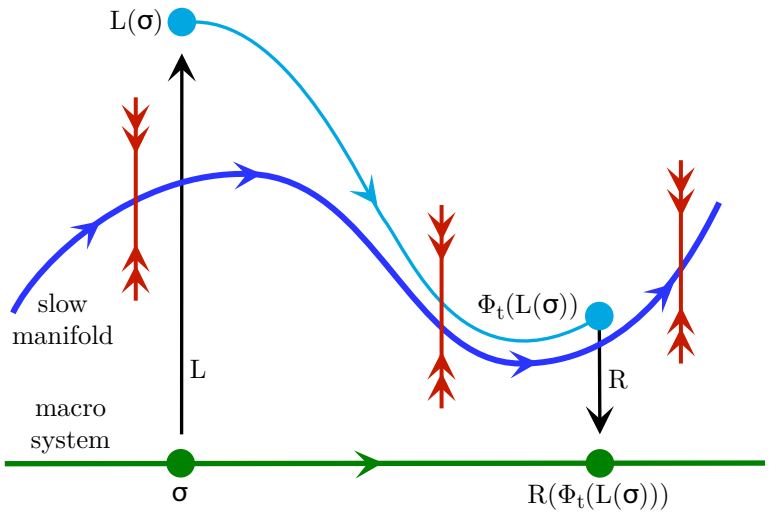

Figure 2.3: Sketch showing the equation-free modeling approach applied to a slow-fast system. An initial macrostate σ is lifted to a point that may possibly not lie on the slow manifold. Evolving this state for a short time will bring the profile close to the manifold. The evolved microstate can then be restricted back to the macro level or further evolved in time along the slow manifold.

transverse attraction towards this D-dimensional slow manifold Brunovský (1996), Jones (1995), Kuehn (2015). Once a system is known to be slow-fast, the goal is to choose macro-level variables that parametrize the slow manifold as best as possible.

The process of equation-free modeling uses two operators: lifting and restriction. The lifting operator $\mathcal{L} : \mathbb{R}^D \to \mathbb{R}^N$ maps a given *macrostate* to a corresponding *microstate*. Ideally, the lifting operator maps onto the slow manifold, but this is not easy to accomplish directly since the slow manifold may not be known. However, exploiting both time-scale separation and the assumption that the slow manifold is attracting, we only need to evolve the lifted microstate for a short time duration, using the time evolution $\Phi_t : \mathbb{R}^N \to \mathbb{R}^N$, to guarantee that the profile is close to the slow manifold, and can then use the resulting microstate as the image of the macrostate under lifting; see Figure 2.3 for an illustration. The additional short time evolution is often referred to as the *healing step*. The restriction operator $\mathcal{R} : \mathbb{R}^N \to \mathbb{R}^D$ takes a *microstate* and maps it to a low-dimensional *macrostate*.

Once the macrolevel parametrization is known, the restriction operator is usually much easier to define. In the traffic model, for instance, the restriction operator is defined to be the standard deviation of the headways of a microstate Marschler et al. (2014a). For consistency, we require that $\mathcal{R} \circ \mathcal{L} = \mathbb{I}_{\mathbb{R}^D}$, where $\mathbb{I}_{\mathbb{R}^D}$ is the identity map for the macrolevel variables.

The equation-free framework has many benefits as an approach to study macrolevel behavior. Once the operators are defined, traditional numerical analyses of the macrosystem can be conducted without constantly simulating the microsystem Kevrekidis and Samaey (2009). Since equilibria can exist in the macrostate without existing in the microstate, equation-free methods can even carry out macro-level bifurcation analyses that would be impossible with only the microsystem Kevrekidis and Samaey (2009), and we will demonstrate this in §2.4.

2.3 Lifting and restriction operators via diffusion maps

We first review diffusion maps Coifman et al. (2005), Marschler et al. (2014b) and then use this approach to construct lifting and restriction operators from a given data set. The goal of diffusion maps is to embed a large data set in a high-dimensional space $\mathbb{R}^N$ into a low-dimensional space $\mathbb{R}^D$ with $N \gg D$ so that the local geometry of the data is preserved.

2.3.1 Diffusion maps

Given a high-dimensional data set $X = \{X_m \in \mathbb{R}^N \mid m = 1, \ldots, M\}$ of m observations, we first calculate the pairwise distances between the data points X_m. Although any metric can be used, we use the Euclidean norm to define $d_{ij} = \|X_i - X_j\|$. Next, we define an affinity matrix $\mathcal{D} \in \mathbb{R}^{M \times M}$ such that a smaller distance corresponds to a high affinity and a larger distance corresponds to a small affinity. We use a Gaussian kernel to construct $\mathcal{D}$ from the pairwise distances via

$$\mathcal{D}_{ij} = \exp\left(\frac{-d_{ij}^2}{\epsilon^2}\right) = \exp\left(\frac{-\|X_i - X_j\|^2}{\epsilon^2}\right)$$

where the parameter ϵ should be chosen so that it reflects the spatial distance on which we want to resolve geometric features of the data set. Choosing ϵ too small treats all data points as singletons; picking an ϵ that is too large ignores the differences between data points: either way, we lose all geometric information. In our application in §2.4, we select

$$\epsilon = 5\,\text{median}(d_{ij})_{i>j}$$

to be five times the median of the pairwise distances d_{ij}, which yielded good results. Other strategies for selecting ϵ are discussed in Marschler et al. (2014b). We will discuss at the end of this section how we can measure the effectiveness of a given choice of ϵ for reducing the dimension quantitatively.

Next, we convert the affinity matrix $\mathcal{D}$ into a Markov transition matrix $\mathcal{M} \in$

$\mathbb{R}^{M\times M}$ by normalizing each row via

$$\mathcal{M}_{ij} := \frac{\mathcal{D}_{ij}}{\sum_{m=1}^{M} \mathcal{D}_{im}}.$$

For each fixed $0 < D < M$ and each choice of D eigenvalues $\lambda_1, \ldots, \lambda_D$ with associated eigenvectors $\psi_1, \ldots, \psi_D \in \mathbb{R}^M$ of the matrix $\mathcal{M}$, we follow Erban et al. (2007), Laing et al. (2006) and map the data set X into $\mathbb{R}^D$ via

$$X_m \longmapsto (\psi_{1,m}, \ldots, \psi_{D,m}) \in \mathbb{R}^D, \quad m = 1, \ldots, M, \tag{2.3.1}$$

where $\psi_{\ell,m}$ denotes the m^{th} element of the eigenvector $\psi_\ell \in \mathbb{R}^M$ for $\ell = 1, \ldots, D$.

The final step is to select the finite set of eigenvectors to represent the data set. A common choice is to choose the eigenvectors that belong to the D largest eigenvalues of $\mathcal{M}$, where D is chosen, for instance, to indicate a gap in the eigenvalues. This approach ignores the fact that not all eigenvectors add significantly new geometric information. Hence, we instead follow the algorithm proposed in Dsilva et al. (2015) to determine the dimension D and the eigenvectors that provide an optimal embedding. The idea is to pick eigenvectors recursively and use local linear fits with the previously selected set of eigenvectors to see whether the new eigenvector adds sufficient information to be included. Assume that we selected the first $j-1$ eigenvectors and set $\Psi_{j-1,m} := [\psi_{1,m}, \ldots, \psi_{j-1,m}]^T \in \mathbb{R}^{j-1}$ for $m = 1, \ldots, M$. Let ψ_j be the eigenvector of $\mathcal{M}$ with the largest eigenvalue that we have not considered yet. We then compute the local fit parameters

$$(\alpha_{j,m}, \beta_{j,m}) := \underset{\alpha\in\mathbb{R},\beta\in\mathbb{R}^{j-1}}{\arg\min} \sum_{i\neq m} \exp\left(\frac{-\|\Psi_{j-1,m} - \Psi_{j-1,i}\|^2}{\epsilon^2}\right) \left(\psi_{j,i} - \left(\alpha + \beta^T \Psi_{j-1,i}\right)\right)^2$$

where "local" refers to data points whose Gaussian distance is small, and the accompanying cross-validation error for the linear fit given by

$$r_j := \sqrt{\frac{\sum_{m=1}^{M}(\psi_{j,m} - (\alpha_{j,m} + \beta_{j,m}^T \Psi_{j-1,m}))^2}{\sum_{m=1}^{M} \psi_{j,m}^2}}.$$

Note that small values of r_j indicate that ψ_j is locally well approximated by $\psi_1, \ldots, \psi_{j-1}$, so that including ψ_j will not improve the embedding. We therefore include only eigenvectors with large r_j values in our diffusion-map embedding and stop its recursive definition once the values of r_j stay close to zero. A good choice of ϵ will result in a steep transition of the sequence r_j from values close to one to values close to zero.

2.3.2 Lifting and restriction operators

In general, equation-free modeling requires the definition of lifting and restriction operators that depend on the specific system we want to solve. Previous approaches rely on in-depth understanding of the relationship between the microscopic and macroscopic states. Here, we provide an algorithm for the construction of lifting and restriction operators that depends only on a given set of data points or observations and on a given diffusion-map embedding.

Restriction. First, we describe how we construct the restriction operator based on a given data set and the accompanying diffusion-map embedding. Assume $X \in \mathbb{R}^{N \times M}$ is an existing data set, and $\mathcal{M} \in \mathbb{R}^{M \times M}$ denotes the associated Markov transition matrix with eigenvalues λ_j and eigenvectors ψ_j for $j = 1, \ldots, M$. Assume also that we picked an embedding dimension D and that we ordered the eigenvalues

λ_j and eigenvectors ψ_j so that the restriction operator $\mathcal{R}$ evaluated on a point $X_m \in \mathbb{R}^N$ in the data set is defined by

$$\mathcal{R}(X_m) := (\psi_{1,m}, \ldots, \psi_{D,m}) \in \mathbb{R}^D, \qquad m = 1, \ldots, M, \tag{2.3.2}$$

where $\psi_{\ell,m}$ denotes the m^{th} component of ψ_ℓ. We need to extend the definition of $\mathcal{R}$ so that $\mathcal{R}(X_{\text{new}})$ is defined for each $X_{\text{new}} \in \mathbb{R}^N$. One option is to add the new data point X_{new} to the existing data set and recalculate for the new $(M+1)$-dimensional data set, but this is cumbersome and very expensive. Instead, we follow Coifman et al. (2008) and use the Nyström extension to extend $\mathcal{R}$ to new data points. This technique takes advantage of the fact that eigenvectors and eigenvalues are related by $\mathcal{M}\psi_\ell = \lambda_\ell \psi_\ell$ or, equivalently,

$$\psi_{\ell,m} = \frac{1}{\lambda_\ell} \sum_{j=1}^{M} \mathcal{M}_{m,j} \psi_{\ell,j}, \qquad \ell, m = 1, \ldots, M. \tag{2.3.3}$$

The embedding for a new data point X_{new} cannot be calculated directly from (2.3.3), but we can modify this equation as follows to approximate the embedding. As in §2.3.1, we define the Gaussian kernel

$$\mathcal{D}_{\text{new},m} = \exp\left(\frac{-\|X_{\text{new}} - X_m\|^2}{\epsilon^2}\right)$$

and use this expression to define

$$\psi_{\ell,\text{new}} := \frac{1}{\lambda_\ell} \sum_{m=1}^{M} \frac{\mathcal{D}_{\text{new},m}}{\sum_{j=1}^{M} \mathcal{D}_{\text{new},j}} \psi_{\ell,m}.$$

Following Liu et al. (2014), Marschler et al. (2014b), Sonday et al. (2009), we then set

$$\mathcal{R}(X_{\text{new}}) := (\psi_{1,\text{new}}, \ldots, \psi_{D,\text{new}}), \tag{2.3.4}$$

which extends the definition of the restriction operator $\mathcal{R}$ to include the new data point X_{new}.

Lifting. Next, we focus on the lifting operator. Given a macrostate in $\phi \in \mathbb{R}^D$, we need to define a lifted microstate $X = \mathcal{L}(\phi) \in \mathbb{R}^N$ so that $\mathcal{R}(X) = \mathcal{R}(\mathcal{L}(\phi)) \in \mathbb{R}^D$ is close to ϕ. Our goal is to construct the lifting operators using only the given data set in $\mathbb{R}^N$ and the parametrization via diffusion maps. Earlier work Erban et al. (2007), Laing et al. (2006), Sonday et al. (2009) approached this problem using simulated annealing, but this approach is computationally expensive. Instead, we define a lifting operator by solving an optimization problem.

To set up the algorithm, we choose an integer K with $K \geq D + 1$. Given the data points $X_m \in \mathbb{R}^N$ with $m = 1, \ldots, M$, we define

$$\phi_m := \mathcal{R}(X_m) = (\psi_{1,m}, \ldots, \psi_{D,m}) \in \mathbb{R}^D, \qquad m = 1, \ldots, M.$$

Given a macrostate $\phi_{\text{target}} \in \mathbb{R}^D$, we first find the K macrostates ϕ_{m_k} with $k = 1, \ldots, K$ that are closest to ϕ_{target} in $\mathbb{R}^D$. We then define the lifted state to be

$$\mathcal{L}(\phi_{\text{target}}) := \sum_{k=1}^{K} a_k X_{m_k},$$

where the coefficient vector $(a_1, \ldots, a_K) \in \mathbb{R}^K$ is determined as the solution to the optimization problem

$$(a_1, \ldots, a_K) := \underset{(b_1,\ldots,b_K)\in[0,1]^K}{\arg\min} \left\{ \left\| \phi_{\text{target}} - \mathcal{R}\left(\sum_{k=1}^{K} b_k X_{m_k} \right) \right\| \text{ subject to } \sum_{k=1}^{K} b_k = 1 \right\}. \tag{2.3.5}$$

Thus, we define the lifted microstate to be the element in the convex hull of the K microstates whose restriction is closest to the specified targeted macrostate. Note

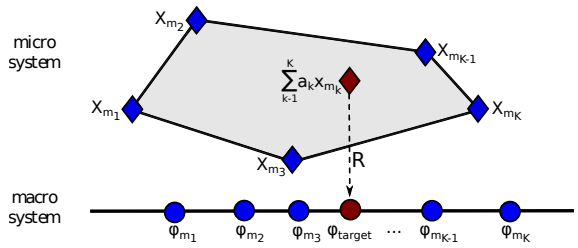

Figure 2.4: Shown is a visualization of the lifting operator. We solve for a linear combination of the K microstates corresponding to the K nearest macrostates such that the linear combination restricts to ϕ_{target}.

that (2.3.5) will always have a solution since $\mathcal{R}$ is continuous and the domain is compact. In general, (2.3.5) may have multiple solutions. In practice, choosing larger values of K ensures that zero is achieved as a minimum and evolving solutions forward will bring them close to the underlying attracting manifold: as discussed in the next section, we did not encounter any difficulties with potential discontinuities of the lifting operators during arclength continuation.

2.4 Case study: Traffic model

In this section, we will use the traffic model introduced in §2.1 to, first, demonstrate the accuracy and efficiency of the lifting and restriction operators we defined in §2.3.2 and, second, use these operators to compute bifurcation diagrams using equation-free modeling.

2.4.1 Traffic model

We write the traffic model introduced in §2.1 as the first-order system

$$\begin{aligned} \frac{dx_n}{dt} &= y_n, \\ \frac{dy_n}{dt} &= \frac{1}{\tau}\left[V(x_{n+1}-x_n)-y_n\right], \qquad n=1,\dots,N \end{aligned} \tag{2.4.1}$$

with $x_n \in \mathbb{R}/L\mathbb{Z}$, where the velocity function $V(d)$ is given by

$$V(d) = v_0(\tanh(d-h)+\tanh(h)). \tag{2.4.2}$$

We note that (2.4.1) is posed on the $2N$-dimensional space $(\mathbb{R}/L\mathbb{Z}\times\mathbb{R})^{2N}$, and we denote the solution of (2.4.1) with initial condition $P=(x_n,y_n)_{n=1,\dots,N}$ evaluated at time t by $\Phi_t(P)$. We record that (2.4.1) respects the action of the discrete symmetry group $\mathbb{Z}_N$ given by

$$(\mathbb{R}/L\mathbb{Z}\times\mathbb{R})^{2N} \longrightarrow (\mathbb{R}/L\mathbb{Z}\times\mathbb{R})^{2N}, \quad (x_n,y_n)_{n=1,\dots,N} \longmapsto (x_{(n+l)\mathrm{mod}N}, y_{(n+l)\mathrm{mod}N})_{n=1,\dots,N} \tag{2.4.3}$$

for $l\in\mathbb{Z}_N$, which corresponds to relabeling the cars consecutively. Throughout, we will fix the parameters as in Table 2.1, and focus on the emergence of free-flow and traffic-jam solutions as the parameter v_0, which appears in the velocity function, varies.

The free-flow solution of the microsystem (2.4.1) is defined by

$$x_n(t) = (n-1)\frac{L}{N} + tV\left(\frac{L}{N}\right) \mod L, \qquad y_n(t) = V\left(\frac{L}{N}\right), \qquad n=1,\dots,N.$$

It captures traffic flows where all cars keep the same distance L/N from each other and travel with the velocity dictated by the constant headway. Traffic jams are

Table 2.1: Parameter descriptions and values

Parameter	Description	Value
N	number of cars	30
L	length of the road	60
τ^{-1}	inertia of the car	1.7
h	desired safety distance between cars	2.4
v_0	optimal velocity parameter	Uniform([0.96, 1.1])
A	amplitude of initial conditions in the dataset	Uniform([0, 4.5])
t_{stop}	time evolved to create dataset	Exp(mean=700, shift=200)
K	number of points used for lifting	3 ($n = 1$)
		8 ($n = 2$)
t_{skip}	healing evolution time	300
δ	time evolved for finite difference approximation	240
s	continuation step size	0.0025 ($m = 5,000$)
		0.01 ($m = 1,000$)
ν	integer multiple of period sought	7

captured by the traveling-wave ansatz

$$(x_n(t), y_n(t)) = (x_*(n - ct), y_*(n - ct)), \qquad n = 1, \dots, N,$$

where $(x_*(\xi), y_*(\xi))$ are N-periodic functions that describe the profile of the traveling wave, and c is its speed. Substituting this ansatz into (2.4.1) shows that the N-periodic profile $(x_*(\xi), y_*(\xi))$ and the wave speed c need to satisfy the system

$$\begin{aligned} -c\frac{dx_*}{d\xi}(\xi) &= y_*(\xi) \\ -c\frac{dy_*}{d\xi}(\xi) &= \frac{1}{\tau}\left(V(x_*(\xi + 1) - x_*(\xi)) - y_*(\xi)\right) \end{aligned} \tag{2.4.4}$$

of delay differential equations. Note that the wave speed c is a variable that needs to be solved for as part of (2.4.4).

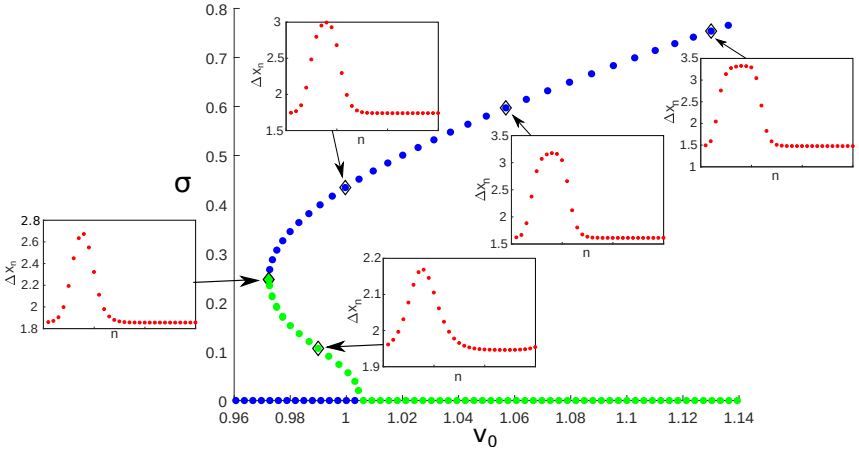

Figure 2.5: Shown is the zero set of the function $\mathcal{F}^{\text{tw}}$ obtained by pseudo-arclength continuation. The associated traveling-wave profiles are shown for selected points (marked with black diamonds) on the bifurcation branch. Note that the scale on the vertical axes of the insets varies to better illustrate the shape of the profiles. The algorithm detects a fold point at $(v_0, \sigma) \approx (0.97, 0.25)$, where stability changes: blue dots mark stable states, and green dots mark unstable states.

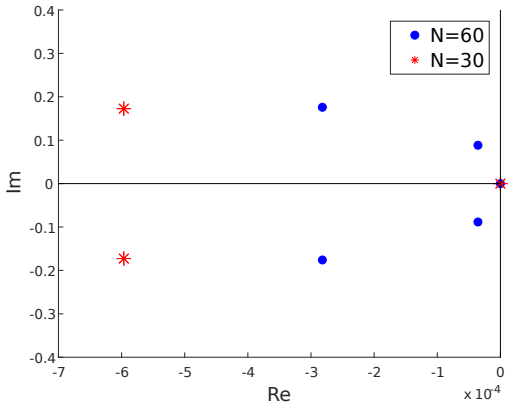

Figure 2.6: Shown are the Floquet spectra of the linearization of the traffic system (2.1.1) about a traveling-wave solution for the fold points for $N = 30$ and $N = 60$ ($v_0 = 0.97, 0.88$ and $h = 2.4, 1.2$, respectively). Note that the nonzero Floquet exponents are much closer to zero for $N = 60$ than for $N = 30$, which indicates that the slow-fast time-scale separation becomes less pronounced as N increases.

2.4.2 Computing bifurcation diagrams using the microsystem

First, we use pseudo-arclength continuation to compute and continue traveling waves, and their wave speeds, as N-periodic solutions to (2.4.4) as the parameter v_0 is varied. We will use the headways $u(\xi) := x(\xi+1) - x(\xi)$ instead of the positions $x(\xi)$ as variables. We will argue that N-periodic traveling waves correspond to regular roots of the function

$$\begin{aligned} \mathcal{F}^{\mathrm{tw}} : \quad & C^2(\mathbb{R}/N\mathbb{Z}) \times \mathbb{R}^2 \longrightarrow C^0(\mathbb{R}/N\mathbb{Z}) \times \mathbb{R}^2 \\ & (u,c,d) \longmapsto \begin{pmatrix} \xi \mapsto c^2\tau \frac{d^2u}{d\xi^2}(\xi) - c\frac{du}{d\xi}(\xi) - V(u(\xi+1)) + V(u(\xi)) + d \\ L - \sum_{n=0}^{N-1} u(n) \\ \int_0^N \left\langle \frac{du}{d\xi}(\xi), u_*(\xi) - u(\xi) \right\rangle d\xi \end{pmatrix} \end{aligned} \tag{2.4.5}$$

for each fixed value of v_0. The first component $\mathcal{F}_1^{\mathrm{tw}}$ is the delay-differential equation (2.4.4) written as a function of the headways; the additional term d accounts for the fact that the first component (with $d = 0$) has mass zero, so that $\int_0^N \mathcal{F}_1^{\mathrm{tw}}(u,c,d)(\xi)d\xi = 0$ for all (u,c,d). The second component of $\mathcal{F}^{\mathrm{tw}}$ ensures that the headways add up to the length L of the ring road. Finally, the last component is a phase condition that selects a unique profile amongst the family of spatial translates of a given solution $u_*(\xi)$; during continuation, $u_*(\xi)$ is normally taken to be the solution obtained at a previous continuation step. The following result gives conditions that guarantee that the set of roots of $\mathcal{F}^{\mathrm{tw}}$ consists of regular zeros.

Theorem 2.4.1. *Let*

$$\begin{aligned} F : H^2(S^1) &\to L^2(S^1), \\ u &\mapsto u''(x) - u'(x) - V(u(x+1)) - V(u(x)), \end{aligned}$$

and

$$G : H^2(S^1) \to (S^1),$$
$$u \mapsto \sum_{j=0}^{n-1} u(j).$$

Assume $u^* \in H^2(S^1)$ *satisfies* $F(u^*) = 0$ *and* $N(DF(u_*)) = \mathbb{R}k$ *with* $Gk \neq 0$. *Then, the linearization of the map*

$$\mathcal{F} : H^2(S^1) \times \mathbb{R} \times \mathbb{R} \to L^2(S^1) \times \mathbb{R} \times \mathbb{R},$$
$$(u, \mu, c) \mapsto (Fu + \mu, Gu, \int_0^L \langle u'_*(s), u'(s) \rangle ds),$$

around the point (u^*, μ^*, c^*) *is invertible.*

Proof. Consider the linearization of $\mathcal{F}$ about (u^*, μ^*, c^*) given by

$$J : \begin{pmatrix} v \\ m \\ \xi \end{pmatrix} \mapsto \begin{pmatrix} F_u(u^*)v + m(u^*_x - 2c\tau u^*_{xx}) + \xi 1 \\ Gv \\ \langle u^*_x, v \rangle \end{pmatrix}$$

To show J is invertible, it is sufficient to show that the kernel is trivial.

Suppose $w = \begin{pmatrix} v \\ \epsilon \\ \mu \end{pmatrix}$ satisfies $Jw = 0$. This implies that

$$F_u(u^*)v + m(u^*_x - 2c\tau u^*_{xx}) + \xi 1 = 0.$$

We claim that $F_u(u^*)v + m(u_x^* - 2c\tau u_{xx}^*) \in R(F_u(u^*))$ and $\xi 1 \in R(F_u(u^*))$ so that

$$F_u(u^*)v + m(u_x^* - 2c\tau u_{xx}^*) = 0,$$
$$\xi 1 = 0.$$

If we differentiate F with respect to c and evaluate at u^*, then

$$\begin{aligned} 0 &= -2c\tau u_{xx}^* - c^2\tau u_{xxc}^* + u_x^* + cu_{xc}^* + V'(u^*(x+1))u_c^*(x+1) - V'(u^*(x))u_c^* \\ &= -2c\tau u_{xx}^* + u_x^* + F_u(u^*)u_c^*. \end{aligned}$$

Thus,

$$F_u(u^*)u_c^* = 2c\tau u_{xx}^* - u_x^*,$$

so $(u_x^* - 2c\tau u_{xx}^*) \in R(F_u(u^*))$ as claimed. Additionally, note that

$$R(F) \subset \{u \in H^2(S^1) : \int_0^n u(x)dx = 0\} = (\mathbb{R}1)^\perp,$$

so $\xi 1 \in R(F_u(u^*))$ as claimed. Since

$$F_u(u^*)v - mF_u(u^*)u_c^* = F_u(u^*)v + m(u_x^* - 2c\tau u_{xx}^*) = 0,$$

it follows that $\xi = 0$ and $v = au_x^* - mu_c^*$ for some $a \in \mathbb{R}$.

Next, consider $Gv = 0$. By the above argument and the assumption that $Gu_c^* = 0$, we have that

$$0 = Gv = aGu_x^* - mGu_c^* = aGu_x^*.$$

We claim that $Gu_x^* = 0$. Define $q(x) = \sum_{j=0}^{n-1} u^*(x+j)$ so that $Gu^* = q(0)$. Since $F(u^*) = 0$, we have that $l''(x) - l'(x) = 0$ for all x, which has the solution $l(x) = k +$

$e^x(l(0)-k)$. We know that l must be n-periodic, so it follows that $l(x) = l(0) = Gu^*$. Thus, $l'(x) = 0$ for all x, which implies $Gu^*_x = l'(0) = 0$ as desired.

Finally, consider the last component of Jw :

$$\begin{aligned} 0 &= \langle u^*_x, v\rangle \\ &= a\left\|u^*_x\right\|^2 - m\langle u^*_x, u^*_c\rangle \\ &= a\left\|u^*_x\right\|^2, \end{aligned}$$

so $a = 0$. Thus, we have that $v = 0$, and subsequently, $w = 0$. □

Theorem 2.4.2. *Fix v_0. If $\mathcal{F}^{\mathrm{tw}}(u_*, c_*, 0) = 0$, the null space of $D_u\mathcal{F}^{\mathrm{tw}}_1(u_*, c_*, 0)$ is two-dimensional and spanned by $u'_*, v \in C^2(\mathbb{R}/N\mathbb{Z})$ with $\sum_{j=0}^{N-1} v(n) \neq 0$, and $D_c\mathcal{F}^{\mathrm{tw}}_1(u_*, c_*, 0)$ is not in the range of $D_u\mathcal{F}^{\mathrm{tw}}_1(u_*, c_*, 0)$, then $D_{(u,c,d)}\mathcal{F}^{\mathrm{tw}}(u_*, c_*, 0)$ has a bounded inverse.*

Proof. We give only a brief outline of the proof. The key observations are that $D_u\mathcal{F}^{\mathrm{tw}}_1$ is Fredholm of index zero and that elements in its range have mass zero. Since u'_* is contained in the null space of $D_u\mathcal{F}^{\mathrm{tw}}_1$, it is not difficult to show that the null space is at least two-dimensional, and we assumed that its dimension is indeed two and that the null space is spanned by u'_*, v. Using this information and the remaining assumptions, it is now straightforward to prove that the null space of the full linearization $D_{(u,c,d)}\mathcal{F}^{\mathrm{tw}}_1(u_*, c_*, 0)$ is trivial, which proves the theorem since this operator is also Fredholm with index zero by the bordering lemma (see (Beyn, 1990, Lemma 2.3)). □

Theorem 2.4.2 indicates that we can use arclength continuation with a secant predictor and Newton's method as corrector to compute branches of traveling waves

by solving $\mathcal{F}^{\text{tw}}(u, c, d; v_0) = 0$, where $\mathcal{F}^{\text{tw}}$ depends on v_0 through the optimal velocity function $V(u) = V(u; v_0)$. We implemented this algorithm in Fourier space to take advantage of spectral convergence.

The result of the numerical continuation is visualized in Figure 2.5 using the standard deviation σ of the headways $u_*(n) := x_*(n+1) - x_*(n)$. Figure 2.6 shows the Floquet spectra of the linearization of the microscale system about sample traveling waves for $N = 30$ and $N = 60$. In both cases, $\lambda = 0$ is an eigenvalue of multiplicity two. We observe that the gap between the nonzero Floquet exponents and the eigenvalues at the origin is much smaller for $N = 60$ than for $N = 30$. In fact, the gap will shrink to zero as N goes to infinity due to the presence of a conservation law in the continuum limit and, in particular, the slow-fast time-scale separation will become less prominent as N increases. For this reason, we focus on the case $N = 30$ in the remainder of this paper.

2.4.3 Constructing embeddings using diffusion maps

Our goal is to use diffusion maps to construct an embedding of the essential dynamics of the microsystem (2.4.1) into a low-dimensional space and identify macroscopic variables that parametrize the reduced dynamics.

Construction of the data set. We construct the data set X to which we apply the diffusion-map approach as follows. We generate $m = 5000$ initial conditions of the form

$$x_n(0) = \frac{L(n-1)}{N} + A \sin\left(\frac{2\pi n}{N}\right), \quad y_n(0) = V\left(\frac{L}{N}\right), \qquad n = 1, \ldots, N,$$

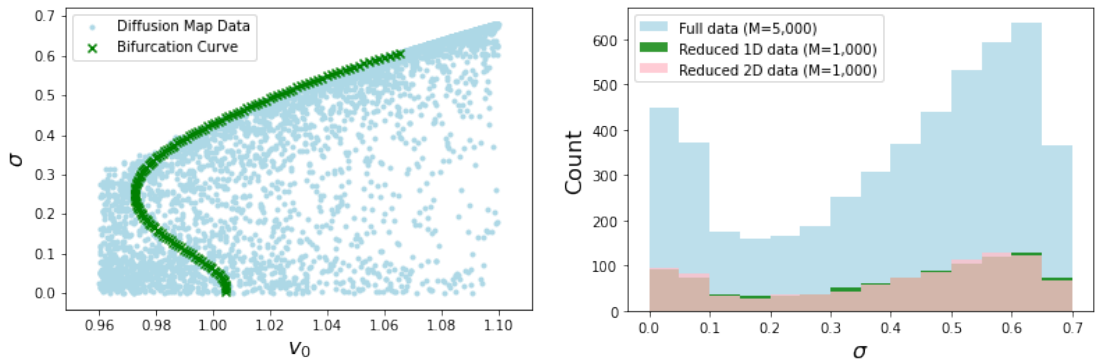

Figure 2.7: Shown are the data points generated for the diffusion map.

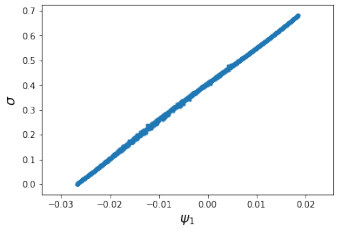

Figure 2.8: Plotted are the values of the restriction operator $\mathcal{R}(X_m) = \psi_{1,m}$ against the standard deviation σ evaluated at the data points X_m for $m = 1, \ldots, M$. Since there is a positive linear relationship between the diffusion map embedding and standard deviation of each data point, the two represent the same feature.

and draw values for the coefficient $A \in [0, 4.5]$ and the parameter $v_0 \in [0.96, 1.1]$ randomly using the uniform distribution on these intervals. Each corresponding trajectory of (2.4.1) is evolved using an ODE solver until a stopping time t_{stop} is reached that is drawn from a shifted exponential curve with mean 700 and shift 200. The collection of the M end states of the headways in $\mathbb{R}^N$ defines the data set X. We emphasize that we do not include information about the velocities in our data set. By varying the parameters and the time captured, our data set X is comprised of varying wave shapes as well as some free-flow profiles. Figure 2.7 shows that this sampling results in good coverage of the space encapsulating the bifurcation diagram.

Reduction to one dimension: factoring out the discrete symmetry. First, we explicitly factor out the discrete symmetry (2.4.3) present in the model (2.4.1).

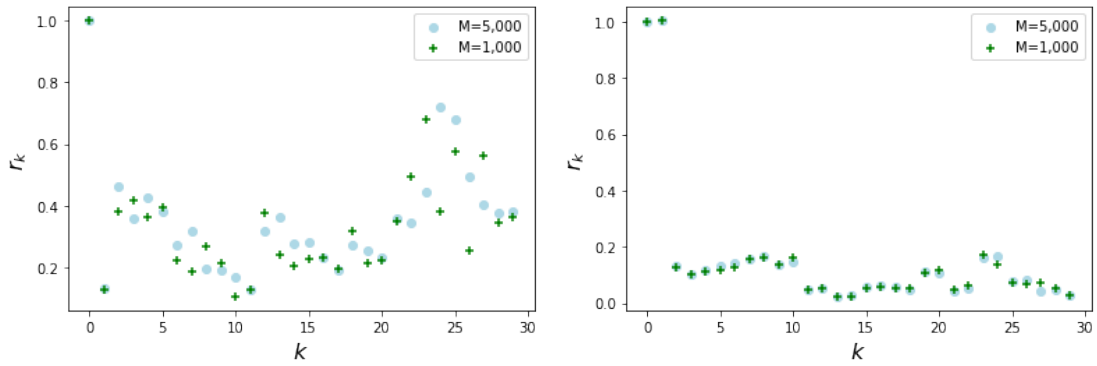

Figure 2.9: Shown are the local linear fit coefficients r_k as functions of the index k for the diffusion maps for $D = 1$ (left panel) and $D = 2$ (right panel).

We accomplish this by shifting the indices in each profile in our data set using the symmetry (2.4.3) so that the maximum headway $\max(x_{n+1} - x_n)_{n=1,\dots,N}$ inside the profile is achieved at $n = 10$. If all solutions converge to, or at least resemble, traveling waves, factoring out the discrete symmetry effectively factors out the one-dimensional phase of all traveling waves and should therefore reduce the effective dimension of the embedding by one. Applying the diffusion-map approach outlined in §2.3.1 to the set X_{align} of aligned elements in $\mathbb{R}^N$, we indeed find that we can take $D = 1$ as the embedding dimension so that the resulting restriction operator (2.3.2) maps the aligned data set X_{align} into $\mathbb{R}$ (Figure 2.9). Figure 2.8 shows that the embedding parametrizes the data set using essentially the standard deviation of the headways. In particular, the diffusion-map approach proposed here automatically generates the parametrization introduced previously in Marschler et al. (2014a) based on *a priori* knowledge of the dynamics.

Reduction to two dimensions. Next, we apply the diffusion-map approach from §2.3.1 directly to the original data set X without any alignment or other adjustments. In this case, we can take $D = 2$ as the embedding dimension, and the resulting restriction operator (2.3.2) therefore maps the data set X into $\mathbb{R}^2$ (Figure 2.9).

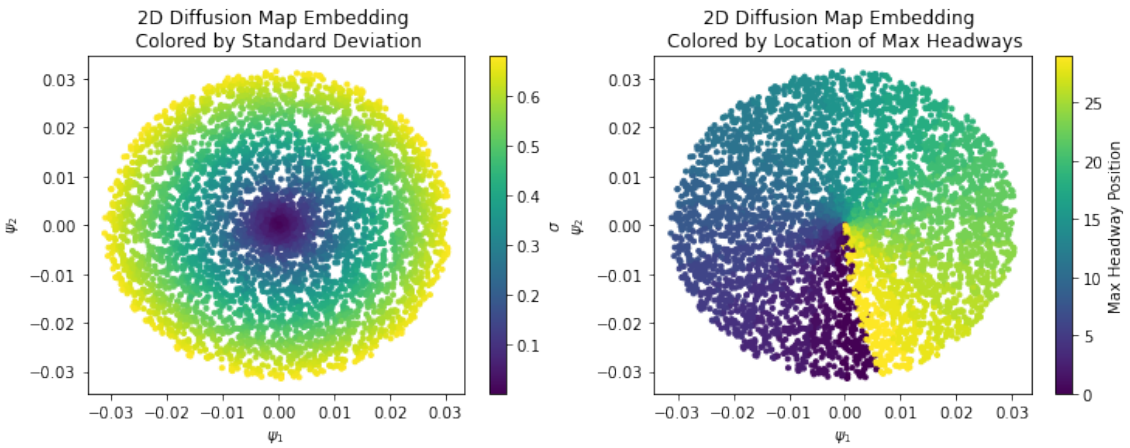

Figure 2.10: Shown is the image of the data set X under the diffusion-map embedding into $\mathbb{R}^2$ where the colors indicate the standard deviation (left panel) and the wave peak position (right panel) of the pre-images $X_m \in X$.

Figure 2.10 shows that the planar embedding parametrizes the data set through polar coordinates where the radial direction corresponds to the standard deviation of headways and the angular direction captures the location of the peak of solutions along the circular ring road.

Reducing the number of data points to $M = 1000$. To see if we can produce the same results with fewer data points, we also reduce the original diffusion map from $M = 5000$ to $M = 1000$. We use the 5000 point diffusion map to inform the down-sampling of the data. For the one-dimensional case, we sort the embedding to be numerically ascending, uniformly sample the $M = 1000$ data points corresponding to those embeddings, and then recompute the diffusion map on those points. Since the two-dimensional diffusion map resembles a disc, we covert the embedding to polar coordinates, uniformly sample 2000 points radially, and then further uniformly sample $m = 1000$ points with respect to the angular component. Finally, we recompute the diffusion map using the data points corresponding to the sampled embeddings. Figure 2.7 shows that this down-sampling approach preserves the dis-

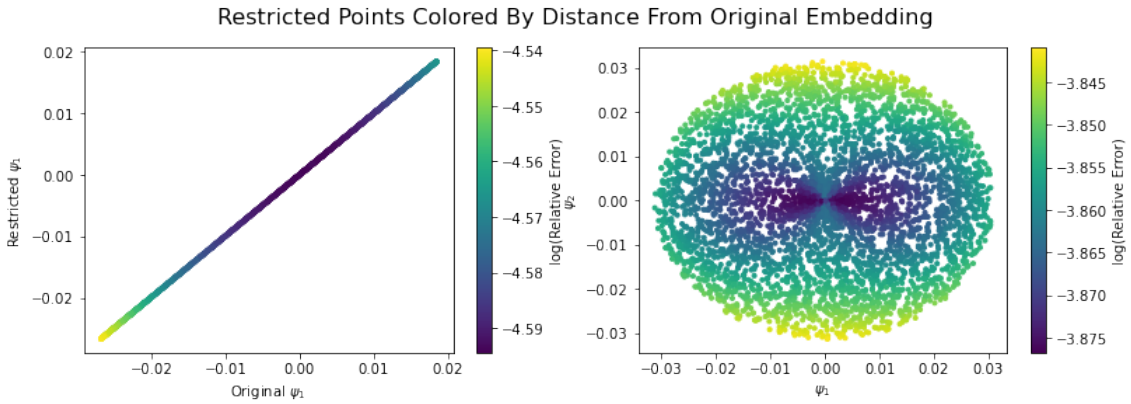

Figure 2.11: To test the accuracy of the restriction operator defined via the Nyström extension, we visualize the differences $\mathcal{R}(X_m) - \mathcal{R}_{X\setminus\{X_m\}}(X_m)$ (see main text for their definitions). Left panel: For $D = 1$, we plot the one-dimensional coordinates of $\mathcal{R}(X_m)$ and $\mathcal{R}_{X\setminus\{X_m\}}(X_m)$ against each other as m varies and indicate the logarithm of the relative error of their difference using colors. Right panel: For $D = 2$, we plot the images $\mathcal{R}(X_m)$ as m varies and visualize the logarithm of the relative error of their difference using colors.

tribution of σ values in the dataset. We also observe the same relationship with σ and the location of the maximum headways in the embeddings.

2.4.4 Lifting and restriction operators

We now test the accuracy of the lifting and restriction operators that we defined in §2.3.2 based on the diffusion-map embedding constructed in §2.4.3.

First, we test the accuracy of the restriction operator defined via the Nyström extension. For each fixed element $X_m \in \mathbb{R}^N$ of our data set X (or X_{align}), we first calculate the image $\mathcal{R}(X_m)$ of X_m under the restriction operator (2.3.2). Separately, we compute the restriction operator $\mathcal{R}_{X\setminus\{X_m\}}$ by applying the algorithm outlined in §2.3.2 to the data set $X \setminus \{X_m\}$, obtained from X by removing the element X_m, and apply this restriction operator to the removed element X_m to obtain $\mathcal{R}_{X\setminus\{X_m\}}(X_m)$ via the Nyström extension (2.3.4). The norm of difference $\mathcal{R}(X_m) - \mathcal{R}_{X\setminus\{X_m\}}(X_m)$

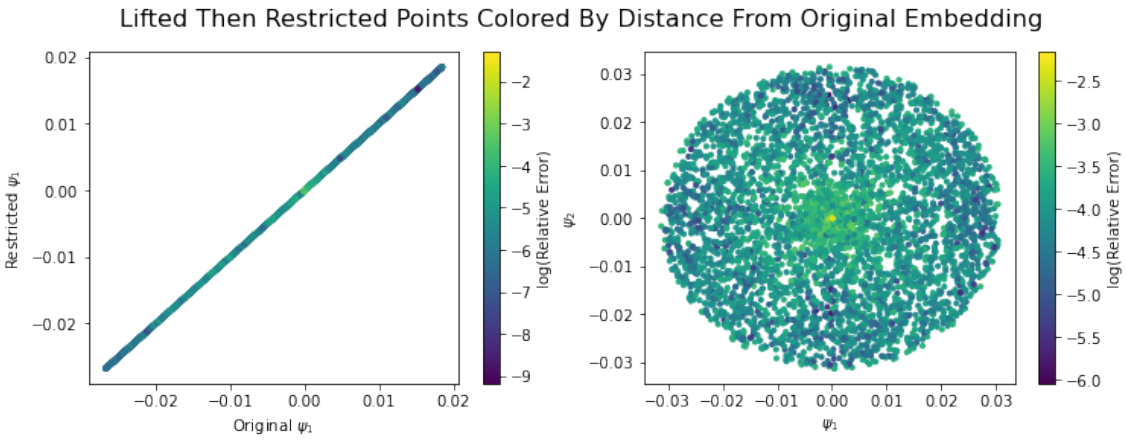

Figure 2.12: Shown are the norms of the difference $\mathbb{I}_{\mathbb{R}^D} - \mathcal{R} \circ \mathcal{L}$ evaluated on the image $\mathcal{R}(X_{\text{align}})$ for $D = 1$ (left panel) and on $\mathcal{R}(X)$ for $D = 2$ (right panel).

measures the accuracy of our approach, and Figure 2.11 shows that the magnitude of the difference of these two images for both one- and two-dimensional embeddings is less than 10^{-5}. For $D = 1$, the average relative error is about 0.26%, while it is about 1.38% for $D = 2$. Plotting the errors as function of the reduced macrosystem, we see that the error is smallest in the center of domain; see Figure 2.11. Since the Nyström extension essentially takes a linear combination of the existing points, this observation is not unexpected.

Next, the theoretical approach to equation-free modeling assumes that $\mathcal{R} \circ \mathcal{L} = \mathbb{I}_{\mathbb{R}^D}$ is the identity in the macrovariables in $\mathbb{R}^D$. To test this property, we calculate the sup-norm of the map

$$\mathbb{I}_{\mathbb{R}^D} - \mathcal{R} \circ \mathcal{L} : \quad \mathcal{R}(Y) \subset \mathbb{R}^D \longrightarrow \mathbb{R}^D, \quad \phi \longmapsto \phi - \mathcal{R}(\mathcal{L}(\phi))$$

for $Y = X, X_{\text{align}}$ and plot the results in Figure 2.12 separately for the aligned and the original data sets. In the lifting operator, we set the number of interpolating points to $K = 3$ for $D = 1$ and $K = 8$ for $D = 2$. We observe very little difference in accuracy based on the value of K but chose these values as they show the greatest

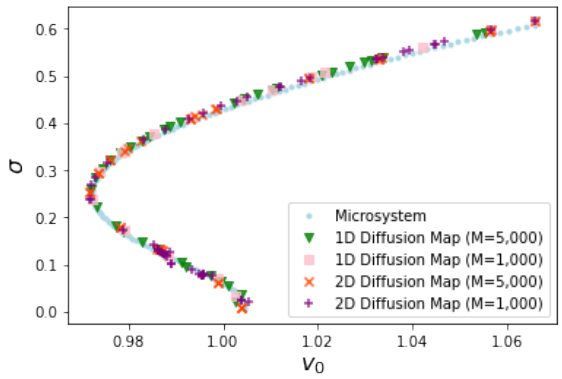

Figure 2.13: Shown is the bifurcation diagram in σ coordinates computed for the microsystem directly, the one-dimensional, two-dimensional, full, and reduced diffusion maps.

accuracy. For $D = 1$, the average relative error is less than 0.27%. For $D = 2$, this error increases to 1.63%. We observe the least accuracy near the embeddings corresponding to low values of σ, likely due to the fact that these traffic jam solutions are unstable and thus have more variation in sampled profiles.

2.4.5 Computing bifurcation diagrams using the macrosystem

In the preceding sections, we presented a data-driven approach to constructing lifting and restriction operators based on embeddings derived from diffusion maps and validated this approximation. In this section, we use an equation-free modeling approach based on these operators to compute the bifurcation diagram of traveling-wave solutions of (2.4.1) in the reduced n-dimensional embedding space separately for $D = 1$ and $D = 2$.

One-dimensional reduction: continuing fixed points. First, we focus on the lifting and restriction operators constructed from the aligned data set X_{align}, where

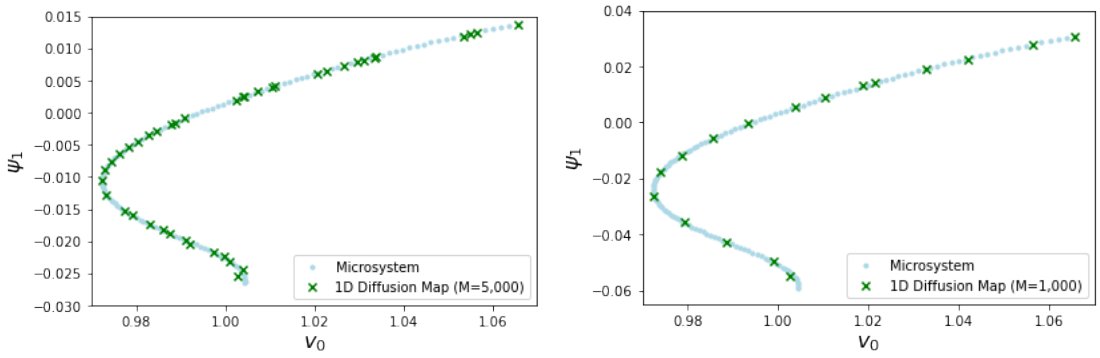

Figure 2.14: Shown is a comparison of the bifurcation diagram computed using the equation-free approach for the reduced system with $D = 1$ associated with the aligned data set and the diagram computed using pseudo-arclength continuation for the microsystem. Left: $M = 5,000$ points, Right: $M = 1,000$ points

we explicitly factored out the discrete symmetry. As shown above, the restriction operator maps into $\mathbb{R}$, and the parametrization of the macrosystem corresponds to the standard deviation of the headways. Since the phase is effectively factored out, we focus on computing and continuing equilibria of the macrosystem defined implicitly using an equation-free model.

We denote by $\Phi_t(P; v_0)$ the solution of the microsystem (2.4.1) with parameter value v_0 that belongs to the initial condition $P = (x_n, y_n)_{n=1,\ldots,N} \in \mathbb{R}^N$. The equation-free macrosystem is then defined by the finite-difference quotient

$$\frac{d\phi}{dt} = F(\phi, v_0) := \frac{\mathcal{R}\left(\Phi_{t_{\text{skip}}+\delta}(\mathcal{L}(\phi); v_0)\right) - \mathcal{R}\left(\Phi_{t_{\text{skip}}}(\mathcal{L}(\phi); v_0)\right)}{\delta},$$

where $\phi \in \mathbb{R}$ denotes the macrovariable, and the time steps $t_{\text{skip}}, \delta > 0$ are chose to obtain time-scale separation in the dynamics. We then compute and continue roots of the function $F(\phi, v_0)$ using pseudo-arclength continuation with a secant predictor of step size s and a Newton corrector in the space $(\phi, v_0) \in \mathbb{R}^2$.

The results are shown in σ coordinates in Figure 2.13 and the diffusion map coordinates Figure 2.14 for both the full diffusion map and the reduced diffusion map.

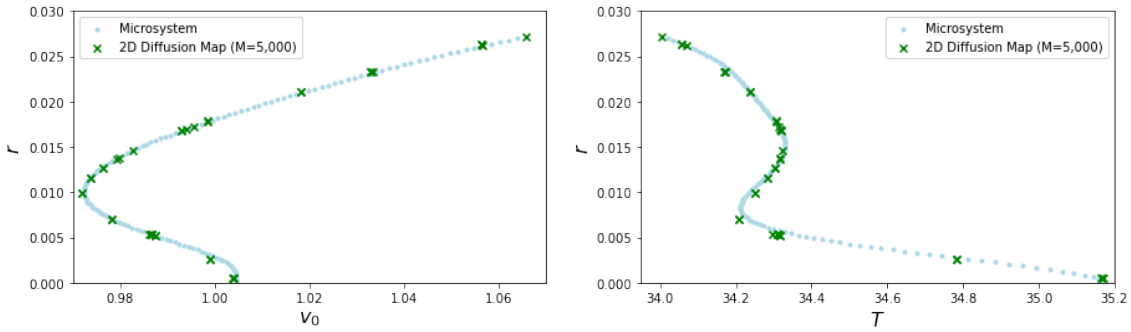

Figure 2.15: Shown is a comparison of the bifurcation diagram of periodic orbits computed using the equation-free approach for the reduced system with $D = 2$ associated with the data set X and the diagram computed using pseudo-arclength continuation for the microsystem. The left and right panel show projections into (v_0, r) and (T, r), respectively.

Both bifurcation diagrams resemble the diagram computed previously in Marschler et al. (2014a) and the diagram in Figure 2.5 computed using continuation in the microsystem. We notice that the reduced diffusion map is more robust to continuation parameter choices, and we hypothesize that this difference results from the artificial alignment of the profiles in the one-dimensional diffusion map. In the $M = 5,000$ diffusion map, aligning the data introduces noise around each value of σ since there will be many profiles originating from different positions to potentially lift. Downsampling the diffusion map to $M = 1,000$ reduces some of that noise, thus resulting in a more robust method.

Two-dimensional reduction: continuing periodic orbits. Next, we focus on the lifting and restriction operators constructed from the original data set X, which results in a two-dimensional macrosystem given by

$$\frac{d\phi}{dt} = F(\phi, v_0) := \frac{\mathcal{R}\left(\Phi_{t_{\text{skip}}+\delta}(\mathcal{L}(\phi); v_0)\right) - \mathcal{R}\left(\Phi_{t_{\text{skip}}}(\mathcal{L}(\phi); v_0)\right)}{\delta}, \tag{2.4.6}$$

where $\phi \in \mathbb{R}^2$ again denotes the, now two-dimensional, macrovariable. Traveling waves of the microsystem (2.4.1) correspond to periodic orbits of the planar system (2.4.6). We denote by $\tilde{\Phi}_t(\phi; v_0)$ the solution operator of the macrosystem (2.4.6) and define the one-dimensional Poincare section $\mathbb{R}\phi_{\text{PS}}$ for a nonzero vector $\phi_{\text{PS}} \in \mathbb{R}^2$ in the macrosystem. Periodic orbits with period T of (2.4.6) can then be found as solutions of the system $\mathcal{F}(r, T, v_0) = 0$ given by

$$\mathcal{F}: \quad \mathbb{R}^3 \longrightarrow \mathbb{R}^2, \quad (r, T, v_0) \longmapsto \tilde{\Phi}_{t_{\text{skip}}}(r\phi_{\text{PS}}; v_0) - \tilde{\Phi}_{t_{\text{skip}}+\nu T}(r\phi_{\text{PS}}; v_0).$$

We solve this system for (r, T, v_0) using again pseudo-arclength continuation with a secant predictor and a Newton corrector. Since the periods are roughly $T \approx 34.5$, we choose to evolve for $\nu = 7$ periods to match $\delta = 240 \approx \nu T$ used in the one-dimensional system.

The results are illustrated in σ coordinates in Figure 2.13 and in diffusion map coordinates in Figure 2.15. The full diffusion map diagram is computed accurately in (r, T, v_0)-space and deviates from the diagram obtained from the microsystem only slightly near the fold point. The reduced diffusion map diagram shown in Figure 2.16 is less accurate, particularly on the stable branch. In this case, downsampling the data likely left gaps in the (r, T, v_0) space of interest, leading to less accurate results.

2.5 Conclusions

We considered large-dimensional dynamical systems, referred to as the microsystem, with a slow-fast time-scale separation. Equation-free modeling attempts to reduce the dynamics of the microsystem to an implicitly defined, lower-dimensional

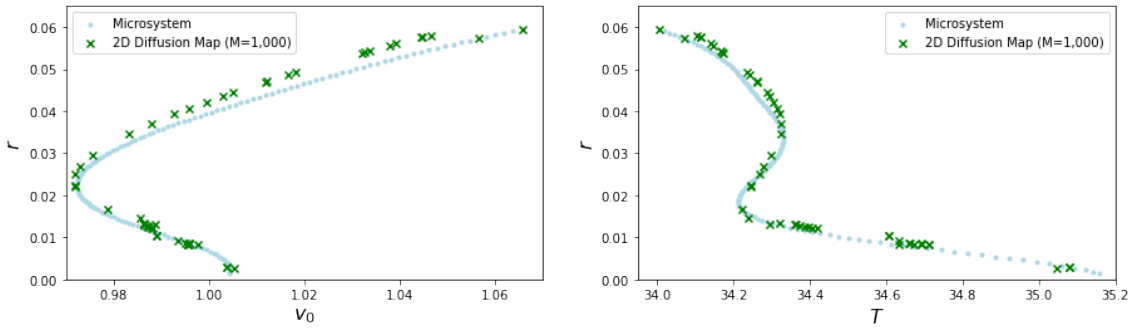

Figure 2.16: Shown is a comparison of the bifurcation diagram of periodic orbits computed using the equation-free approach for the reduced system with $D = 2$ associated with the reduced data set X and the diagram computed using pseudo-arclength continuation for the microsystem. The left and right panel show projections into (v_0, r) and (T, r), respectively.

macrosystem that captures the dynamics on the slow manifold of the microsystem corresponding to the slow time scale. The macrosystem can then be used, for instance, to carry out direct numerical simulations with larger step sizes compared to simulations of the microsystem or to compute and continue stationary solutions or periodic orbits of the macrosystem that may correspond to more complex patterns in the microscopic variables. In previous applications, the macroscopic variables were identified based on insights into the microsystem: for instance, in the context of traffic-flow models, it makes sense to use the standard deviation from the free-flow solution to capture traffic-jam solutions.

For this paper, our goal was to develop an application-independent approach to equation-free modelling that does not rely on being able to make an explicit ansatz for the macroscopic variables. We focused on a data-driven approach and used diffusion maps to embed the data set into a lower-dimensional space, identify macroscopic variables that parametrize the low-dimensional space, and construct lifting an restriction operators that connect the micro- and macroscopic systems. Our case study demonstrated that these operators can be used to continue traffic-flow patterns as steady states or as periodic orbits in the macroscopic system.

It would be interesting to see whether this approach can be used to compute and continue more complex patterns that cannot be characterized directly in the microsystem: examples are patches of turbulent fluid surrounded by laminar flow, for instance, or other spatially chaotic structures whose overall shape that may be parametrized by appropriate macroscopic variables.

In our application, we found that the most important aspect was appropriately sampling data to build the diffusion map so that all regions of interest in phase space are well-covered. Another avenue for future research is to find a better method for identifying the underlying data set, for instance by finding better sampling techniques for the initial data and the stopping times.

Chapter Three

JUSTFAIR

This chapter is based on work presented in Ciocanel et al. (2020a).

3.1 Introduction

In the United States, the public's right to access court proceedings and court records is frequently debated and litigated Ardia (2016). However, with respect specifically to criminal trials, the public's right of access is unambiguous. A 7-1 majority of the Supreme Court affirmed this right in the case *Richmond Newspapers, Inc. v. Virginia*, 448 U.S. 555 (1980). The opinion of the plurality held that "the First Amendment guarantees of speech and press, standing alone, prohibit government from summarily closing courtroom doors." Federal court proceedings may be closed in special cases involving classified information, trade secrets, ongoing investigations, or other complications. By default, however, most proceedings are open Administrative Office of the U.S. Courts (2020b), Fenner and Koley (1981).

While members of the public have the right to attend criminal trials, they are not necessarily well-positioned to exercise this right. Any individual could in principle choose to attend a federal criminal court proceeding and observe the defendant, the judge, the crimes allegedly committed, the evidence provided, the verdict, and, if applicable, the sentence given. However, no individual can attend the hundreds of thousands of proceedings that take place in federal courts every year. Stated differently, the public does not have access to high quality, large-scale information about the federal criminal justice system, and therefore, much of what happens in criminal courts remains opaque. The opacity of the federal criminal court system impedes the public in a number of ways, including in its ability to assess sentencing equity. As we explain later, while federal sentencing guidelines that recommend

appropriate penalties exist, they are not binding. Judges have broad discretion in their choice of sentences.

Sentencing disparity is an active area of research in legal scholarship. One stream of this research investigates associations between the severity of sentences and demographic characteristics of defendants. A study of over 77,000 federal offenders finds, after controlling for numerous factors, that judges give longer sentences to individuals who are Black, male, or low-income as compared to members of other racial, gender, and income groups Mustard (2001). More recent work confirms sentencing gaps biased against men and Black individuals Rehavi and Starr (2014), Yang (2015), Starr (2015). For example, Rehavi and Starr construct a multi-agency-based dataset that focuses on offender and offense information. Their data includes almost 37,000 cases of black and white male US citizens who were arrested for violent, property/fraud, weapons, and public order offenses and referred to federal prosecutors for potential prosecution between fiscal years 2006 and 2008. This dataset provides very detailed demographic and arrest/prosecution information for the offenders, but is not linked to judge identity or information. A second stream of research focuses not on characteristics of defendants, but rather, on those of judges Schanzenbach and Tiller (2008), Yang (2014, 2015), Cohen and Yang (2019). For instance, Cohen and Yang (2019) dissects the role of political affiliation in federal criminal cases and finds that as compared to Democratic-appointed judges, Republican-appointed judges give longer sentences to defendants who are Black or male.

While this second, judge-focused stream of research does examine associations between judge demographics and sentencing disparities, it does not shed light on the sentencing decisions of individual judges. As the authors of Schanzenbach and Tiller (2008) explain, "The unavailability of judge data is one of the most frustrating aspects of the study of federal sentencing and has significantly impeded scholarly

evaluation." The U.S. government does publish certain data sets related to federal criminal sentencing, available from the United States Sentencing Commission. However, as Schanzenbach and Tiller (2008) also notes, "the Commission has refused to identify the sentencing judge in the data."

Despite the U.S. government's decision not to identify sentencing judges, studies like Schanzenbach and Tiller (2008), Yang (2014, 2015), Cohen and Yang (2019) have been possible because the investigators have assembled their own databases to identify judges. The painstaking work of Schanzenbach and Tiller (2008) manually assembles a database of 2,265 records using information from the United States Sentencing Commission and the Public Access to Court Electronic Records System. However, while the data set does link individual judges with sentencing decisions, no individual judges are discussed in Schanzenbach and Tiller (2008), nor is the data set publicly accessible. The studies in Yang (2014, 2015), Cohen and Yang (2019) are much larger scale, involving hundreds of thousands of sentencing decisions, each linked to information about the sentencing judge. To build these databases, the authors combine data from government databases with data from a proprietary source, the Transactional Records Access Clearinghouse (TRAC), which provides information about judges' sentencing habits obtained via Freedom of Information Act requests. This dataset consisted of almost 547,000 cases linking federal sentencing data with judge information for defendants sentenced between 1999 and 2015 Cohen and Yang (2019). To connect between defendant and crime characteristics and sentencing judge information, the authors matched records in these databases based on "district court, sentencing year, sentencing month, sentence length in months, probation length in months, amount of total monetary fines, whether the case ended by trial or plea agreement, and whether the case resulted in a life sentence" Cohen and Yang (2019). However, TRAC's assembled data is proprietary and is governed by a data

usage agreement Transactional Records Access Clearninghouse (2002). Presumably for these reasons, the data sets used in Yang (2014, 2015), Cohen and Yang (2019), which incorporate information from TRAC, are not public, and these manuscripts do not disclose information about the sentencing decisions of individual judges.

In summary, the public has a constitutional right to observe criminal proceedings, and in principle could tabulate the sentencing decisions of every federal judge. However, due to practical constraints, individuals cannot exercise this right on a large scale. The U.S. government publishes data about criminal sentencing, but this data does not identify the judge associated with any particular sentencing data. Past legal scholarship has linked sentencing decisions with individual judges, but the proprietary data source used prohibits the identification of these judges.

To remedy this situation, we present JUSTFAIR: Judicial System Transparency through Federal Archive Inferred Records, a database of criminal sentencing decisions made in federal district courts. We have compiled this data set from public sources including the United States Sentencing Commission, the Federal Judicial Center, the Public Access to Court Electronic Records system, and Wikipedia. With nearly 600,000 records from 2001 - 2018, JUSTFAIR is the first large-scale, free, public database that links information about defendants and their demographic characteristics with information about their federal crimes, their sentences, and crucially, the sentencing judges. Our publicly-available database extends the one in Cohen and Yang (2019) to fiscal years 2002-2018 and uses similar variables (district court, sentencing date, prison sentence length, probation length, fine amount, offense codes, etc.) to match records in the appropriate public sources.

The rest of this paper is organized as follows. In Legal Background, we provide a layperson's overview of the federal court system and a brief history of federal criminal

sentencing. In Data Sources and Preprocessing, we describe in detail our procedures for acquiring data from four sources and for preparing that data for merging. In Data Consolidation, we explain how we merge United States Sentencing Commission data first with Federal Judicial Center data to obtain court docket numbers, then with the Public Access to Court Electronic Records system to obtain the initials of the sentencing judge for each case, and finally, with information from Wikipedia (as well as from the Federal Judicial Center) to obtain the full name of each judge. Finally, we conclude with a brief summary of JUSTFAIR and some directions for future work.

3.2 Legal Background

3.2.1 Structure of the Federal Judicial System

For a layperson's review of the federal court system, see, *e.g.*, Administrative Office of the U.S. Courts (2016). In our work, our primary concern is the principal federal trial court system, which comprises 94 district courts. Each U.S. state has one or more federal district courts. For instance, Alaska has one district, whereas California, New York, and Texas each have four districts. Additionally, the District of Columbia and Puerto Rico each have one district court. Altogether, the U.S. states, District of Columbia, and Puerto Rico account for 91 of the 94 aforementioned courts. The remaining three courts are associated with the territories of Guam, the Northern Mariana Islands, and the U.S. Virgin Islands. While these three territorial courts are established in a different part of the Constitution than the 91 districts, they have similar jurisdiction, and so it is common to refer to them as district courts. Nonetheless, there are some differences between these territorial courts and the other district courts. Excluding the three territorial courts, each

district also has a specialized bankruptcy court. Additionally, there are two special courts that have nationwide jurisdiction, namely the U.S. Court of Federal Claims and the U.S. Court of International Trade.

Decisions made by judges in the 94 district courts can be appealed. Each district court belongs to one of twelve circuits, and each circuit has an appeals court. A thirteenth circuit, the Federal Circuit, hears appeals for certain special types of cases. The Supreme Court sits above the federal circuit courts of appeals. While the U.S. Constitution and federal laws mandate special types of cases that the Supreme Court must hear, much of the Supreme Court's case load is chosen by the court itself, whenever at least four justices agree to hear a case. One important role of the court is to resolve situations where two or more circuit courts of appeals issue legal rulings that are in conflict.

Excluding judges on the three territorial courts, federal district judges are chosen as specified in Article III of the Constitution. That is to say, they are appointed by the president of the United States and must be confirmed by a vote of the Senate. These Article III judgeships are lifetime appointments. Once judges are confirmed, they cannot be removed from office except through impeachment by the House of Representatives followed by conviction in the Senate. Judges in the three territorial courts are not governed by Article III of the Constitution, and thus do not have lifetime appointments.

Federal district courts also have magistrate judges. A district magistrate judge is chosen not by the president, nor by Congress, but rather, by a majority vote of the federal judges in the district. Magistrate appointments last for eight years and are renewable. The specific duties of magistrate judges are not prescribed by federal law. Each district court has the authority to delegate to its magistrates "additional

duties as are not inconsistent with the Constitution and laws of the United States" Administrative Office of the United States Courts (2010).

3.2.2 Federal Criminal Sentencing

For an overview of federal sentencing, see United States Sentencing Commission (2018). Prior to 1987, federal district judges had the discretion to impose, nearly without restriction, the criminal sentences they deemed appropriate. Put concretely, two individuals with similar criminal histories who were convicted of committing similar crimes under similar circumstances could receive vastly different sentences depending on the outlook of the presiding judge.

However, in 1984, Congress passed the Sentencing Reform Act as part of a larger piece of legislation, the Comprehensive Crime Control Act. The Sentencing Reform Act established the U.S. Sentencing Commission which, in 1987, formally adopted federal sentencing guidelines in order to mitigate sentencing disparities. These guidelines set uniform sentencing policies for felonies and for the most serious class (Class A) of misdemeanors. The guidelines are a detailed set of instructions for algorithmically calculating the appropriate range of a sentence. The algorithm accounts for the type of crime, aggravating and mitigating factors, criminal history, and more.

These sentencing guidelines were originally interpreted as mandatory. However, in *United States v. Booker*, 543 U.S. 220 (1985), the Supreme Court ruled that these sentencing guidelines cannot be mandatory and must be considered merely as advisory. Today, federal district court judges are required to calculate the sentence recommended by the guidelines, but they can still issue a different sentence that they deem appropriate, subject to mandatory minimum sentences when applicable.

Generally speaking, magistrate judges do not impose criminal sentences. The primary conditions under which a magistrate judge might impose a criminal sentence are when the conviction is for a Class A misdemeanor and the defendant waives their right to be sentenced by a federal district judge.

3.3 Data Sources and Preprocessing

3.3.1 Overview of Data Sources

One of our primary goals is to facilitate study of the criminal sentencing decisions of individual federal district judges. JUSTFAIR includes information about federal criminal cases including:

- demographic characteristics of the sentenced individual,
- sections of the law relevant to the conviction,
- factors influencing the recommended sentence,
- components of the sentence, including (potentially) a fine, probation, and prison time,
- sentencing date,
- federal district of the court, and
- name, time period of service, appointing president of the sentencing judge, as well as other background information.

To assemble JUSTFAIR, we consolidate data from five sources:

- the United States Sentencing Commission Database,
- the Federal Judicial Center Integrated Database,
- the Public Access to Court Electronic Records system,
- Wikipedia, and
- the Federal Judicial Center Biographical Directory of Article III Federal Judges.

Fig 3.1 illustrates the data pipeline used to create JUSTFAIR. First, we obtain detailed information about criminal cases and sentences (pink section of figure) from the United States Sentencing Commission database. This database includes demographic information about convicted individuals and detailed information about sentences given. However, while this database provides the sentencing date and federal judicial district of the case, it does not identify the sentencing judge. To obtain information about the sentencing judge requires a case docket number. We obtain these case docket numbers from the Federal Judicial Center Integrated Database and merge them with the United States Sentencing Commission database (pink section of figure). Then, from the Public Access to Court Electronic Records system, we retrieve the initials of the sentencing judge associated with each docket number (blue section of figure). In order to map judge initials to an identified judge, we retrieve judge names from two different sources (green section of figure). From Wikipedia, we scrape the names of current and former federal district judges and extract their initials. We also use the Federal Judicial Center Biographical Directory of Article III Federal Judges as an alternative means to obtain the names of current judges as well as their demographic, education, and nomination information. In the following

subsections, we describe our five data sources in greater detail, and we explain the procedures we use to preprocess the data and merge it to create the JUSTFAIR database.

3.3.2 United States Sentencing Commission

The United States Sentencing Commission (USSC) is an independent agency of the federal judicial system created by Congress via the Sentencing Reform Act of 1984. USSC's mission includes the directive "to collect, analyze, research, and distribute a broad array of information on federal crime and sentencing issues, serving as an information resource for Congress, the executive branch, the courts, criminal justice practitioners, the academic community, and the public" United States Sentencing Commission (2011). In service of this mission, USSC maintains publicly accessible data sets, including files which provide information about sentences given to individuals in federal district courts. We refer to these as individual offender files (IOFs) to distinguish from other public files that tabulate information on corporate offenders United States Sentencing Commission (2019). USSC provides a detailed description of IOF contents in their code book United States Sentencing Commission (2020).

Seventeen IOFs are available, one each for the fiscal years 2002 - 2018 (inclusive). These files come in .dat format, and range in size from 1.19 gigabytes to 4.07 gigabytes, with a median size of 2.02 gigabytes. In total, the IOFs comprise 35.08 gigabytes. We convert these files to .csv format for convenience. Each file has tens of thousands of records (rows), ranging from approximately 64,000 to approximately 86,000 with a median of approximately 73,000. Each file also has thousands or tens of thousands of variables (columns). The differing numbers of variables from file to file are due to changes in sentencing policies and record keeping over time, as detailed

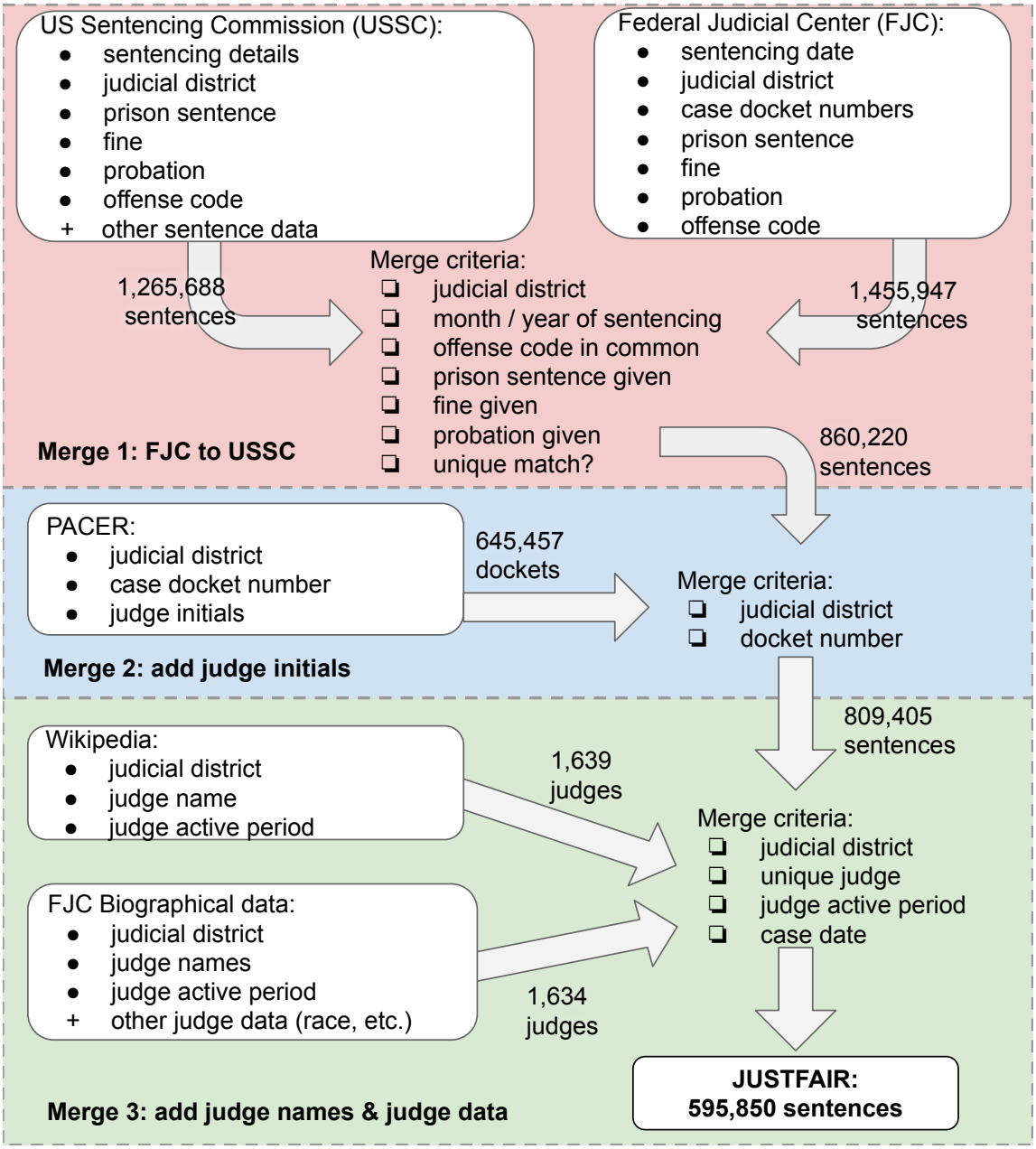

Figure 3.1: **JUSTFAIR data pipeline.** We combine information from the United States Sentencing Commission (USSC), the Federal Judicial Center (FJC), the Public Access to Court Electronic Records System (PACER) and Wikipedia in order to create our database, Judicial System Transparency through Federal Archive Inferred Records (JUSTFAIR). With nearly 600,000 federal district court records from 2001 - 2018, JUSTFAIR is the first large-scale, free, public database that links information about criminal defendants and their demographic characteristics with information about their federal crimes, their sentences, and crucially, the sentencing judges.

in Kitchens (2019a,b), Kyckelhahn (2019), Syckes (2020), Reedt et al. (2013). As we will explain momentarily, a majority of these columns are mostly empty.

To make the data set amenable to study by the public, we reduce it to a more tractable size. First, we construct the set of column names in each IOF and we intersect these sets (ignoring case, which is inconsistent across years). There are 2,384 variable names that are shared across all 17 files. Additionally, there are 16 variables that we retain for possible downstream analysis: `SENTDATE`, `SENTMON`, `SENTYR`, `SENTTOTO`, `BOOKER2`, `BOOKER3`, `BOOKERCD`, `BOOKPOST`, `DEFCONSL`, `OFFTYPE2`, `OFFTYPESB`, and `NWSTAT1 - NWSTAT5`. We eliminate all other variables. We then examine the amount of data present in the retained variables. Many of these columns consist primarily of missing data. We retain variables that have at least 25% of data present, a cutoff chosen by inspection of the histogram of the percentage of data present in each column (not shown). After eliminating the mostly-empty columns from each file, we again intersect variable names across all IOFs. In the previous two steps, however, we force inclusion of the 16 variables enumerated above. Altogether, we identify 163 columns to retain. We reduce all IOFs to these shared columns and concatenate the 17 files. The result is a 0.5 gigabyte .csv file with 1,265,688 records and 163 variables. In our data set, we retain the variable `USSCID` which is a unique identifier in the USSC database. Because we retain this identifier, any user wishing to recover variables we eliminated from the raw IOFs can do so in a straightforward manner.

We make several modifications to our USSC data. First, we reconcile a change in the coding of sentencing dates. For the years 2001 - 2003, the month and year when each sentence was given are coded in one variable called `SENTDATE`. From 2004 onwards, the month and year are stored separately in variables `SENTMON` and `SENTYR`. We extract the months and years from `SENTDATE` in the 2001 - 2003 data to have a

consistent format for dates.

Second, we create modified versions of the variable SENTTOT. This variable stores the total prison sentence given in months. Because of this choice of units, the values of SENTTOT are not necessarily integers. Later, we merge our USSC data with data from the Federal Judicial Center Integrated Database by matching on several variables, including sentence length. However, in the latter database, all values of prison sentence length are integers. Therefore, to facilitate merging, we create variables roundSENTTOT and floorSENTTOT, where the former rounds SENTTOT to the nearest integer and the latter truncates any fractional portion of SENTTOT.

Third, we reconcile SENTTOT with another variable, TOTPRISN (total months of imprisonment, not including time served). TOTPRISN may take on a number of special values, including 9996, and 9998, which correspond to life imprisonment and the death penalty, respectively. We carry these values over to SENTTOT to retain information about these penalties.

Fourth, we clean SENTTOT for certain records. In particular, we note 157,484 records for which SENTTOT has missing data but a closely related variable, SENTTOT0 has the value of zero. According to the USSC data codebook United States Sentencing Commission (2020), SENTTOT codes zero-length sentences as missing, while SENTTOT0 does not. Because Kitchens (2019b) recommends using the variable SENTTOT for analysis, but because we need information about zero-length sentences, we recode SENTTOT as zero for the 157,484 aforementioned records. We also create a new variable, SENTTOTFLAG, which flags the records that have been recoded.

Fifth, we create modified versions of the variables NWSTAT1 - NWSTAT5. These variables record each portion of the U.S. legal code under which a defendant has

been convicted. These codes include, potentially, the relevant title, section, and subsection of the U.S. legal code. For instance, one code in the data set is 181956A1, which refers to a subsection of U.S. Code Title 18, Section 1956 on money laundering. We truncate the `NWSTAT` codes to include only the title and section information so that, for example, 181956A1 becomes 181956. We refer to our truncated variables as `USSCcode1` - `USSCcode5`, and these will validate matching with other data sources downstream.

Finally, we set any empty fields in the database to NA (short for "not available") to represent missing data. After the preprocessing steps described above, our final USSC data set has 1,265,688 records and 171 variables.

3.3.3 Federal Judicial Center Integrated Database

The Federal Judicial Center (FJC) is an agency of the U.S. federal judicial system, established in 1967 with a tripartite mission that includes "research and study of the operation of the courts." To this end, the FJC maintains and provides public access to an Integrated Database Federal Judicial Center (2020) consisting of seven files with information on civil cases, criminal cases, appeals cases, and bankruptcy cases. The criminal cases are spread across two files, the second of which covers the time period from 1996 to present. We download this data, which FJC provides as a 2.7 gigabyte tab-delimited text file Federal Judicial Center (2019). The raw file comprises 4,783,129 records and 144 variables. FJC provides a detailed description of file contents in their code book Federal Judicial Center (2016).

We make several modifications to the FJC data that we download. First, we create new variables for sentencing date. The raw variable `SENTDATE` encodes the

day, month, and year of sentencing. From it, we extract two new variables, `SENTMON` and `SENTYR`, which store the month and year separately for commensurability with USSC data downstream.

Second, because USSC data begins in calendar year 2001, we restrict our FJC data to this same time period. This restriction drastically reduces the number of records. In fact, the raw FJC file not only includes data from prior to 2001, but it includes over three million records whose sentencing dates are listed as January 1, 1900. We exclude these records. By retaining FJC records with sentencing dates in 2001 or later, we reduce the data set to 1,455,947 records.

Third, we address issues with the coding of the federal judicial district for each record as stored in the variable `DISTRICT`. The coding of judicial districts in FJCIDB differs from that in USSC for eight districts, namely the Middle District of Florida, the Southern District of Florida, the Northern District of Georgia, the Middle District of Georgia, the Southern District of Georgia, the Eastern District of Louisiana, the Middle District of Louisiana, and the District of Alaska. In the FJC Integrated Database, these are coded as 3A, 3C, 3E, 3G, 3J, 3L, 3N, and 7-, respectively. We replace these codes with the numbers 30 - 35, 96, and 95, respectively, to be commensurate with the `DISTRICT` variable in USSC. Additionally, values of -8 signify missing information, and we replace these with NA.

Fourth, we recode the sentencing variable `PRISTOT`, an integer specifying the total prison sentence given in months. While most values are positive integers, the data also contains certain special negative values. Values of -1 signify imprisonment of less than one month and values of -2 signify that no prison sentence was imposed. We replace these values with zeros for commensurability with USSC data. Values of -3 represent sealed sentences, and we replace them with NA since it is not possible

to know the sentence. Values of -4 and -5 represent life imprisonment and the death penalty, respectively. We replace these values with 9996 and 9998 to match the USSC database. There are also 38,375 records with PRISTOT equal to -8, signifying missing data. For these records, we examine the variables PRISTIM1 - PRISTIM5 which give the amount of prison time to which a defendant was sentenced for up to five sections of the U.S. criminal code. For all 38,375 records, the value of PRISTIM1 is zero, and the values of PRISTIM2 - PRISTIM5 are either -8 (signifying missing data) or zero. In summary, the PRISTIM variables for these records do not indicate any prison sentence, and so we set the value of PRISTOT to be zero for them. We also create a new variable, PRISTOTFLAG, which flags the records that have been recoded.

Fifth, we recode the sentencing variable PROBTOT, which specifies the number of months of probation given with a sentence. We perform this recoding in a manner very similar to that used above for PRISTOT. Values of -1 represent probation of less than one month and we replace these with zeros for commensurability with the USSC data. Values of -8 represent missing data, of which there are 353,000 occurrences. For these records, we examine the variables PROBMON1 - PROBMON5 which give the amount of probation time to which a defendant was sentenced for up to five sections of the U.S. criminal code. For all 353,000 records, the value of PROBMON1 is zero, and the values of PROBMON2 - PROBMON5 are either -8 (signifying missing data) or zero. In summary, the PROBMON variables for these records do not indicate any probation time, and so we set the value of PROBTOT to be zero for them. We also create a new variable, PROBTOTFLAG, which flags the records that have been recoded.

Sixth, we create ten new variables, FJDcode1 - FJDcode10, which are modified versions of the variables FTITLE1 - FTITLE5 and TTITLE1 - TTITLE5. The first five variables record portions of the U.S. code under which a case was filed, and the latter five record portions of the U.S. code under which a case was disposed.

For commensurability with the `USSCcode1` - `USSCcode5` variables we created in our preprocessed USSC database, we eliminate punctuation from these strings and reduce them to include only the title and section of the U.S. code.

Finally, we set any empty fields in the database to NA to represent missing data. After the preprocessing steps described above, our FJC data set contains 1,455,947 records and 158 variables.

3.3.4 PACER and Juriscraper

Public Access to Court Electronic Records (PACER) is the records access service of the United States federal courts. This database was established in 1988 and was made publicly accessible online beginning in 2001 Johnson (2009). For each court case it contains, PACER includes a breadth of information ranging from a roster of involved parties, to chronologies of the judicial process, to scanned court documents, to, potentially, judicial opinions issued. PACER content for each district court is maintained by that court on its own website. However, there is a centralized PACER Web portal Administrative Office of the U.S. Courts (2020a) with a search feature that connects to individual court sites. Currently, access to PACER records is for-fee at the rate of $0.10 per page of information accessed.

Our queries to PACER require us to have a case docket code. Constructing this code requires the variable `CASLGKY`, which originates in the FJC data. `CASLGKY` is a 15 character string of the form "CCDDOYYSSSSSTTR" where:

- Characters CC are digits specifying the federal judicial circuit of the court where the case occurred.

- Characters DD are digits or letters specifying the federal judicial district.
- Character O is a digit or letter indicating which office within the judicial district handled the case.
- Characters YY are the last two digits of the year in which the case was filed.
- Characters SSSSS represent the five-digit sequence number of the case (essentially, an ID number assigned to the case by each court).
- Characters TT are either "CR" or "MJ," specifying, respectively, a criminal case handled by a federal district judge or one handled by a magistrate judge.
- Character R is the one-digit reopen sequence number indicating whether the record corresponds to an original court proceeding or the re-opening of a previous proceeding.

From `CASLGKY` we construct a federal docket number. Each docket number has the form "O:YY-CR-SSSSS" where the characters O, YY, and SSSSS are as defined above, and "CR" indicates that we are searching only for cases handled by district judges. We use our docket number and the `DISTRICT` variable to query Juriscraper, a tool of the FreeLaw Project Free Law Project (2020). Juriscraper provides an API to retrieve metadata from district court websites.

A Juriscraper query on a docket number returns a case title. This case title begins with a version of the docket number, followed by information. The version of the docket number that is returned may include additional letters that we did not submit as part of our search query. For the vast majority of districts, these are the initials of the judge (we discuss exceptions later). For example, we use Juriscraper to query the docket number 2:01-CR-00071 to the District of Maine. This is a criminal

case from the second court within the district, filed in 2001, having sequence number 00071. The query returns PACER's full case title, which is "2:01-cr-00071-DBH USA v. BLAKE (closed 01/03/2002)." The appended letters "DBH" are the initials of a judge within the district, in this case, Hon. David Brock Hornby. Some PACER case titles include two sets of initials, for example, "1:00-cr-00425-RWR-AK." We store the first set of initials in the variable `pacer_first_inits`. If the second set of initials is present, we store it in variable `pacer_second_inits`.

In cases that have multiple defendants, a single Juriscraper query may return more than one result. In fact, we submit 744,401 queries to Juriscraper (see Merging Federal Judicial Center Integrated Database), and these queries return 3,394,237 results with 645,457 unique values of `CASLGKY`.

3.3.5 Wikipedia

To obtain information about federal district judges, we scrape Wikipedia. We begin with Wikipedia's page containing a master list of U.S. district courts Wikipedia contributors (2020), and from it, we retrieve the link to each individual court's Wikipedia page. These individual court pages each contain tables with information about the judges currently and formerly on the court. We scrape these tables, retaining information such as judge name, lifespan, years active, and link to the Wikipedia page for the judge (if available).

Eventually, we will use information about judges' initials from PACER records in order to infer the full name of the judge. To enable this inference, we must extract judge initials from our scraped Wikipedia data. Before performing this extraction, we clean the scraped judge names. First, we remove strings in judge names that

do not contribute to the judge's initials. These strings include Roman numerals "I" through "V" as well as "Jr.," "Sr.," and footnotes or numbers such as "[1]" that were present in some scraped records. Second, we deal with apostrophes in names, based on a manual investigation of judge initials within PACER. For instance, in PACER, Hon. Beverly Reid O'Connell has initials "BRO." Thus, we eliminate all apostrophes from our Wikipedia names. Third, we deal with hyphens in names, also based on manual investigation. For instance, Hon. Martin Leach-Cross Feldman has initials "MLCF" in PACER. Thus, we replace hyphens in Wikipedia names with spaces. Fourth, we remove parenthetical nicknames appearing for some judges, such as Hon. Wilhelmina Marie (Mimi) Wright. Fifth and last, we manually edit a small number of entries. Specifically, we encode Hons. James Arnold von der Heit, John W. deGravelles, and Frederick van Pelt Brian as "JAH," "JWG," and "FPB," respectively. With these data cleaning procedures complete, we extract initials from each judge's name simply by breaking each name into space-separated tokens and retaining their first letters. We record these initials in the variable `judge_inits`.

Some judges have a middle name listed on their individual Wikipedia page, but not in the table on their district's Wikipedia page. For example, Hon. Abdul Karim Kallon's name is given simply as Abdul Kallon in the table on the Wikipedia page for the Northern District of Alabama. To account for these cases, we access all available individual district judge Wikipedia pages (linked from the tables in the courts' Wikipedia pages) and scrape each judge's full name. We extract initials from these names in a manner similar to that used above. This second version of a judge's initials could be identical to the first version, could be a more complete version that incorporates additional middle names, or could be "N" for Null if that judge does not have an individual Wikipedia page (we record this in the variable `judge_inits_full`).

Finally, we add to our data set the federal judicial district in which each judge serves or served. There are in fact two numbering schemes for judicial district used in the USSC data, namely, the variables `DISTRICT` and `CIRCDIST`; see United States Sentencing Commission (2020). We add both codes to each judge record in our Wikipedia data set. In total, our scraped Wikipedia data comprises 3,140 former and current district judges, but after filtering the data for judges who were active from 2001 to present, 1,639 judges remain.

Federal Judicial Center Biographical Directory of Article III Federal Judges

An alternative source for information on federal district judges is a database maintained by FJC Federal Judicial Center (2013). This database, called the Biographical Directory of Article III Federal Judges, provides a publicly-available .csv file where each row is the record for an individual judge. The record includes information about the judge's demographics, federal judicial service, education, professional career, confirmation process, and more Federal Judicial Center (2013).

For consistency with the databases described above, we extract the `CourtName` field for each judge and convert it to the `CIRCDIST` numbering scheme for these judicial districts. Because a small number of judges serve for two districts in their state, we duplicate the rows for these judges and record the corresponding `CIRCDIST` for each entry. We also add variables for the overall starting and ending years of service of each judge, based on information recorded in the `CommissionDate` and `TerminationDate` fields in this data set. Finally, since the `LastName`, `MiddleName`, and `FirstName` are separate in this data set, we extract the initials from each name

and concatenate them into an `Initials` field.

This data set of judges consists of 3,799 former and current district judges. After filtering the data for judges active after 2001, 1,634 judge entries remain. A difference between this FJC data set and the Wikipedia judges data set described above is that the FJC judge data set does not include judges from the three territorial courts since they are not governed by Article III of the Constitution. In addition, both the Wikipedia and the FJC judge data sets each contain entries of several judges (not included in the other data set) who have not yet been commissioned and will begin service in 2020. These judges will not appear in our final data set since our USSC data is restricted to fiscal years 2002-2018.

3.4 Data consolidation

3.4.1 Merging Federal Judicial Center Integrated Database

Each record in the USSC database has a unique identifier stored in the variable `USSCIDN`. However, this identifier is not used in the FJC Integrated Database. Therefore, to link information, we use an approach established in Yang (2014, 2015), Cohen and Yang (2019). We perform a data merge on critical variables that are shared between these two data sets, namely federal judicial district, month of sentencing, year of sentencing, prison sentence given, fine given, and probation given. Within our preprocessed USSC data, this information is stored in the variables `DISTRICT`, `SENTMON`, `SENTYR`, `SENTTOT`, `FINE`, and `PROBATN`, respectively. Within our preprocessed FJC data, it is stored in `SENTMON`, `SENTYR`, `DISTRICT`, `PRISTOT`, `FINETOT`, and `PROBTOT`.

Records that have values of NA for any of these variables cannot be matched, and so we drop them from our data sets. These deletions reduce the USSC data from 1,265,688 records to 1,245,827 records, and they reduce the FJC data from 1,455,947 records to 1,452,288 records.

We then carry out a basic merge of the two data sets across the aforementioned six core variables, retaining only unique matches. That is to say, for each record in the USSC data, we check if there is exactly one record in the FJC data that has the same federal district, sentencing month and year, and prison, fine, and probation penalties. If a match is made, we merge the records. As an additional step to validate the matching, we require that for any merged record, at least one of the values of `USSCcode1 - USSCcode5` must be included in the values of `FJDcode1 - FJDcode10`. That is to say, we check that the two data sets have at least some overlap in how they report the parts of the U.S. code relevant to the court case. After carrying out these steps, we successfully merge 658,891 records, leaving 586,936 USSC records unmatched.

Recall that in the USSC data, the prison sentence variable `SENTTOT` is real-valued while in the FJC data, the prison sentence variable is integer-valued. Therefore, for the remaining unmatched records in USSC, we repeat the merging steps described above, but using the variable `roundSENTTOT` rather than `SENTTOT` in the USSC data. This procedure produces an additional 145,237 matches for a total of 804,128 matches, leaving 441,699 USSC records unmatched.

We repeat the merging procedure yet again, but using the variable `floorSENTTOT` rather than `SENTTOT`. This procedure produces an additional 56,092 matches for a total of 860,220 matched USSC records. There remain 385,607 USSC records that we are unable to match.

3.4.2 Merging PACER

Having merged USSC and FJC data, we now integrate judge's initials from PACER to each case record. To understand how we perform this merge, it is important to remember two facts. First, the variable `CASLGKY` in our merged USSC/FJC carries, for our purposes, the same information as the pairing of district and docket number that we use to query PACER. In the discussion below, we explain our merging procedure using `CASLGKY`. Second, the basic units of data in USSC/FJC and in our PACER data are different. In USSC/FJC, each record describes the sentencing outcomes for an individual. However, criminal cases may contain more than one defendant, meaning that `CASLGKY` need not be unique within our USSC/FJC data. In fact, we find 86,251 values of `CASLGKY` that are not unique. The most frequently appearing one occurs 89 times, indicating a criminal case that had 89 different defendants (as verified through inspection of the defendant number variable, `DEFNO`). Equally, as mentioned before, searching for a particular case in PACER may return multiple records due to multiple defendants.

Thus, we merge the USSC/FJC data with PACER data in the following manner. First, we select a record from the USSC/FJC data. We search our PACER results for every record with the same value of `CASLGKY`. If we find that these PACER records all have the same judge initials, we assign these initials not just to the record we selected, but to any other record in our USSC/FJC data that shares the same value of `CASLGKY`. In other words, when there are multiple defendants, we are only able to assign judge initials to USSC/FJC records when all defendants are sentenced by the same judge. Otherwise, we are not able to assign judge initials, as there is no reliable to way to map a particular defendant in USSC/FJC to a particular defendant in PACER.

Our merging procedure produces the following results. We begin with the 860,220 records in our USSC/FJC data, which comprise 697,555 unique values of `CASLGKY`. In our 3,394,237 PACER records, there are 645,457 distinct values of `CASLGKY`. Of these, 524,393 are unique, and as a result, give us unique judge initials. An additional 106,436 values of `CASLGKY` appear more than one time in our PACER data but still have unique judge initials. Overall, when we merge our USSC/FJC data with our PACER data, we are able to assign initials to 809,405 of the 860,220 records in the former.

3.4.3 Merging Wikipedia

Our next merging step involves matching the judges' full names to their initials for each defendant case in the USSC/FJC/PACER data set. Recall that our scraped Wikipedia data includes each judge's district and two versions of the judges' inferred initials: one from the district table and another from their individual Wikipedia page, if available. Therefore, we merge the full names to initials by district. In particular, we use the variable `CIRCDIST` from USSC United States Sentencing Commission (2020), which we have also added to the Wikipedia judges data set.

There are two primary cases to consider: first we match judges with initials that are unique within their district, and second, we match judges with initials that are not. Since our data set covers criminal cases after 2001, we identify only 15 duplicated judge initials out of 1,639 judge initials-district pairs. In addition, we also take into account the fact that Juriscraper queries return two sets of judge initials in the PACER full case title in 187,799 cases of our 809,405 records. For instance, the case title "1:00-cr-00425-RWR-AK USA V. Hill et all (closed 04/26/2002)" gives initials "RWR" and "AK" for federal judges (and potentially magistrates) assigned to this

case. Before proceeding with the two merging stages accounting for each case, we separate the USSC/FJC/PACER data set into cases where judge initials correspond to unique judge initials in a district (comprising the vast majority of cases), and cases where they correspond to duplicated judge initials in a district.

In the first merging stage, we include only unique judge initials-district combinations from the Wikipedia data set. We first merge USSC/FJC/PACER data with the judges data based on the variables `CIRCDIST` and the first set of judge initials from PACER. This merge uses the variable `judge_inits` in the Wikipedia data set and yields 552,878 matches. We repeat the same merging process for the remaining unmatched cases but instead use the second form of the initials from the Wikipedia data set, `judge_inits_full`. This second merge yields an additional 47,629 matches. We then repeat the above steps for the remaining unmatched cases but instead use variables `CIRCDIST` and the second set of judge initials from PACER. As before, we first use the variable `judge_inits` in the Wikipedia data set, which yields 262 additional matches. We then use the variable `judge_inits_full` in the Wikipedia data set, which yields an additional 19 matches.

In the second merging stage, we separately match each duplicate judge initials-district combination in our Wikipedia data set with the USSC/FJC/PACER data set. Since there are two judges corresponding to one district-judge initials pair, we determine factors that allow us to distinguish between cases assigned to either of the judges. These factors include judge activity periods that do not intersect, availability of data only for one of the judges' activity periods, and judges that have only served from 2019 on, a period which is not included in our data. In several duplicate cases, one of the judges has an additional middle name (provided in their individual Wikipedia page and stored in `judge_inits_full`), so we verify that they already were matched to USSC/FJC/PACER cases in the first merging stage. We

then assign the remaining criminal cases associated with this judge initial-district combination to the second judge. There are several exceptions we encounter when dealing with these duplicate judge initials-district combinations. For instance, one of these combinations does not appear in our USSC/FJC/PACER data set. There is also one duplicate combination where we could find no distinguishing factor: there are two judges with initials "CAB" in the Southern District of California, Hons. Cathy Ann Bencivengo and Cynthia Ann Bashant, with largely overlapping activity periods. There is also a duplicate combination where we are only able to assign cases to each of the judges outside the intersection of their activity periods. Finally, we find an additional duplicate judge initials-district pair when considering the full judge initials from Wikipedia individual pages.

Overall, there are 600,788 matches of the USSC/FJC/PACER data set with judges from Wikipedia based on unique judge initials-district pairs and 6,985 additional matches from the duplicate judge initials-district combinations. By analyzing the cases that were not matched to judges in this merge, we identify eight districts for which none of the cases have matches to judges from our Wikipedia data set. In six of these districts (the Southern District of West Virginia, the Southern District of Texas, the District of the Northern Mariana Islands, the District of Guam, the Northern District of Illinois, and the Middle District of Tennessee), no judge initials are reported in PACER, making the connection to the judges' identities impossible. In the Eastern District of North Carolina, almost 45% of the cases report only one letter for judge initials and therefore cannot be matched. The majority of the remaining cases have letter combinations that do not correspond to existing judge names. We thus conclude that the letters in this district's docket numbers may correspond to other notes about the criminal case. Similarly, very few matches are possible for the Northern District of Texas, where almost 96% of the cases report

only one letter for judge initials, and in the Western District of Oklahoma, where almost 83% of the cases report only one letter for judge initials. Finally, in the District of New Hampshire, we find that the judge initials reported in PACER do not include the judges' middle initials. We address this issue by removing the middle initials from variable `judge_inits_full` and are able to match an additional 2,685 cases, corresponding to 99.2% of this district's cases in our data set.

We complete an additional validation step by checking whether variable `SENTYR` in this final merged data set is within the assigned judges' activity period. We find that 589,433 of the cases satisfy this requirement, and we accept these as the product of this stage of our work.

3.4.4 Merging Federal Judicial Center Biographical Data

Rather than using Wikipedia to map judge initials to names, an alternative approach is to use information from the FJC biographical database. This approach has the advantage that the resulting data set contains additional information about federal judges. At the same time, a disadvantage is that judges from the three U.S. territories are excluded because they are not Article III judges. In addition, this data set only includes one set of judge initials. As in the case of merging with the Wikipedia data set, we also use the district variable `CIRCDIST` from USSC to merge the judge full names to their initials.

We restrict the data set to judges active after 2001 and identify 13 duplicates out of 1,634 judge initials-district pairs. These duplicate pairs are a subset of the 15 duplicate pairs we identified in the Wikipedia data set. As in the previous section, we consider separately judges whose initials are unique within their district and those

whose initials are not.

In the first merging stage we include only unique judge initials-district combinations from the FJC judge data set. We merge USSC/FJC/PACER data with the judges data based on the variables `CIRCDIST` and the first set of judge initials from PACER, which yields 584,000 matches. We then repeat this merge for the remaining unmatched cases but instead use variables `CIRCDIST` and the second set of judge initials from PACER. This process yields 241 additional matches.

In the second merging stage, we separately match each duplicate judge initials-district combination in our FJC judge data set with the USSC/FJC/PACER data set, restricted to cases where judge initials are duplicated in a district. Since the duplicate judge initials-district combinations in this data set are a subset of the ones identified in the Wikipedia data set, we use the same factors described in the previous section to distinguish between cases assigned to either judge.

Overall, we find 584,241 matches of the USSC/FJC/PACER data set with judges from the FJC biographical data based on unique judge initials-district pairs, and 6,196 additional matches from the duplicate judge initials-district combinations. As in the previous section, we treat matches for the District of New Hampshire separately, since judge initials reported in PACER do not include middle name initials. By removing the middle initials from variable `Initials` in the FJC judge data set, we are again able to match an additional 2,685 cases to judges in this district.

To validate the combined matched data set, we check whether variable `SENTYR` is within the corresponding judges' activity period. We find that 573,171 of the cases satisfy this requirement, and therefore these cases are the final product of this merging step.

3.4.5 Finalizing JUSTFAIR

We combine the results of the previous two merging steps to provide a final data set that links cases and defendant information to judge information from both the Wikipedia district court pages and the FJC biographical data. In comparing the matching of each USSC/FJC/PACER defendant case to judges' full names and characteristics using Wikipedia versus FJC judge data, we find only 12 discrepancies. These discrepancies all come from cases where PACER records contain two sets of initials; see PACER and Juriscraper. For example, a case in the Eastern District of Michigan is assigned initials "AC-LJM" in PACER. The first set of judge initials, "AC," corresponds to Hon. Avern Cohn (as recorded in Wikipedia). In the FJC biographical data, this judge is recorded as Avern Levin Cohn, and therefore the merge cannot be done using the "AC" initials stored in `pacer_first_inits`. Instead, this case gets merged to Laurie Jill Michelson. Since `pacer_first_inits` is more likely to correspond to the federal judge assigned to the case, we removed 8 cases from the data set merged with Wikipedia and 4 cases from the data set merged with FJC, which correspond to merges using the second set of judge initials, `pacer_second_inits`.

Of the 589,433 cases matched with judges from Wikipedia, 22,683 cases are not matched to any judge in FJC. These cases correspond to 45 judges, 3 of whom serve on the District Court of the Virgin Islands. These three judges cannot be found in the FJC data, which only lists Article III federal judges. For the remaining 42 judges, the initials extracted from Wikipedia are consistent with how these judge initials are recorded in PACER; however the initials extracted from FJC for those judges are not. This inconsistency can occur because of a missing middle name, additional middle name, different writing of the last name, or similar issues. For example, a

federal judge in the Eastern District of Pennsylvania is recorded in Wikipedia as Hon. Franklin Van Antwerpen, whereas in the FJC biographical data their name is recorded as Franklin Stuart Van Antwerpen. Similarly, of the 573,171 cases matched with judges from FJC, 6,425 cases are not matched to judges in Wikipedia. All these judges have slight differences in initials as compared to the same judges in the Wikipedia data set.

We assemble the final data set so that each row provides the case, sentencing, and defendant information, as well as judge information extracted from both Wikipedia and from the FJC biographical data (when available). This combined data set contains 595,850 cases and 557 variables. We add a final variable (column) called `DataSource` to this data set, which takes the value 1 if the case is matched with both judge data sources, value 2 if the case is matched with the Wikipedia court page judge source only, and value 3 if the case is matched with the FJC biographical data judge source only.

3.5 Data Quality and Validation

As described in previous sections, in the creation of our JUSTFAIR data set, we take several measures to ensure data quality. When merging the USSC and FJD databases, we validate merged records by checking for common offense codes. When assigning names of judges to sentencing records based on initials, we use two independent sources. And finally, we require that the date of sentencing for a case fell within the time period during which the inferred judge was active.

We now perform a validation procedure on our data in order to estimate accuracy.

We measure the accuracy of judge inferences using bulk data available from the FreeLaw CourtListener project Free Law Project (2020). CourtListener provides free, archived access to the text of 3.6 million judicial opinions collected from PACER. To create our validation set, we randomly sample CourtListener records for which the docket number includes the string "CR," suggesting a criminal case, and for which there is an identified judge in the record. After using a small number of records to iteratively fine-tune our validation process and data pipeline, we randomly select 400 records to use as a validation set. In the discussion below, it is important to note that judicial opinions are typically not directly associated with criminal sentencing proceedings. Instead, judicial opinions are associated with procedural issues that occur during the case, such as whether certain evidence is admissible. Thus, an underlying assumption of our judge identification procedure (which we will see is occasionally violated) is that the judge involved in a proceeding is the same as the sentencing judge.

To validate based on a CourtListener record, we first search our scraped PACER results for a record with a matching district and approximately matching case title. If we obtain such a match, we can infer a `CASLGKY` identifier which then allows us to locate the case in JUSTFAIR. However, there are many reasons that a sampled CourtListener record might not be found in JUSTFAIR. It is possible that the information about the case could not be unambiguously matched during one of our merging steps, and hence is not included in our final data. Alternatively, the record sampled from CourtListener might in fact not be a criminal case despite the appearance of the string "CR" in its docket number. Or, despite the proceeding being a criminal matter, we may not be able to locate a sentencing associated with it. Finally, the CourtListener record might be related to a sentencing that is outside the time period of our study.

For the CourtListener records for which we can locate a related record in JUSTFAIR, we check whether the CourtListener judge for the procedural opinion matches the JUSTFAIR judge for the criminal case associated with the procedural opinion. If there is a match, we accept our JUSTFAIR record as correct. If there is not a match, we attempt to use alternative methods including PACER searches and news searches to see if the JUSTFAIR judge might nonetheless be correct. For instance, a minor procedural matter in a criminal case might be handled by a magistrate judge who would be listed in the CourtListener record, and yet the sentencing decision would be made by a full-fledged district judge.

We sample 400 records from CourtListener. Of these, 243 do not appear in JUSTFAIR for the reasons mentioned above. For the 157 CourtListener validation cases that do appear in JUSTFAIR, JUSTFAIR has the correct judge for 154 of them. Our three incorrect cases merit some discussion.

One of our incorrect inferences pertains to *United States v. Christy* from the District of New Mexico. The sampled record is an opinion for a procedural matter related to suppression of evidence in a child pornography and sexual exploitation trial. The opinion was authored by Hon. James O. Browning, and a search of the news indicates that he was, in fact, the sentencing judge. In JUSTFAIR, we have listed the judge as Hon. James Aubrey Parker. Parker was indeed involved with this criminal case, but only in post-conviction proceedings that occurred several years after sentencing. Thus, our inference is incorrect.

A second incorrect inference is for *United States v. Rajaratnam*, a high-profile securities fraud case from the Southern District of New York. The sampled record is an opinion from a post-conviction proceeding in which the defendant moved for acquittal on procedural grounds. The opinion was authored by Hon. Richard J.

Holwell, and a news search confirms that Holwell was the sentencing judge. However, in JUSTFAIR, we have listed the judge as Hon. Loretta A. Preska, who was involved with the case only in post-conviction proceedings some years later. This is an error similar to the one discussed above for the *Christy* case.

Finally, our third incorrect inference is for *United States v. Council*, a firearm possession case from the Eastern District of Virginia. The judge listed in CourtListener is Hon. James R. Spencer, who, indeed, was the sentencing judge. JUSTFAIR, however, infers the judge to be Hon. Henry E. Hudson. As with the previous two incorrect cases, Hudson became involved with the case several years after sentencing.

In summary, we sample 400 cases from CourtListener and find 157 of them in JUSTFAIR. Of these 157, we correctly infer the sentencing judge for 154 of them. The three incorrect cases (which we have manually corrected in our data) all correspond to situations where a judge other than the sentencing judge became involved in proceedings post-sentencing. Overall, our measured case-to-judge accuracy rate is 98.1%. A standard statistical calculation yields that a 95% confidence interval for the accuracy rate within the entire JUSTFAIR data set is 94.5% to 99.6%.

3.6 Conclusion

With nearly 600,000 records, JUSTFAIR is the first large scale, free, public database that links information about defendants and their demographic characteristics with information about their federal crimes, their sentences, and crucially, the identity of the sentencing judge. We have inferred the sentencing records in JUSTFAIR by using data from the United States Sentencing Commission, the Federal Judicial Center,

the Public Access to Court Electronic Records system, and Wikipedia. We have performed manual data checks to validate our inferences of sentencing judge, and we estimate 98% accuracy within our data set.

We remind the reader of some limitations of our data. First, merging information from the United States Sentencing Commission and the Federal Judicial Center requires matching the federal judicial district and sentencing date of a case and the prison time, probation, and monetary fine imposed on the defendant. This matching is possible, then, only when the aforementioned combination of variables is unique. Second, our judge identification procedure depends on strings of letters obtained from querying PACER. If those letters are not available, or if they do not represent judge initials, we cannot infer a judge. As discussed previously, the result is that JUSTFAIR does not contain cases from the Southern District of West Virginia, the Southern District of Texas, the Northern District of Illinois, the Middle District of Tennessee, the Eastern District of North Carolina, the District of the Northern Mariana Islands, and the District of Guam, and contains limited results for the Northern District of Texas and the Western District of Oklahoma.

The United States Constitution guarantees the public's right to access criminal court proceedings. At present, the judicial branch releases criminal case data in a manner that obfuscates the identity of sentencing judges. Perhaps ironically, our two primary data sources, the United States Sentencing Commission and the Federal Judicial Center, are physically co-located in the Thurgood Marshall Federal Judiciary Building in Washington, D.C. Before JUSTFAIR, the information necessary to identify sentencing judges lived behind a paywall and was therefore inaccessible to many people. We have made the JUSTFAIR data set public Ciocanel et al. (2020a) and we have created an interactive online visualizer for exploration of the data Cart (2020). By doing so, we hope to correct the inequity of access to information about

sentencing judges.

While the government has chosen to release case data piecemeal, the JUSTFAIR project demonstrates that it is possible to put the pieces back together. Given this capability, we encourage the government simply to release full data with identified judges. Doing so would provide the public with a more complete and accurate data set than JUSTFAIR, resulting in increased transparency.

Beyond providing insight into our federal criminal justice system, JUSTFAIR enables a more granular examination of federal sentencing disparities than has previously been possible. Moving forward, we anticipate analyzing district level and individual judge level data to assess the presence or absence of sentencing disparities, especially along racial lines.

Chapter Four

Judicial Racial Disparities in Federal Criminal Sentencing

This chapter is based on work presented in Smith et al. (2021).

4.1 Introduction

Existing research shows that the U.S. criminal justice system is rife with racial inequality and suggests that part of this inequality operates at the sentencing stage (Commission, 2018). Average racial disparities are substantial for criminal sentencing, with federal judges giving Black and Hispanic defendants harsher sentences than similar white defendants. These disparities are especially apparent among men: according to the U.S. Sentencing Commission, federal judges give white men sentences 19% and 5% shorter, respectively, than the sentences they give Black and Hispanic men who commit similar crimes.

Still, average racial disparities do not express the variability in judges' racially disparate sentencing patterns. Presumably, the conditional difference in sentences across racial groups is not the same for every judge.

We answer four primary questions about federal criminal cases: (1) What is the average disparity between the sentences of observationally equivalent Black and white defendants? (2) What is the average disparity between the sentences of observationally equivalent Hispanic and white defendants? (3) How much do conditional Black-white sentencing disparities vary across judges? (4) How much do conditional Hispanic-white sentencing disparities vary across judges?

4.1.1 Literature on Aggregate Disparities

Racial disparities are a common and well-researched issue in the sentencing practices of U.S. federal courts. Most research on aggregate disparities in federal sentencing has focused on the disparity between Black and white defendants. What research does exist on other groups finds that Asian American defendants are sentenced similarly to white defendants (Johnson and Betsinger, 2009). Due to unique federal jurisdiction in Indian Country, Native American defendants face much harsher sentences for crimes prosecuted federally than defendants of other races who commit them elsewhere and who would only face state prosecution (Droske, 2008). On the whole, though, Native American defendants do not receive harsher sentences than other defendants in the federal courts (Ulmer and Bradley, 2018). Young Native men, though, receive the harshest sentences of almost any group (Franklin, 2013). Black defendants consistently receive harsher sentences than white defendants (Feldmeyer and Ulmer, 2011, Mustard, 2001, Rachlinski and Wistrich, 2017). Young Black men, in particular, are sentenced most harshly (Doerner and Demuth, 2010). Yang (2015) finds that Black-white disparities in sentencing increased after *United States v. Booker*, in which the Supreme Court held the U.S. Sentencing Guidelines to be advisory rather than mandatory. However, others find that no such increase occurred (Starr, 2013). While scholars debate whether *Booker* increased Black-white disparities in sentencing, there is little debate on whether such disparities have existed both before and after the decision. These disparities also show up in state sentencing (Abrams et al., 2012) and state bail decisions (Arnold et al., 2018).

The literature on Hispanic-white sentencing disparities is more variable. Young Hispanic men do receive particularly harsh sentences (Doerner and Demuth, 2010). Still, some argue that the Hispanic-white sentencing disparity is caused by the

harsher sentences imposed on noncitizens since Hispanic defendents are disproportionatly noncitizens (Light, 2014, Light et al., 2014).

4.1.2 Literature on Interjudge Differences in Disparities

In addition to aggregate disparities, this paper addresses *differences across* federal judges in their sentencing disparities based on race. The literature on interjudge differences in racial disparities is much sparser than the literature on aggregate racial disparities. Some relevant evidence exists among judges at a lower level. Namely, judges in the Circuit Court of Cook County: in terms of interjudge differences in racial disparities, Abrams et al. (2012) find that Cook County judges differ substantially vis-à-vis incarceration rates but not vis-à-vis sentence lengths. The field also knows how much judges differ in terms of overall sentencing severity. Both Scott (2010) and Yang (2014) find that interjudge differences in overall sentencing severity are wide and that the differences have generally grown Post-*Booker*.

4.2 Methods

4.2.1 Data

4.2.1.1 Source

We analyze the JUSTFAIR (Judicial System Transparency through Federal Archive Inferred Records) database of criminal sentencing decisions in federal courts (Ciocanel et al., 2020b). JUSTFAIR is compiled from five public sources, includes almost

600,000 records from fiscal years 2001-2018, and links information about defendants, their federal crimes, their sentences, and the sentencing judges as summarized in Section 3.4. Crucially to the current work, JUSTFAIR also includes the sections of the law relevant to the conviction and factors influencing the recommended sentence. Because JUSTFAIR is a new database, we describe its limitations in detail in Appendix A.

We have extended the JUSTFAIR data pipeline to include the available 2018-2019 fiscal year federal sentencing data, which is updated yearly in the USSC datafile for individual offenders and quarterly in the FJC Integrated Database. Our merging of offender, sentence, and judge information proceeded largely as described in Section 3.4 and by Ciocanel et al. (2020b). Several USSC variables denoting the total prison sentence, the coding of the offense type, and the post-Booker reporting categories changed in the USSC database. Therefore we adjust the data processing approach of Ciocanel et al. (2020b) to remain consistent with the variables in JUSTFAIR. This process extends the dataset we use in this study by over 30,000 cases.

In our analysis, we include cases only from 2006 and later. We are principally concerned with racial disparities observed after the *Booker* decision, and so cases before 2006 lie outside of our target population. We further exclude immigration cases because of the unique use of fast-track sentencing in these cases and the extreme concentration of these cases in several southwestern district courts (Hartley and Tillyer, 2012). We also do not consider cases for which we cannot infer the sentence length. Specifically, in the JUSTFAIR database, there are about 13,000 cases that resulted in a sentence length of zero according to the continuous sentence total variable but resulted in prison time according to the categorical variable of imprisonment type. These two features contradict each other, so we consider the 13,000

cases to be missing the outcome. Therefore, we remove these cases from the analysis. After these pre-processing steps, the analytic sample represents about 380,000 cases corresponding to 1116 judges. The median number of cases per judge is 263 (μ = 358.7, σ = 389.6).

4.2.1.2 Measures

The outcome of interest is the length of the sentence assigned to the defendant. We cap the sentence length at 470 months because life sentences are generally coded as 470 months. We log-transform the outcome because its distribution has a positive skew. Because the log of zero is undefined, we add one to every sentence before log-transforming so that no incarceration outcomes (zero-month "sentences") have a defined value in our estimation. It is challenging to address zero-month sentences in analyses where sentence length is the dependent variable, and there is no perfect solution to the issue. Adding one to all sentences is a common approach (Fischman and Schanzenbach, 2012, Light, 2021, Rehavi and Starr, 2014, Yang, 2015). This approach appreciates that no incarceration outcomes are real outcomes that can contribute to racial disparities while also appreciating that a no incarceration outcome entails less incarceration time than any other sentence. Nevertheless, we acknowledge that this approach masks nuance in the difference between a no incarceration outcome and a one-month sentence. (Bushway and Piehl, 2001). Accordingly, we interpret the results in terms of percentage changes rather than in terms of linear changes in months of prison time.

The case-level characteristic of principal interest is the defendant's race and ethnicity, which we measure with' Hispanic,' 'Non-Hispanic Black' (henceforth 'Black'), 'Non-Hispanic White' (hereafter 'white'), and 'Other Race/Ethnicity.' We collapse

the groups in this way because, unfortunately, the sample sizes of the various groups in the final category are not large enough for us to perform statistically informative cross-judge analyses. Using these definitions of race and ethnicity, the median judge imposed sentences in 41 cases with Hispanic defendants ($\mu = 97.3$, $\sigma = 210.6$) and in 69 cases with Black defendants ($\mu = 114.1$, $\sigma = 137.7$).

We include an extensive set of control variables:

- the guideline minimum sentence, with any statutory minimum sentences taken into account,
- the defendant's criminal history points,
- crime type, namely 'violent crime,' 'drug-related crime,' 'embezzlement, fraud, theft,' or 'other,'
- whether the case was settled by plea agreement or trial,
- sentencing year, and
- defendant demographics, namely sex, age, U.S. citizenship status, and educational attainment.

The most important of these control variables is the guideline minimum sentence, which is standardized based on factors like the severity of the crime. Thus, at least in a theoretical world with no racial disparities, we would expect roughly equal sentences imposed on two defendants from different racial groups but who have the same guideline minimum sentence.

While the defendant's demographics are extralegal factors that judges ought not to consider in their sentencing, we nevertheless control for these demographics be-

cause they may capture variation in relevant legal factors. For example, if a judge hears cases where many Hispanic defendants have a unique situation not captured by observed legal variables, and if this unique situation is also associated with educational attainment, then controlling for educational attainment accounts for some variation due to the unobserved prevalence of this situation.

4.2.2 Analytic Strategy

4.2.2.1 Empirical Model

We estimate a hierarchical linear model of log-transformed sentence lengths where cases (Level 1) are nested within federal judges (Level 2). In particular, we estimate the following model:

$$\text{(Level 1)} \quad \log(Y_{ijk}) = \beta_{0j} + \beta_{\mathbf{100}}\mathbf{X_{ij}} + \beta_{\mathbf{1j}}\mathbf{C_{ij}} + \epsilon_{ij} \tag{4.2.1}$$

$$\text{(Level 2, Intercepts)} \quad \beta_{0j} = \gamma_{000} + \zeta_{0j} \tag{4.2.2}$$

$$\text{(Level 2, Slopes)} \quad \beta_{1jk} = \delta_{100k} + \eta_{1jk} \tag{4.2.3}$$

where Y_{ijk} is the length of the sentence corresponding to case i by judge j for a defendant of race k, β_{0j} is the intercept for cases heard by judge j, $\mathbf{X_{ij}}$ is a vector of defendant control characteristics corresponding to case i by judge j, $\beta_{\mathbf{100}}$ is a

vector of slopes between control characteristics and log transformed sentence length, $\mathbf{C_{ij}}$ is a vector of racial group binary indicators, $\beta_{\mathbf{1j}}$ is a vector of random slopes between logged sentence length and each racial group indicator among cases heard by judge j, ϵ_{ij} is an idiosyncratic error term, γ_{000} is the overall intercept, ζ_{0j} is the deviation of the intercept for cases heard by judge j from the overall intercept, β_{1jk} is the random slope corresponding to racial group k represented in $\beta_{\mathbf{1jk}}$, δ_{100k} is the overall slope between log transformed sentence length and racial group indicator k, and η_{1jk} is the deviation of the slope for cases heard by judge j from the overall slope corresponding to racial group k. Our model assumes that ϵ_{ij}, ζ_{0j}, and η_{1jk} each are normally distributed with mean zero.

We are primarily interested in the values of η_{1jk} for each judge. For example, if $k = Black$ and the reference category is white defendants, the value of $\eta_{1j,Black}$ answers the following question: how much greater is the Black-white sentencing disparity when judge j hears cases compared to when the average judge hears cases, controlling for defendant and case characteristics? This value does not necessarily reflect directly unfair treatment based on race. One reason for this is that a judge may hear cases in which Black and white defendants are unbalanced with respect to unobserved legal characteristics, even above and beyond observed control variables. The general approach of conditioning on observed factors is nevertheless common in research on racial sentencing inequalities (Feldmeyer and Ulmer, 2011, Light, 2021), in part because cases do not seem to be randomly assigned to judges even within districts. Regarding variability in $\eta_{1j,Black}$ values, we note that, because our model estimation shrinks random slopes and intercepts for judges with a small number of cases, such judges do not cause random slopes and intercepts to appear unduly variable.

The model includes judge random effects instead of judge fixed effects because the

former accommodate random slopes, which are the primary parameters of interest in this study. We forgo a hybrid model because it is less parsimonious and does not influence the estimated random slopes (Schunck, 2013), our primary interest. The model excludes district fixed effects because judges' racial disparities should be evaluated based on all other judges' disparities, rather than just those in the same district. Still, we estimated an alternative model with district fixed effects, and the results did not differ in any substantial way[1]. Thus, even if the two specification options were equally compelling theoretically, our preferred specification is more parsimonious and does not affect the main results.

4.2.3 Limitations of the JUSTFAIR dataset

In Ciocanel et al. (2020b), we detail several measures we took to ensure the quality of the JUSTFAIR database. We validated merged records by checking for additional common variables such as offense codes in the original datasets, used two independent sources to identify judge names from judge initials, and excluded records where sentencing dates fall outside the judges' activity periods. In addition, we carried out a manual data validation procedure for randomly sampled cases from the assembled dataset (Ciocanel et al., 2020b). The validation set confirms the accuracy of the merging procedure: very few cases violate the assumption that the judge involved in a proceeding is the same as the sentencing judge. Nevertheless, while we corrected these cases in the validation set, there could be rare additional instances where a judge other than the sentencing judge became involved in proceedings post-sentencing, leading to an incorrect inference in the database.

[1] All estimates of interest—average conditional racial disparities and interjudge standard deviations in conditional racial disparities—differed by less than 0.01. See the online supplementary materials for the corresponding code.

While JUSTFAIR is, to our knowledge, the largest publicly-available database of federal sentences currently available, it nevertheless remains an incomplete database of the cases sentenced in 2001–2018 (extended in this work to 2019) due to issues of data quality and merging challenges. For example, when merging USSC and FJC sentence and defendant information, JUSTFAIR only retains cases with unique matching. Similarly, when retrieving sentencing judge initials from PACER, criminal cases may include more than one defendant; however, JUSTFAIR only keeps records where all defendants with the same court docket number are associated with the same judge. JUSTFAIR also cannot include information on sealed federal cases. Other factors that prevent JUSTFAIR from being complete include the inability to distinguish between judges who have the same initials and district and the full or partial lack of judge initials in certain districts or records. In particular, due to such data quality issues, JUSTFAIR contains no sentencing data from the Eastern District of North Carolina, the Southern District of West Virginia, the Southern District of Texas, the Middle District of Tennessee, the Northern District of Illinois, the District of Guam, and the District of the Northern Mariana Islands. The database also includes limited sentencing data (less than 33% of the starting USSC cases) from the Northern District of Texas, the Southern District of California, the District of Oregon, the District of New Mexico, the Western District of Oklahoma, and the Northern District of Florida.

Non-random assignment of cases to judges presents another limitation, not just in the case of the JUSTFAIR database but rather whenever federal judge effects are of interest. The Administrative Office of the US Courts claims that "The majority of courts use some variation of a random drawing" (Administrative Office of the United States Courts, 2020). This statement means that, "to assure equitable distribution of caseloads and avoid judge shopping" (Administrative Office of the United States

Courts, 2020), the district courts have a rotation plan for cases assigned to judges. However, the US Courts Office also mentions that special expertise and geography are also considered in case assignments. Previous studies have exploited the random assignment of cases to judges in their analyses of sentencing equity (Anderson et al., 1999, Abrams et al., 2012, Cohen and Yang, 2019). Establishing random assignment has been considered key for these studies. Doing so ensures that each judge receives a similar combination of cases and defendants in terms of observable characteristics so that unobservable characteristics can also be assumed to be similar across judges.

For instance, Abrams et al. (2012) analyze racial sentencing disparities in felony cases from Cook County, Illinois. To verify random assignment, they use a Monte Carlo simulation methodology for felony data from 1995 to 2001. Specifically, they construct a randomly assigned dataset across characteristics such as defendant race, age, gender, and crime category and use it to establish random assignment for their dataset (Abrams et al., 2012). Similarly, the study of the proprietary federal sentence dataset in Cohen and Yang (2019) relies on the assumption that cases are randomly assigned to judges in the same district court, focusing on observed case and defendant characteristics across Republican-appointed vs Democrat-appointed judges. Motivated by whether the political affiliation of a judge's appointing president influences racial and gender gaps in sentencing decisions, the authors use a joint F-test to test whether there are significant differences in defendant characteristics by judge political affiliation as well as by judge tenure and gender. Conditioning on sentencing year and district court fixed effects, they find that cases are randomly assigned to sentencing judges from each political affiliation (Cohen and Yang, 2019).

Our setting is different from these studies, given that we are 1) analyzing the large, publicly-available JUSTFAIR dataset of federal sentences from 2001-2018 (Ciocanel et al., 2020b) and 2) interested in whether the cases in this dataset are assigned

randomly to individual judges within each district (rather than to judges with different characteristics). We find that when testing for random assignment using an F-test method similar to Cohen and Yang (2019) or using a Monte Carlo simulation method similar to Abrams et al. (2012), nearly every district shows evidence of nonrandom judge assignment. Specifically, almost every district shows statistically significant between-judge differences in at least one case characteristic, such as defendant race, defendant sex, and offense type. Due to nonrandom assignment, some unknown component of between-judge variability in conditional racial disparities arises from differences in unobserved factors that should reasonably influence the sentence. Observed covariates probably make this component much smaller than it would be otherwise, but we cannot know how substantial it remains after including these covariates.

4.3 Results

4.3.1 Aggregate Disparities

Conditional on observed case and defendant characteristics, judges assign white defendants sentences about 13% shorter than Black defendants' and 19% shorter than Hispanic defendants', on average. In contrast, defendants in other racial groups receive sentences about 10% conditionally shorter than white defendants' sentences, on average (Table 4.1). Given the tiny standard errors and p-values of all these estimates, sampling error is a weak candidate for explaining these patterns.

Our aggregate results align with previous studies, though they perhaps point to even more aggregate disparities than found in earlier timespans with different

Table 4.1: The average estimated racial disparities in logged sentence length. Each percent change in the first column represents what percent greater/lesser the conditional average sentence length is for defendants from the corresponding racial group, relative to white defendants. This quantity comes from the following formula, where δ_{100k} represents the estimated coefficient corresponding to racial group k, as shown in the "Estimate" column: Percent Change $= 100 \times (e^{\delta_{100k}} - 1)$.

	% Change	Estimate	Std. Error	p-value
Intercept		2.32	0.02	$<2 \times 10^{-16}$
Black	12.77	0.12	0.01	$<2 \times 10^{-16}$
Hispanic	19.15	0.18	0.01	$<2 \times 10^{-16}$
Other Race	-10.28	-0.11	0.02	5.82×10^{-11}

sampling mechanisms. In an investigation of a 2000–2002 pre-*Booker* USSC dataset, Feldmeyer and Ulmer (2011) find that, after controlling for a considerable set of legal and extralegal factors, Black defendant sentences are 6% longer and Hispanic defendant sentences are 1% longer on average than white sentences. In an analysis of a 2006–2008 multi-agency-based dataset, Rehavi and Starr (2014) similarly find that, in the aggregate, Black defendants receive sentences that are 9% longer than those of similarly situated white defendants who commit the same crimes. Finally, using a federal sentencing dataset similar to ours but covering cases from 2000 to 2010, Yang (2015) finds that Black offenders receive 1.9 months longer sentences, and Hispanic offenders over 1.9 months longer sentences than those of similar white offenders post-*Booker*.

4.3.2 Interjudge Variability in Disparities

How much do judges vary in their estimated racial disparities in sentencing? The estimates in the previous section show only the average extent of estimated racial disparities. We turn now to a discussion of how much judges are spread around this average. Table 4.2 shows that the spread is considerable for each racial group,

with random slope standard deviations ranging from 0.21 to 0.31 log-transformed years. The random intercepts, estimating judges' sentencing severity against white defendants, also vary to a similar extent.

Table 4.2: Dispersion of random effects. Random intercepts—ζ_{0j} in Equation (4.2.2)—represent judges' estimated sentencing severity toward white defendants. Random slopes—η_{1jk} in Equation (4.2.3)—represent judges' estimated racial disparities compared to the average judge's estimated racial disparities.

	Std. Dev.
Random Intercepts	0.35
Random Black Slopes	0.21
Random Hispanic Slopes	0.23
Random Other Race Slopes	0.31

Because interpreting the foregoing standard deviations is challenging given the log-transformed dependent variable, we present a more intuitive account of the interjudge variability in Table 4.3. As noted before, the average judge assigns Black defendants sentences conditionally 13% longer than white defendants'. However, a judge one standard deviation above average in terms of estimated Black-white disparity gives Black defendants sentences conditionally 39% longer than white defendants'. Furthermore, a judge two standard deviations above average assigns Black defendants sentences that are conditionally 71% longer than white defendants'. Similarly, the average judge assigns Hispanic defendants sentences that are conditionally 19% longer than white defendants'. A judge one standard deviation above average in terms of estimated Hispanic-white disparity assigns Hispanic defendants sentences 49% longer than white defendants'. Finally, a judge two standard deviations above average assigns Hispanic defendants sentences conditionally 87% longer than white defendants'.

To test whether the variance in random slopes is significantly different from zero, we conduct a likelihood ratio test that compares our model to a model that excludes random slopes but is otherwise identical. The p-value resulting from this

Table 4.3: The percent change in sentence length conditionally associated with the defendant being in different racial groups relative to being white by the average judge (column 1), by a judge one standard deviation above average (column 2), and by a judge two standard deviations above average (column 3), in terms of estimated racial disparities.

	% Change (Avg.)	% Change (1 SD)	% Change (2 SD)
Black	12.77	38.94	71.20
Hispanic	19.15	49.35	87.20
Other Race	-10.28	22.05	66.05

test is less than 2.2×10^{-16}, suggesting conditional racial disparities in sentencing do vary across judges. The Bayesian Information Criterion for the model with random slopes is 1,209,679, compared to a much less favorable value of 1,211,157 for the model without random slopes. In short, if all judges' conditional racial disparities were the same, the foregoing values would be surprising.

4.4 Conclusion

Using the JUSTFAIR database of district court criminal sentences, we computed a hierarchical linear model of log-transformed sentence length with cases nested within federal judges. We controlled for recommended sentence, crime type, presence of a plea agreement, sentencing year, defendant demographics, and more. Our model calculated that the percentage by which judges' sentences given to Black and Hispanic defendants are conditionally greater than the sentences judges given to white defendants.

Our findings suggest that average racial disparities are sizeable and that judges vary nontrivially in terms of how racially disparate their sentencing patterns are. For Black-white disparities, Black defendants on average receive 13% longer sen-

tences than observationally equivalent white defendants receive, and a judge one standard deviation above average in Black-white disparities gives 39% longer sentences to Black defendants. For Hispanic-white disparities, Hispanic defendants on average receive 19% longer sentences than observationally equivalent white defendants receive, and a judge one standard deviation above average in Hispanic-white disparities gives 49% longer sentences to Hispanic defendants. The online supplementary materials for this paper include a replication package with the data and code we used for these findings.

We emphasize some words of caution in interpreting our results. First, because cases are not randomly assigned to judges, we cannot know to what extent unobserved factors that systematically differ across judges drive the interjudge differences we find. A second limitation is our inability to observe every case or even every district. Like most social science datasets, the JUSTFAIR database contains a sample rather than the population. Thus, one cannot know what determines whether a case in the sampling frame is missing. If sample inclusion is unrelated to our measures of interest, then lacking the entire population leaves our estimates more susceptible to random sampling error than they would be otherwise. If sample inclusion is related to our measures of interest, then lacking the full population leaves our estimates more susceptible to systematic and random sampling errors than otherwise. In light of these limitations, our results should be seen as imperfect approximations for the degree of aggregate racial disparities and the interjudge variation therein.

Future research might extend analyses to state courts. While our work has examined criminal sentencing in federal district courts, we know that most sentencing in the United States occurs in state courts. To understand interjudge variability in race-based sentencing disparities, we must gather and analyze data from all 50 state court systems, as well as districts, territories, and other possessions. This task will

be challenging given that each of these systems has its own sentencing frameworks, data-keeping practices, and levels of transparency. Still, it will be worthwhile to help move the justice system towards greater equity.

Chapter Five

Everything is Relative: Auditing with Optimal Transport

This chapter is based on work presented in Kwegyir-Aggrey et al. (2021).

5.1 Introduction

Public attention on algorithmic discrimination has shifted research attention from simply learning accurate models to training models that meet a broader set of social and technical demands. This attention shift has created a new practice of algorithmic auditing: careful examination of algorithms' performance with respect to social rather than traditional computational metrics. The most well-known algorithmic audits compute standard performance metrics (e.g., true positive rate, false negative rate) for different protected groups to reveal group-wise disparities in performance Angwin et al. (2016), Buolamwini and Gebru (2018), Obermeyer et al. (2019). As auditing has become more prevalent, many questions have arisen regarding ways to make these audits more informative and transparent Raji and Buolamwini (2019), Wilson et al. (2021), Burrell (2016).

In response to these audits, scholars have attempted to develop algorithms that learn fair classifiers Hardt et al. (2016), Calders et al. (2009), Zafar et al. (2017). These efforts have operated under the assumption that if an algorithm is provably fair according to some statistical fairness guarantee, it will behave fairly in practice. In numerous cases, however, this assumption has been highly contested Corbett-Davies and Goel (2018), Hoffmann (2019). Issues surrounding the practicality of these guarantees Lohaus et al. (2020), Hu and Chen (2020), their ability to safeguard various identities and their intersections Morina et al. (2019), and their efficacy in representing fairness and equity based in broader historical contexts Hanna et al. (2019) has generally left practitioners unsatisfied with these approaches, leaving them

in search of alternate methodologies to evaluate the performance of their automated-decision making tools.

Auditing algorithms has revealed not only the problems with the individual systems but also with the metrics used Raji et al. (2020). We consider auditing to be an inquisitive process that requires more than computing bias metrics; a successful audit must also provide insight regarding the structure of disparities. This insight may illuminate problems (to act as a formalizer per Abebe et al. (2020)), give the designers an appropriate path to select an intervention, or allow for the declaration of clear guidelines around the limits of its use.

Here, we introduce optimal transport on the output of a classifier as an auditing tool. In supervised learning settings, several works have proposed using optimal transport to de-bias training data to compute fairer classifiers downstream Silvia et al. (2020), Zehlike et al. (2020), Gordaliza et al. (2019), Jiang et al. (2019), Feldman et al. (2015), Johndrow et al. (2019), Chiappa and Pacchiano (2021). We present a new use: an auditing framework that can conduct audits at the individual, subgroup, and group levels while also investigating the structure of bias exhibited by some classifier. We show that flexibility within optimal transport techniques makes them well suited for real-world audit problems, where the needs of each auditor may vary greatly, and versatility is a crucial design goal.

Contributions. This work tackles several critical challenges in current algorithmic auditing research. Primarily, we develop a tool that goes beyond simply certifying if a classifier is fair or not by also discerning *how* it is unfair. Specifically, we define a new approach to the study of algorithmic bias that uses the Wasserstein distance between output distributions and its transport map to investigate disparities at the individual, subgroup, and group levels. Our audits recover disparities

previously found in the COMPAS and German credit datasets and offer explorations of their structure. We claim that our framework is sufficiently adaptable to perform audits in the context of any supervised learning system. Furthermore, it can interface with human stakeholders to better inform their understanding of algorithmic discrimination than current practices based on rigid and conflicting fairness definitions based on classical performance metrics.

5.2 Background

Our audit seeks to understand *if* and *under what circumstances* a classifier produces unfair outcomes. Current practices require that a practitioner choose a fairness definition well suited to the problem domain. In many cases, however, the goals of a sociotechnical intervention may be difficult to distill to singular or even multiple fair definitions Hoffmann (2019), Selbst et al. (2019). Instead of fixing a fairness notion, in our framework, we allow the auditor to supply an example set of outcomes deemed to be fair, which we refer to as a *fair reference*. Determining how a classifier differs from the fair reference reduces the problem of comparing the behavior of the fair reference to that of the classifier being audited. In Section 5.4, we take this reference to be a dataset or classifier.

Preliminaries. Allow $\mathcal{X} \subseteq \mathbb{R}^k$ to be some domain and Y to be a set of potential decisions. To begin, we consider the binary classification case where $Y = \{0, 1\}$ and $Y = 1$ denotes the positive decision. Allow $f : \mathcal{X} \to [0, 1]$ to be a function the denotes a model's belief that an individual $x \in \mathcal{X}$ will achieve the positive outcome, or rather $Pr(Y = 1|X = x)$. Allow $d(x)$ to be a decision rule to create prediction $\hat{Y}$, typically through some threshold $\tau \in [0, 1]$ where $d(x) = \mathbf{1}_{f(x)>\tau}$.

5.2.1 Moving Beyond Thresholds to Comparing Distributions

Commonly evoked fairness notions, such as Demographic Parity and Equalized Odds, evaluate model behavior in the binary decision space rather than studying the model scores before thresholding. Comparing model beliefs before applying a decision function can reveal more detailed information regarding the structure of the function's predictions. Furthermore, by conducting analysis in the outcome space rather than the input space, we can more easily study differences in classification probabilities since it is difficult to find a good similarity metric for individuals in the feature spaceDwork et al. (2012), Ilvento (2019). However, differences in model beliefs can anchor our analysis of potentially disparate representations for different groups in the feature space. This hypothesis is closely related to the idea of distribution matching Quadrianto and Sharmanska (2017), Quadrianto et al. (2009), where the distribution of classifier output is evaluated with respect to some target distribution. Next, allow $g : \mathcal{X} \to [0, 1]$ to reflect the beliefs of a target reference. We write F and G to be the random variables corresponding to the distribution over risk scores output by functions f,g respectively Then, we can view the distributions of model beliefs as probability measures μ_f and μ_g, where for some score $\alpha \in [0, 1]$ we have $\mu_f(\alpha) = Pr(F = \alpha)$.

The choice in fair reference is crucial, as it allows the practitioner to examine fairness along some group/subgroup lines or at the individual level. Several existing works demonstrate how to construct hypothetical fair distributions Jung et al. (2019), Agarwal et al. (2018), Wang et al. (2019). In Example 5.2.1, Blue College could hope to admit a representative cohort with specified levels of representation. This comparative setup allows for a more flexible understanding of the context of the

application. Furthermore, we can analyze the score-generating process by comparing outcome distributions rather than the final decisions alone.

Example 5.2.1. Consider the fictitious Blue College who want to audit their admissions data to see if they meet their goal of creating representative cohorts. Blue College has records of the scores $f(x)$ they have assigned to applicants $x \in X$ and demographic information. Similarly, they have a proposed scoring function $g(\cdot)$, which meets their admissions goals. Since Blue wants to admit representative cohorts, not all student groups must be admitted with the same frequency, which means an audit must provide more than average group performance measures. Blue College computes their observed and intended outcome distributions over admission scores μ_f and μ_g, and uses our Wasserstein framework to observe how these distributions differ and if the college is meeting its goals. Later, we will show how the optimal transport map, a key component in computing the Wasserstein distance, can also show the college where exactly their current practice differs from the intended policy.

5.2.2 Optimal Transport

To measure the difference between sets of outcomes, we use optimal transport, which seeks the most cost-effective way to transform one probability distribution into another Peyré et al. (2019). The Kantorovich optimal transport problem seeks to find a minimal cost mapping between two probability distributions Peyré et al. (2019). Referring back to the problem of moving a sand pile to fill in a hole, Kantorovich optimal transport allows us to split the mass of a grain of sand instead of moving the whole grain; therefore, the mappings need not be 1—1. For probability measures μ and ν defined on metric spaces $\mathcal{X}^1$ and $\mathcal{X}^2$, respectively, this optimal transport

problem finds a minimal coupling π that attains

$$\min_{\pi \in \Pi(\nu,\mu)} \int_{\mathcal{X}^2 \times \mathcal{X}^1} c(x^1, x^2) d\pi(x^1, x^2), \tag{5.2.1}$$

where $c(x, y)$ is a cost function and $\Pi(\mu, \nu)$ is the set of couplings of μ and ν given by

$$\Pi(\mu,\nu) = \{\pi \in P(\mathcal{X}^1 \times \mathcal{X}^2) : \pi(A \times \mathcal{X}^2) = \mu(A)\ \forall A \subset \mathcal{X}^1, \\ \pi(\mathcal{X}^1 \times B) = \nu(B)\ \forall B \subset \mathcal{X}^2\}. \tag{5.2.2}$$

Intuitively, the cost function says how many resources it will take to move x_i^1 to x_j^2, and the coupling entry π_{ij} assigns a probability $\pi(x_i^1, x_j^2)$ that x_i^1 should be moved to x_j^2. When the spaces of interest are the same metric space with set $\mathcal{M}$, distance d, and cost function $c(x, y) = d(x, y)^p$, the optimal transport distance (Equation 5.2.1) is equivalent to the $p-$th Wasserstein distance:

$$W_p(\mu,\nu) = \left(\inf_{\pi \in \Pi(\mu,\nu)} \int_{\mathcal{M} \times \mathcal{M}} d(x, y)^p d\pi(x, y) \right)^{\frac{1}{p}}. \tag{5.2.3}$$

The Wasserstein distance measures the distances between probability distributions on a metric space and is commonly used in machine learning applications.

Given observed data $x_1^1, x_2^1, \ldots x_{n_1}^1 \in \mathbb{R}^d$ and $x_1^2, x_2^2, \ldots x_{n_2}^2 \in \mathbb{R}^d$, we define the marginals as discrete empirical distributions:

$$p^j = \sum_{i=1}^{n_j} p_i^j \delta_{x_i^j} \text{ for } j = 1, 2,$$

where δ_x is the Dirac measure. Then, the cost function is given as a matrix $C \in$

$\mathbb{R}^{n_1 \times n_2}$, e.g. $C_{ij} = \left\|x_i^1 - x_j^2\right\|$, and the set of couplings are the matrices

$$\Pi(p^1, p^2) = \{\Gamma \in \mathbb{R}_+^{n_1 \times n_1} : \Gamma \mathbb{1}_{n_2} = p^1,\ \Gamma^T \mathbb{1}_{n_1} = p^2\}. \tag{5.2.4}$$

The discrete coupling Γ relates measures p^1 and p^2: Each row of Γ_i tells us how to split the mass of data point x_i^1 onto the points x_j^2 for $j = 1, \ldots, n_2$, and the condition $\Gamma \mathbb{1}_{n_2} = p^1$ requires that the sum of each row Γ_i equals p_i^1, the probability of sample x_i^1. The discrete optimal transport finds a coupling that minimizes the cost of moving samples through the linear program:

$$\min_{\Gamma \in \Pi(p^1, p^2)} \langle \Gamma, C \rangle. \tag{5.2.5}$$

Although this problem can be solved with minimum cost flow solvers, it is usually regularized with entropy for more efficient optimization and empirically better results Cuturi (2013). Entropy diffuses the optimal coupling, meaning that more masses will be split. Thus, the numerical optimal transport problem is

$$\min_{\Gamma \in \Pi(p^1, p^2)} \langle \Gamma, C \rangle - \epsilon H(\Gamma), \tag{5.2.6}$$

where $\epsilon > 0$ and $H(\Gamma)$ is the Shannon entropy defined as

$$H(\Gamma) = \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \Gamma_{ij} \log \Gamma_{ij}.$$

Equation 5.2.6 is a strictly convex optimization problem, and for some unknown vectors $u \in \mathbb{R}^{n_1}$ and $v \in \mathbb{R}^{n_2}$, the solution has the form $\Gamma^* = \text{diag}(u) K \text{diag}(v)$, with $K = \exp\left(-\frac{C}{\epsilon}\right)$, element-wise. This solution can be obtained efficiently via

Sinkhorn's algorithm, which iteratively computes

$$u \leftarrow p^1 \oslash Kv \text{ and } v \leftarrow p^2 \oslash K^T u, \tag{5.2.7}$$

where $\oslash$ denotes element-wise division. This derivation immediately follows from solving the corresponding dual problem for Equation 5.2.6 Peyré et al. (2019).

5.2.2.0.1 Why Wasserstein? We choose the Wasserstein distance over divergence measures such as Kullback–Leibler, Jensen–Shannon, or F-Divergence since it can detect geometric differences and yield an optimal coupling or transport plan between the distributions. We leverage this transport map to understand the structure of observed biases better and offer actionable suggestions to reduce bias in automated systems. Furthermore, we can compute the Wasserstein distance between distributions that are not defined over the same supports, meaning that we can compare models across two seemingly disparate groups of observations through their outcomes. We refer the reader to Arjovsky et al. (2017) for a more thorough treatment of this discussion.

5.2.2.0.2 Previous Applications of Optimal Transport in Fairness. Other works have proposed the Wasserstein metric to design fair statistical tests on the individual Xue et al. (2020) or group level Taskesen et al. (2020) but only use the transport map to detect if a classifier is robust to changes in a protected attribute. Dwork et al. (2012) and Feldman et al. (2015) motivate the use of the Earth Mover's distance, a particular case of the Wasserstein distance, to study algorithmic bias and fairness in the feature space. Since the reception of these efforts, there have been many other works Silvia et al. (2020), Zehlike et al. (2020), Gordaliza et al. (2019), Jiang et al. (2019), Feldman et al. (2015), Johndrow et al. (2019), Chiappa

and Pacchiano (2021) which also use optimal transport to study fairness, primarily by attempting to compute bias-neutral feature representations for downstream classification.

Our work is distinct in two key ways. Primarily, we use optimal transport maps for auditing and studying bias in classification instead of attempting to compute a fair classifier. Our technique is most similar to Black et al. (2020), which uses an optimal transport map to determine if a classifier is sensitive to changes in protected attributes. This style of audit is also explored in Taskesen et al. (2020), Xue et al. (2020), where both works use optimal transport to measure changes in classification under perturbation to input features. Second, we compute the Wasserstein distance on outcomes rather than features, and we use the transport map to audit a classifier with respect to a *fair reference*, which we show offers detailed analysis regarding other aspects of a classifier's bias that are not exposed by these other methods.

Motivating Audits in the Outcome Space. Since finding a suitable distance function in the feature space has proven to be a challenging task Dwork et al. (2012), Ilvento (2019), we conduct our analysis in the outcome space rather than the feature space. To this end, we rely on the universality of classification probabilities in the outcome space, which can be interpreted very similarly between classification problems (i.e., $P(Y = 1|\cdot)$ is a standard choice). Additionally, we note that the performance of the empirical Wasserstein estimator decays rapidly as the dimensionality of the data in the computation increases Peyré et al. (2019). Since classification outcomes are almost always lower dimensional than their respective feature space, our method also enjoys improved convergence properties compared to techniques that compute the Wasserstein distance over features.

5.2.3 Parity-Based Fairness

Below we review standard definitions of group fairness that we will use as points of comparison. First, we introduce demographic parity, which requires that predictions have equal accuracy regardless of protected attribute status.

Definition 5.2.2 (Demographic Parity)**.** A classifier is fair under demographic parity if the classifier's predictions are independent from the sensitive attribute Calders et al. (2009):

$$\mathbb{E}_{\mathcal{D}}[h(X)|\mathcal{S}=1] = \mathbb{E}_{\mathcal{D}}[h(X)|\mathcal{S}=0]. \tag{5.2.8}$$

Note that this is simply a statement of equality and does not reveal the quantity of observed deviation from a fair ideal, should such a deviation exist and the condition remain unsatisfied.

Chouldechova (2016) demonstrated that satisfying demographic parity requires each protected group to have the same distribution of features. Most data distributions violate this condition, and in that case, enforcing demographic parity can harm classifier performance Zafar et al. (2017), Lohaus et al. (2020), Hu and Chen (2020). To avoid this performance loss, difference of demographic parity (DDP) Wu et al. (2019) relaxes the strict independence assumption:

$$DDP(f) := \mathbb{E}_{\mathcal{D}}[h(X)|\mathcal{S}=1] - \mathbb{E}_{\mathcal{D}}[h(X)|\mathcal{S}=0],$$

and fairness is enforced by minimizing this difference.

Although DPP can increase the number of tractable outcomes, it suffers from

many of the same pitfalls as demographic parity. Namely, it relies on binary protected groups, offers no quantification of unfairness [1], and does not provide additional information regarding the structure of observed unfairness.

There are other statistical fairness notions exist to address some of the shortcomings of demographic parity. For example, equal opportunity Hardt et al. (2016) achieves fairness by equalizing the per-group true positive rates of a classifier.

Definition 5.2.3 (Equal Opportunity)**.** A classifier is fair under equal opportunity if its predictions have the same true positive rates across sensitive attribute membership:

$$\mathbb{E}_{\mathcal{D}}[h(X)|Y=1,\mathcal{S}=1] = \mathbb{E}_{\mathcal{D}}[h(X)|Y=1,\mathcal{S}=0].$$

Although equal opportunity differs from demographic parity in that it can be enforced despite differences in group conditioned feature distributions, the two expressions share many of the same limitations, such as the inability to examine the structure of unfairness. We refer the reader to Corbett-Davies and Goel (2018) for a complete analysis of parity-based fairness definitions.

5.2.4 Individual Fairness

Individual fairness Dwork et al. (2012) is summarized by the intuition that "similar individuals should be treated similarly." Although this definition is semantically appealing, choosing a reasonable similarity metric is often task-specific, thus rendering this notion of fairness untenable across domains Chouldechova and Roth (2018).

[1] In Section 5.4.1 we discuss disparate impact, which quantifies a lack of demographic parity

Furthermore, there are many instances, such as Example 5.2.1, where fairness may require treating dissimilar individuals similarly, which is not captured by this definition.

5.3 Method

This section revisits our preliminaries, introduces our optimal transport-based framework, and explains how we conduct our audits. Our domain $\mathcal{X} \subseteq \mathbb{R}^d$ remains unchanged, however now we allow $\mathcal{Y}$ to be any set of outcomes with finite cardinally, not only the binary set. We also introduce the notation of a policy, which maps features to some higher dimensional probability vector of model beliefs, rather than the univariate beliefs proposed in section 5.2.

Definition 5.3.1 (Policy)**.** A policy $\mathcal{F} : \mathcal{X} \to \mathcal{O}(\mathcal{Y})$ is a mapping from the feature space to the space of outcome distributions.

This notion of a policy allows for our framework to support multi-class classification problems where model beliefs are defined over more than one class. The output of the policy function is a classification vector defined over the $|\mathcal{Y}|$-dimensional standard simplex. We also assume there exists a distribution $\mathcal{D}$ over the set $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, which takes on values $(x, y) \sim \mathcal{D}$.

5.3.1 Comparative Audits

Denote the distribution on outcomes generated by observing a policy F on the space $\mathcal{X}$ by $\mu_{\mathcal{F}}^{\mathcal{X}}$. Given policies $\mathcal{F}_1$ and $\mathcal{F}_2$ on subsets $\mathcal{A} \subset \mathcal{X}^1$ and $\mathcal{B} \subset \mathcal{X}^2$, we use the

Wasserstein distance to measure the differences in outcomes they produce:

$$\mathcal{W}_2(\mu^{\mathcal{A}}_{\mathcal{F}_1}, \mu^{\mathcal{B}}_{\mathcal{F}_2}). \tag{5.3.1}$$

We can compare many possible scenarios with this setup, including a desired outcome to a ground truth. By setting $\mathcal{A}$ and $\mathcal{B}$ to be the same population, we measure the impact of different policies. Alternatively, we can take $\mathcal{F}_1$ and $\mathcal{F}_2$ to be the same policy to study its effects on different populations. In a fairness audit, one policy will represent a decision-making system being audited, and the other will be a reference that has been deemed fair. In the context of Example 5.2.1, the populations are the real applicants and synthetic applicants, and the policies are the historical admissions scores and the desired outcomes. The optimal coupling maps each student's outcome onto the nearest outcomes in ideal reference. Figure 5.1 gives an overview of this process.

In practice, we do not have access to the true distributions of the outcomes, so, we approximate the outcome distributions with uniform distributions as is standard in computational optimal transport Peyré et al. (2019) and compute $\mathcal{W}_2(m^A_{\mathcal{F}_1}, m^B_{\mathcal{F}_2})$:

$$m^A_{\mathcal{F}_1} = \frac{1}{n_1}\sum_{i=1}^{n_1} \delta_{\mathcal{F}_1(a_i)} \text{ and } m^B_{\mathcal{F}_2} = \frac{1}{n_2}\sum_{j=1}^{n_2} \delta_{\mathcal{F}_2(b_j)}. \tag{5.3.2}$$

In the above distributions, the outcomes are generated by applying the policies to data $A = \{a_i\}_{i=1}^{n_1} \sim \mathcal{A}$ and $B = \{b_j\}_{j=1}^{n_2} \sim \mathcal{B}$. In this case, the cost matrix is computed by:

$$C_{ij} = \|\mathcal{F}_1(a_i) - \mathcal{F}_2(b_i)\|_2 . \tag{5.3.3}$$

We implement this method with the Python Optimal Transport toolbox (`https://pot.readthedocs.io/en/stable/`).

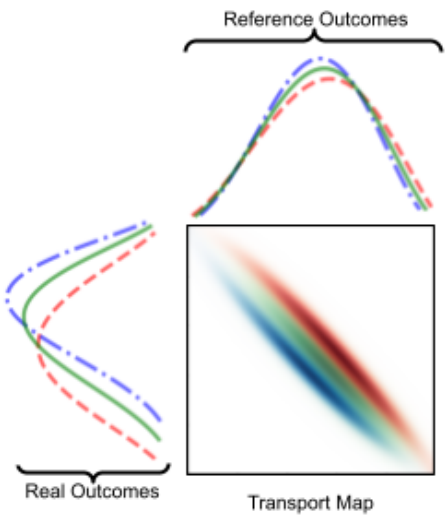

Figure 5.1: We apply our method to audit the admissions outcomes with the ideal representative outcomes as a reference to generate a transport map in the form of a coupling matrix. The red and blue colors in the coupling matrix show how mass should be mapped from real data groups to reference data groups. Each entry π_{ij} is the probability that outcome i from the real dataset should be mapped to outcome j in the synthetic dataset. We exploit the coupling matrix to explore the nature of bias in the admissions policy across groups. By tracing which groups in the reference dataset mass from one group in the real dataset are mapped to, we can understand which groups achieve their ideal outcomes.

This computation reveals the magnitude of difference between the outcomes from policies $\mathcal{F}_1$ and $\mathcal{F}_2$, and the optimal coupling π associated with the Wasserstein distance is a transport plan describing how to split the mass of outcomes from $\mathcal{F}_1$ onto outcomes from $\mathcal{F}_2$ to produce similar behavior. In Example 5.2.1, the entries of the normalized row $n_1\pi_i$ describe how to split the mass of student a_i's outcome $\mathcal{F}_1(a_i)$ onto the outcomes of the reference students $\{F_2(b_j)\}_{j=1}^{n_2}$. Since the outcome vectors result from applying policies to individuals in the datasets, we can also trace the policy mappings and interpret the transport map between groups in the two sets.

Discrete Outcomes and Total Variation. In many cases, the auditor may only have access to discrete decisions instead of complete model beliefs. In this case, Equation 5.3.1 reduces to the Total Variation distance. More for-

mally, the Total Variation between two measure $\delta(\mu, \nu)$ can be written $\frac{1}{2}\delta(\mu, \nu) = \inf_{\pi \in \Pi(\mu,\nu)} \mathbb{E}_{(x,y)\sim\pi} \mathbf{1}_{x=y}$ Villani (2008). We note Total Variation has previously been utilized as a statistical metric for both individual and group fairness works Gordaliza et al. (2019), Dwork et al. (2012). We highlight the distance as a special case of our framework, though it is less expressive than the Wasserstein distance.

5.3.2 Analysis Through the Coupling

Computing the Wasserstein distance between two sets of outcome distributions results in a coupling that maps the outcomes in one domain into outcomes in the other. Recall that an outcome results from applying a policy to an individual. We can trace back these outcomes to the individuals to which they correspond, therefore allowing us to also interpret the coupling as a mapping between individuals. Explicitly, each outcome $\mathcal{F}_1(a)$ corresponds to some individual a. Thus, we can think of the coupling as function that maps individuals in A to individuals in B based on the similarity of their outcomes. In this way, we can examine how policies treat people differently in feature space. By first mapping individuals to one another by the similarities of their outcome, we can then try to account for what differences in the feature space are responsible for this similarity.

We introduce the distance function d on the feature space as a parameter to evaluate relevant dimensions of the feature space to the current audit. This metric is not required to capture an absolute, bias-free distance between individuals because we use it to explore, not evaluate, similar to Xue et al. (2020), and we advocate considering several distances to explore possible relations in feature space fully. We assume that differences in the feature space and accompanying choices in metric are flawed and are likely responsible for disparate classification outcomes. Our method

probes how individuals with seemingly different representations can have similar classifications. Specifically, we use the outcome space as a similarity oracle that grounds our analysis of problems in the feature space.

First, we examine differences between individuals. Allow $\Gamma(a, \pi)$ to be the set of all individuals that $a \in \mathcal{A}$ is mapped to under the coupling π:

$$\Gamma(a, \pi) = \{b \in B : \pi(a, b) > 0\}. \tag{5.3.4}$$

Suppose we can quantify how different individuals are with some distance on the feature space $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. Then, using this set we can define a notion of individual comparative bias.

We then interpret the coupling weights as the magnitude of outcome similarity. Note that the Kantorvich formulation of optimal transport allows mass splitting, so an individual can be mapped to multiple people. We then measure how different individuals are to their mapped counterparts, weighting this measurement by the similarity of their outcomes as computed by the coupling.

Definition 5.3.2 (Individual Audit Score)**.** The bias an individual a experiences across policies is measured by the expectation:

$$u_{\mathcal{X}}(a) = \mathbb{E}_{b \sim \pi(a, \cdot)}[d(a, b)]. \tag{5.3.5}$$

This measurement assigns a high audit score to an individual with a similar outcome under the audit policy to an individual under the fair reference who has a large distance under d. Ideally all individuals are mapped to their current outcome in the fair reference, such that $\pi(a, b) = 1$ iff $a = b$ and therefore $d(a, b) = 0$. Conversely,

this score is high when two individuals have similar outcomes when compared across policies but are very different under the chosen metric. For Example 5.2.1, if we take d to be the distance in GPAs, a high individual bias for a student with a top GPA would indicate that this student is treated as a student with a lower GPA in the reference set, which could indicate discrimination.

We can aggregate this individual audit score over subgroups and groups. Consider a partitioning of the input space $\mathcal{X}$ into discrete groups $\mathcal{G} = (G_1, G_2, ...G_n)$ where $G_i \subseteq \mathcal{X}$ such that $\mathcal{X} = \bigcup_{i=1} G_i$. These groups can be arbitrarily defined and need not correspond to the normative notion of protected attribute group or intersectional subgroups. We say that the bias experienced by some group is then a sum of the individual bias terms for members of that group.

Definition 5.3.3 (Group Audit Score)**.** The bias (comparatively) experienced across policies for a group G is measured by:

$$U_{\mathcal{X}}(G) = \sum_{g \in G} u_{\mathcal{X}}(g). \tag{5.3.6}$$

This measurement is large when many individuals in a group have high individual audit scores, meaning they have similar outcomes to individuals dissimilar to them under the choice in d. Under our analysis, if a group audit score is high, we can infer the classifier is likely performing poorly on that group. Next, we look to compare these group-level scores to see how the bias experienced by one group may be related to the classification outcomes of another. We make this clearer with the following remark.

Remark. We also can compare groups directly rather than across the whole population. If for all $i, j \leq |\mathcal{G}|$ all groups are disjoint $G_i \cap G_j = \emptyset$ then the group bias

measurement can be additively decomposed as

$$U_{\mathcal{X}}(G_i) = \sum_{j \leq |\mathcal{G}|} U_{G_j}(G_i)$$

where each $U_{G_j}(G_i)$ is the amount of bias measured in group G_i that is the result of comparison with individuals in group G_j.

This remark highlights how our comparative framework allows us to quantify and break down group-wise biases into some salient components. Namely, through this decomposition, we see how group-level biases interact with one another, extrapolating this information to investigate the object under audit better.

5.3.3 Conducting Audits with this Flexible Framework

Our auditing framework is purposefully open-ended for applicability in a variety of settings. While we apply the methodology to existing datasets in this work, we anticipate it could be applied as either an internal or external audit. For an internal audit, the designers of the algorithm could define a reference distribution and then compare against the outcomes produced on a testing set as in Example 5.2.1. In Section 5.4.2, we demonstrate how to audit the shortcomings of a classifier with access to a ground truth for comparison, and in Section 5.4.4, we show how to audit the underlying differences between populations that were assigned different classifications.

5.3.4 Computing Individual and Group Bias

Our definitions of individual and group bias assume that we can measure the entire population $\mathcal{X}$. In practice, we have data $A = \{a_i\}_{i=1}^{n_1} \sim \mathcal{A}$ and $B = \{b_j\}_{j=1}^{n_2} \sim \mathcal{B}$, which requires slight changes for computations.

Definition 5.3.4 (Individual Bias)**.** The bias an individual a from group $\mathcal{A}$ experiences in comparison to group B across policies is measured by the expectation:

$$u_B(a) = \sum_{b \in \Gamma(a,\pi)} d(a,b)\pi(a,b). \tag{5.3.7}$$

We need to alter the definition of group bias to account for the fact that we can only compare an individual $a \in G \subset A$ to their counterparts in B.

Definition 5.3.5 (Group Bias)**.** The bias (comparatively) experienced across policies for a group G within population A when compared to population B is measured by:

$$U_B(A \cap G) = \sum_{g \in A \cap G} u_{\mathcal{B}}(g). \tag{5.3.8}$$

Finally, if we decompose the datasets A and B into disjoint groups $A = \{A_1, A_2, \dots A_n\}$ and $B = \{B_1, B_2, \dots, B_m\}$, then the group bias measurement in population $\mathcal{A}$ relative to population $\mathcal{B}$ can be additively decomposed as

$$U_B(A_i) = \sum_{j \leq m} U_{B_j}(A_i)$$

where $U_{B_j}(A_i)$ is the amount of bias measured in group A_i within population $\mathcal{A}$ that is the result of comparison with individuals in group B_j within population $\mathcal{B}$. We

make use of this framework to compare how individuals in the COMPAS dataset are mapped across race and criminal history-based groups in Section 5.4.2 and how individuals in the German credit data are mapped across age and sex-based groups in Section 5.4.4.

In the case that an individual a in group A is very different from everyone in group B, the individual bias in Definition 5.3.2 will be high regardless of who a is mapped to. Thus, to detect such outliers, we can normalize the individual bias by the individual's own worst-case-scenario, i.e. when they have all of their mass mapped to the individual who is farthest from them:

$$u_B^*(a) = \frac{n_1 \sum_{j=1}^{n_2} \pi(a, b_j) d(a, b_j)}{\max_{j=1:n_2} d(a, b_j)}. \tag{5.3.9}$$

This new normalized individual bias can then be used to differentiate between cases when an individual experiences extreme bias or merely has no equivalent comparison in feature space.

5.4 Experimental Results

This section demonstrates the flexibility of our optimal transport framework by auditing simulated data and the COMPAS and German credit datasets. We ran all experiments on personal laptops.

5.4.1 Wasserstein Distance Detects Bias when Disparate Impact Fails

To motivate the Wasserstein distance for fairness evaluation, consider the fictitious Blue College, who are looking to audit their admissions data for bias. Blue's applicants come from two secondary schools: expensive private School A and public School B. Blue College would like to check that their admissions policy is not biased with respect to an applicant's secondary school. Looking at previous admissions data, Blue College observes that they accept 25% of students from either school with the following pattern:

1. *All* students in School A have the same probability of being accepted: $P(Y = 1|X = x) = P(Y = 1) = 0.25$.
2. Students in School B with GPAs in the top 25% of GPAs in their class are accepted ($P(Y = 1|X = x) = 1$), and students in School B with GPAs in the bottom 75% of their class are rejected ($P(Y = 1|X = x) = 0$).

Blue evaluates their decisions with the disparate impact criteria, which provides a measure of demographic parity.

Definition 5.4.1 (Disparate Impact (80% Rule))**.** A model is said to admit disparate impact if:

$$\frac{P(Y = 1|\text{School} = B)}{P(Y = 1|\text{School} = A)} \leq \tau = 0.8. \tag{5.4.1}$$

Blue college accepts 25% of the applicants from both schools, which implies that $\frac{P(Y=1|\text{School}=A)}{P(Y=1|\text{School}=B)} = 1$, and so the model does not admit disparate impact and is

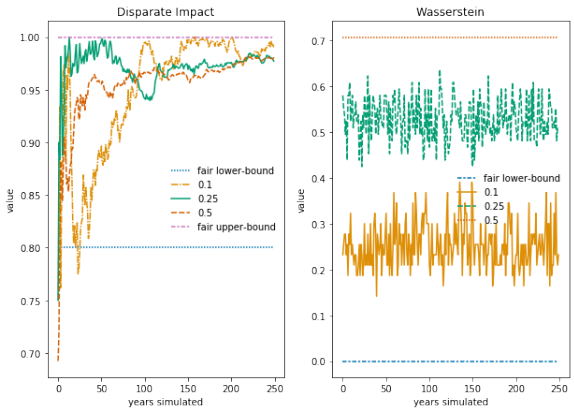

Figure 5.2: Under disparate impact, all rules converge to a fair policy despite structural inequity in the admissions policy. In contrast, the Wasserstein distance recovers the inequities at all rule levels.

considered *fair* with respect to this fairness definition. More generally, if $P(Y = 1|\text{School} = B) \leq P(Y = 1|\text{School} = A)$ then $\tau = 1$ implies this is the *fairest* possible outcome. Clearly, this model is not fair as it puts the bottom 75% of students from School B at a disadvantage, while also offering an advantage to the top 25% of students at school B.

We simulate several years of admissions according to this fictitious model with $\mathcal{A}$ and $\mathcal{B}$ as the students in the two schools, and $\mathcal{F}_1$, and $\mathcal{F}_2$ as the two described admissions patterns. In addition to the 25% rule described above, we also evaluate a 10% and 50% rule. A rule with a lower percentage indicates that Blue College accepts fewer students from both schools, and the 0% rule indicates that no student from A or B is accepted. In Figure 5.2, we plot the Wasserstein distance and disparate impact score as aggregated over the years.

According to disparate impact, for any given year, Blue College may appear fair or unfair. Still, the aggregate admissions data seems fair over time, despite obvious bias in the admissions process. Conversely, the Wasserstein distance reveals a fixed amount of bias over time and shows that the observed bias decreases as the rule tends

towards zero. As the decision rule approaches the trivially fair scenario where Blue College rejects all individuals from both schools, the Wasserstein distance decreases, indicating that the outcomes generated by the policies are moving closer together.

5.4.2 Recovering Racial Discrimination in Recidivism Risk Scores

We use the COMPAS Larson et al. (2016a) dataset to show that the optimal coupling can be used to asses the subgroup accuracy of a classifier. COMPAS is a commercial tool used to predict the recidivism risk of individuals awaiting trial. The COMPAS dataset contains data compiled by Propublica Larson et al. (2016b) on 6172 people scored by the COMPAS algorithm in Broward County, Florida from 2013-2014. Each person has a predicted recidivism score and a binary flag to indicate whether or not they recidivated. There are 3,175 Black defendants, 2,103 white defendants, and 2,809 defendants who recidivated within two years in this sample. It contains nine features: two-year recidivism, number of priors, age above 45, age below 25, female, misdemeanor, ethnicity, and predicted recidivism probability. We one-hot encode the categorical variables and normalize the features before fitting our models.

To audit the COMPAS tool for racial bias, Propublica Larson et al. (2016b) collected COMPAS scores, prior arrest records, and two-year re-arrest records in Broward County, Florida. The accompanying report found that the COMPAS tool assigned Black defendants risk scores that were disproportionately higher than those given to their white counterparts, even after controlling for criminal history. As a result, Black defendants who did not recidivate were often misclassified as "high-risk." In contrast, for white defendants, the opposite was true: those who *did* get

re-arrested were often mistakenly assessed as "low-risk."

We take $A = B$ to be the individuals in the COMPAS dataset and audit the COMPAS predictions (thresholded at 5) with respect to the observed outcomes (two years after initial arrest). This audit considers that COMPAS would be fair if it had high accuracy. For both the observed and predicted cases, we compute a logistic regression model, which takes as input relevant criminal history information and actual and predicted recidivism labels as targets. We used the following features to emulate the COMPAS algorithm: sex, age, race, felony counts, misdemeanor count, and prior offenses. We use these logistic models to estimate the likelihood an individual recidivates under the two scenarios, denoted $\mu^A_{\mathcal{F}_1}$ and $\mu^A_{\mathcal{F}_2}$, respectively, where $\mathcal{F}_1$ denotes the true and $\mathcal{F}_2$ the predicted recidivism policies. By computing $W_2(\mu^A_{\mathcal{F}_1}, \mu^A_{\mathcal{F}_2})$ we obtain the optimal coupling, which we then use to observe the differences between policies.

Using the criminal history information, we compute the individual and group audit scores. First, we observe that the group audit score for Black defendants who are considered high-risk is 5.88% higher than the Black defendants who were considered low-risk. However, for white defendants, we observe a 4.92% decrease in audit scores; the audit scores of white individuals considered low-risk are higher than that of white individuals deemed high-risk. This finding suggests that Black individuals who are considered high-risk and white individuals who are considered low-risk are more likely to attain similar predicted outcomes to an individual who is dissimilar to them in feature space, which is consistent with Propublica's observations. Black individuals who are considered high-risk often have similar criminal histories to white individuals considered low-risk and vice versa.Figure 5.3A shows the group-wise decomposition of audit scores. 43.8% of the outcomes attained by low-risk white individuals are similar to outcomes predicted for high-risk Black in-

dividuals, and 48.3% of high-risk Black individuals are mapped to low-risk white individuals.

5.4.3 Using a Fair Intervention as a Reference

In the above example, we compared the ground truth outcomes to the predicted COMPAS outcomes. Suppose that the creators of COMPAS wanted to see if they could make the predictions fairer. Then, they could test another classifier against the ground truth. To this end, we train a logistic classifier with an equal opportunity constraint based on Zafar et al. (2017) to make the predicted outcomes fairer. The fair classifier increases the true positive rates at test time for the Black and white groups by 10%. Furthermore, in Figure 5.3B, we can see that more of the subgroups are mapped to themselves.

For the positive-predicted group, which we documented initially as discriminatory against mainly Black defendants, we see a decrease in unfairness from the original to the equal opportunity outcome distribution. Moreover, when we decompose the source of remaining audit scores, a larger share of individuals are mapped to defendants in their same predicted risk class. This observed decrease in unfairness suggests that the fair classifier produces outcome distributions less dependent on race than the original COMPAS predictions.

Variable	Group
Savings	Actionable
Credit history	Actionable
Other installments	Actionable
Co-applicant or guarantor	Actionable
Number liable individuals	Actionable
Unemployed	Actionable
Property owned	Actionable
Number existing credits	Actionable
Installment rate	Actionable
Credit amount	Non-actionable
Credit duration	Non-actionable
Credit purpose	Non-actionable
Telephone	Non-actionable
Sex	Immutable
Age	Immuatble
Foreign national	Immutable
Years at job	Immutable
Skilled employee	Immutalbe
Permanent resident since	Immutable

Table 5.1: Actionable, non-actionable, and immutable features in the German credit dataset.

5.4.4 Exploring the Structure of Bias in the German credit dataset

The German credit dataset Hofmann (2000) contains 1000 instances with 26 features each of loan applications at a bank. It includes applicant profiles as well as their classification as a bad ($n_1 = 300$) or good ($n_2 = 700$) credit risk. We one-hot encode categorical variables and then z-score normalize all features before fitting our logistic regression model.

Table 5.1 describes how we classified features as actionable, non-actionable, and immutable as inspired by Chen et al. (2020).

To explore localization of bias with optimal transport, we use the German credit

dataset Hofmann (2000), which contains loan applications at a bank and their classification as good or bad credit risks and is commonly used for algorithmic recourse investigations Gupta et al. (2019), Karimi et al. (2020b), Chen et al. (2020), Rawal and Lakkaraju (2020). We train a logistic regression classifier $\mathcal{F}$ on the features without the sensitive attributes (age and gender) to produce outcome probabilities for each individual. In this case, there is no fair reference. Instead, we seek to understand the difference between those classified as bad applicants, A, and those classified as good, B, under policy $\mathcal{F}$. Figure 5.3C shows that this map recovers age and gender discrimination that has previously been found, namely that the policy favors older and male applicants. In particular 69.85% of men aged $[18, 25)$, 66% of women aged $[18, 25)$, and 63.21% of women aged $[25, 76)$ labeled as bad credit are disproportionately mapped to men aged $[25, 76)$ labeled as good credit.

To understand these disparities, we audit the German credit dataset using multiple distances in the feature space, drawing inspiration from the algorithmic recourse problem, which asks what actionable changes an individual can take to change their classification Karimi et al. (2020a). By exploring this question, we can understand whether an algorithm is discriminating based on actionable (and mutable), non-actionable (but mutable), or immutable (and non-actionable) traits (see Table 5.1 in the appendix for the full breakdown). For each type, we can compute the individual audit scores (Equation 5.3.5) while restricting the distance computation to features of that type. In Figure 5.5, we plot the distribution of the individual audit scores for each type of feature when split by age and sex for those in the bad credit class.

Men aged $[25, 76)$ with a bad credit classification have the highest actionable audit scores, suggesting that this group's actionable features are the furthest from those with a good credit classification. Consistently, women aged $[18, 25)$ with a bad credit classification have the lowest actionable audit scores even though they

receive far fewer good credit ratings. Since the actionable features, credit history, for example, are considered fair to use, this difference reinforces the claim that the policy discriminates along age and gender. For both non-actionable and immutable features, the differences in individual audit scores are more substantial across sex than across age. This result suggests that this classification algorithm disadvantages women for attributes that they cannot change.

For each individual a classified as bad credit, we can compute $\Gamma(a, \pi)$, the group of people with good credit that they were mapped to under the optimal coupling π. The product $n_1 \sum_{b \in \Gamma(a,\pi)} \pi(a, b)b$ projects individual a onto their counterparts in group B. Using this product, we can interpolate an individual's actionable features with those they were mapped to with weight α through

$$(1 - \alpha)a_{\text{actionable}} + \alpha n_1 \sum_{b \in \Gamma(a,\pi)} \pi(a, b)b_{\text{actionable}}. \tag{5.4.2}$$

We can then reapply policy $\mathcal{F}$ to an individual from the bad credit group with their original immutable and non-actionable features and interpolated actionable features to see how their classification changes in response to these actions. Figure 5.4 shows how the probability of being classified as having good credit increases with α, implying that making actionable changes to these features can affect outcomes. As α increases, the distributions shift from centering around low probabilities of receiving a good credit label to more uniform probabilities of receiving a good credit label, implying that changing these actionable features will improve an applicant's odds of success.

We can also examine changes in the actionable features of those who were reclassified to understand which features were the most important (see Figure 5.6 in the appendix). Previous studies tend to focus on recourse accuracy (how many people

were successfully reclassified as a result of actions taken) rather than which features were deemed most important, so we cannot compare our results directly. Our results suggest that having a critical account or credit at other banks and not having a guarantor or co-applicant most positively affect classification, and having no credits or credit repaid duly most negatively affect classification. A savings account has a significant impact on classification; no savings account or more than 100 DM in a savings account impacts classification positively, and having less than 100 DM affects classification negatively.

For each individual labeled as having bad credit, we interpolate their actionable features with the actionable features of those they were mapped to (Equation 5.4.2). Then, we reclassify these individuals. In Figure 5.6, we present the percentage of individuals reclassified as having good credit and record the average change for each actionable feature for varying values of α.

5.5 Discussion

We have presented an optimal transport-based framework for auditing decision-making algorithms to explore bias at various levels and according to flexible notions of fairness through a reference. We leverage the coupling matrix to study the structure of outcome disparities in the feature space. Through the COMPAS example, we have demonstrated that optimal transport can evaluate the accuracy of a classifier on populations and subgroups. Furthermore, the transport map elucidates structures of bias and allows us to quantify it. In studying the German credit dataset, we have shown that optimal transport can audit automated decisions after they are made through investigating aspects of the feature space.

Individual fairness Dwork et al. (2012) is often criticized due to its reliance on the existence of similarity metrics on features Chouldechova and Roth (2018). While our audit scores also rely on a distance in feature space, we can examine bias directly through the coupling matrix. Furthermore, similar to Xue et al. (2020), we treat the distance as a hyperparameter and advocate considering several to fully explore possible relations in feature space as well as leave room for customization of the audit goals. We envision this framework applied in various contexts, from internal audits of proposed plans to external audits. While this method allows a bad-faith audit by defining a dishonest reference, it is more robust to gaming by algorithm designers because it is not a single metric.

The constraint that all mass from one distribution must be transported onto the other limits standard optimal transport. In the context of fairness, this limitation means that outliers and unevenly represented subgroups may not be mapped to their appropriate counterparts. To address this, we could modify our framework to use unbalanced or partial optimal transport. Additionally, computing transport maps between large datasets can be computationally intensive. To address this issue, we could use entropically regularized optimal transport for auditing larger datasets. Future work will focus on addressing these limitations and formalizing the mathematical theory and its connections to definitions of fairness.

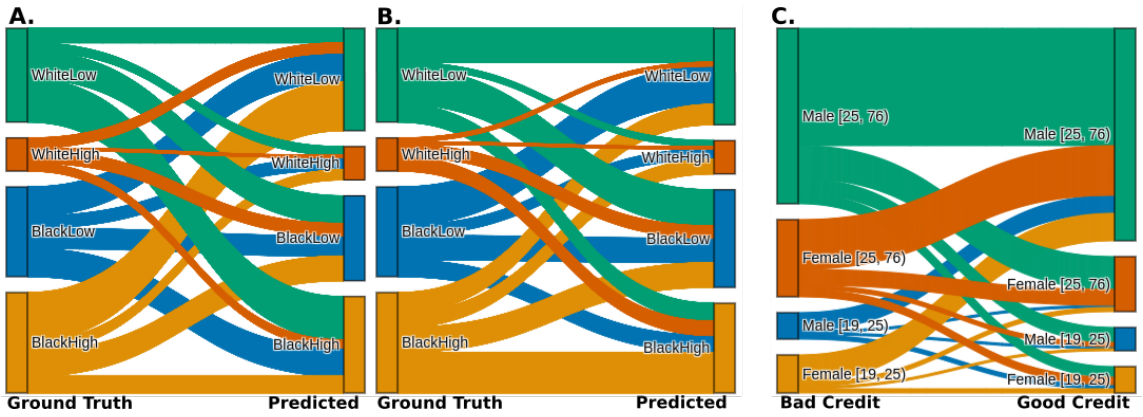

Figure 5.3: Shown is the way the coupling transports groups from one population to another. Each colored flow on the left shows how the mass from one group in the first population is split and transported to groups in the second population on the right. For example, by looking at the blue flow in A, we can see that a large fraction of high-risk white individuals achieve classification outcomes from COMPAS similar to low-risk Black defendants compared to ground-truth recidivism. By comparing the diagrams for **A.** COMPAS's predictions and **B.** equal opportunity classifier's predictions, we can see that the differences are less stark in the outcomes produced by the Equal Opportunity classifier. **C.** The transport map recovers previously found biases in age and gender in the German credit dataset.

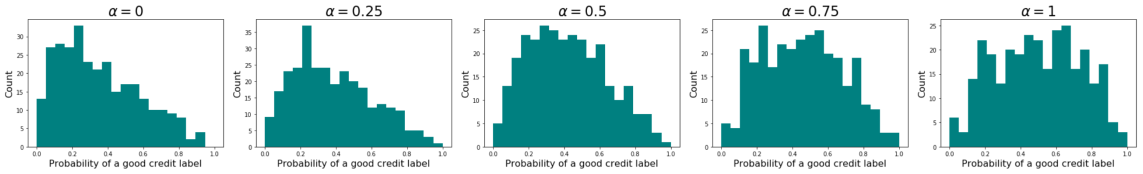

Figure 5.4: As α increases, increasing the weight of the actionable features of those with good credit in the interpolation (Equation 5.4.2), the distribution of the probability of receiving a good credit label for those who had previously been labeled as bad credit shifts from being skewed towards low probabilities to a more uniform distribution.

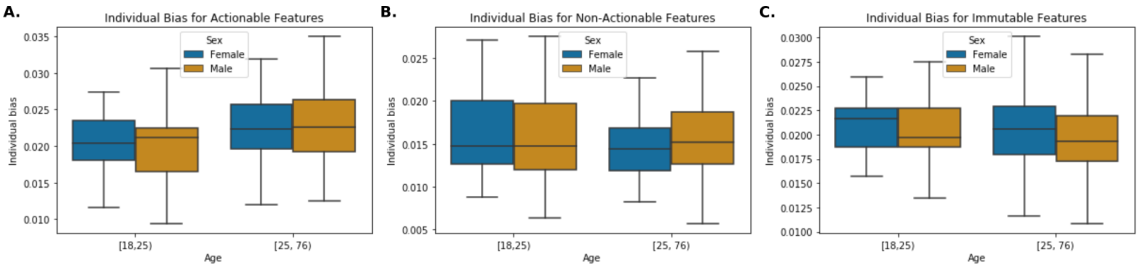

Figure 5.5: Individual audit scores in those labeled as bad creditors broken down by age and sex in the German credit dataset Hofmann (2000) when distances are taken with respect to **A.** actionable, **B.** non-actionable, and **C.** immutable features.

α	1	0.75	0.5	0.25
Percent of flipped classifications	38.64%	27.27%	16.36%	5.45%
Critical account or credits at other banks	**0.716**	**0.565**	**0.401**	**0.262**
No savings account	0.426	0.328	0.236	0.123
No gurantor or co-applicant	0.423	0.248	0.061	-0.015
Owns real estate	0.388	0.274	0.331	0.251
Has guarantor	0.257	0.193	0.071	-0.013
>1000 DM savings account	0.244	0.184	0.084	0.028
500-1000 DM savings account	0.242	0.118	0.057	0.098
100-500 DM savings account	0.133	0.059	-0.026	0.029
Number of existing credits	0.029	0.105	0.155	0.119
Owns a car or other	-0.011	0.053	0.013	-0.032
Has building society savings agreement or life insurance	-0.076	-0.034	-0.071	-0.127
Credits repaid duly until now	-0.084	-0.122	-0.175	-0.149
History of delayed payments	-0.148	-0.101	0.049	-0.084
Has store installments	-0.167	-0.101	0.000	0.000
Number of liable individuals	-0.260	-0.197	0.027	0.107
Has co-applicant	-0.314	-0.135	-0.120	0.000
Has bank installments	-0.374	-0.217	-0.069	0.017
Unemployed	-0.376	-0.170	-0.058	-0.123
Owns no property	-0.382	-0.371	-0.346	-0.124
Installment rate	-0.507	-0.450	-0.286	-0.109
No credits or credit repaid duly	-0.549	-0.419	-0.284	-0.106
Credits repaid duly at this bank	-0.623	-0.397	**-0.248**	0.000
<100 DM savings account	**-0.646**	**-0.434**	-0.235	**-0.176**

Figure 5.6: Average changes in actionable features in the German credit dataset after individuals with bad credit were interpolated (at level α) to their counterparts with good credit.

Chapter Six

Gromov-Wasserstein Alignment of Single-Cell Multi-Omics Data

This chapter is based on work presented in Demetci et al. (2021).

6.1 Introduction

Single-cell measurements provide a fine-grained view of the heterogeneous landscape of cells in a sample, revealing distinct subpopulations and their developmental and regulatory trajectories across time. The availability of single-cell measurements that capture various properties of the genome, such as gene expression, chromatin accessibility, DNA methylation, histone modifications, and chromatin 3D conformation, has increased the need for data integration methods capable of combining disparate data types.

Despite the importance of this task, the heterogeneity among single cells presents unique challenges. Due to technical limitations, it is hard to obtain multiple types of measurements from the same individual cells. As a result, many single-cell multi-omics datasets have distinct measurements from different sets of cells. While certain methods combine different experiments from a single modality (like RNA sequencing) for correcting batch effects Amodio and Krishnaswamy (2018), Welch et al. (2019, 2017), Stuart et al. (2019), Barkas et al. (2019), integrating data from multiple modalities (like RNA and ATAC sequencing) further complicates the task. For example, when we measure different properties of a cell, we cannot *a priori* identify correspondences between features in the two domains. Accordingly, integrating two or more single-cell data modalities requires methods that rely on neither common cells nor features across the data types. This aspect prevents the application of some existing single-cell alignment methods to unsupervised settings because they require some correspondence information Amodio and Krishnaswamy (2018), Welch et al.

(2019, 2017), Stuart et al. (2019), Barkas et al. (2019). For example, Seurat Stuart et al. (2019) requires correspondence information in the form of cells from similar biological state that are shared across the two datasets (known as "anchor points"). Cao *et al.* Cao et al. (2020) have shown that such methods cannot perform good alignments under unsupervised settings.

Multiple approaches have tried to align datasets in an entirely unsupervised fashion. One of the earliest attempts, the joint Laplacian manifold alignment (JLMA) algorithm, constructs eigenvector projections based on k-nearest neighbor graph Laplacians of the data Wang and Mahadevan (2009). The generalized unsupervised manifold alignment (GUMA) Cui et al. (2014) algorithm seeks a 1–1 correspondence between two datasets based on a local geometry matching term. Liu *et al.* Liu et al. (2019) showed that these methods do not perform well on the single-cell alignment task and proposed an alignment algorithm based on the maximum mean discrepancy (MMD) measure, called MMD-MA. Another method, UnionCom Cao et al. (2020), extends GUMA to perform unsupervised topological alignment and makes it more suitable for single-cell multi-omics integration. While MMD-MA aims to match the global distributions of the datasets in a shared latent space, UnionCom emphasizes learning both local and global alignments between the two distributions. Neither method requires any correspondence information, either among samples or features, to perform an alignment. The respective papers demonstrate state-of-the-art performance on simulated and real datasets. Although these results are encouraging, MMD-MA and UnionCom require that the user specify three and four hyperparameters, respectively. Selecting these hyperparameter values can be a barrier to adopting tools in an unsupervised setting with no validation data.

In this paper, we propose an unsupervised alignment method based on optimal transport. Optimal transport finds the most cost-effective way to move data points

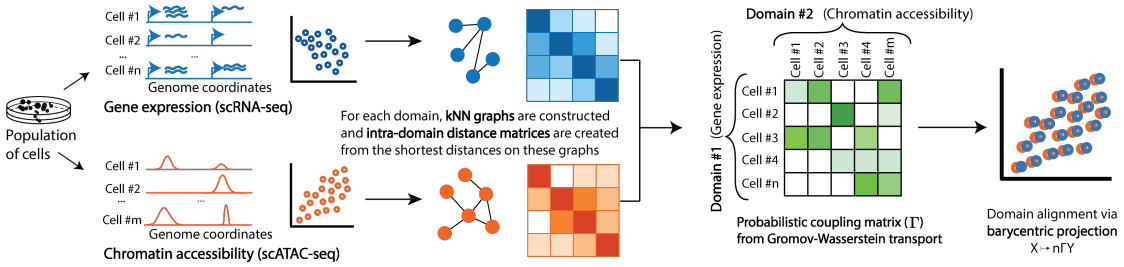

Figure 6.1: **Schematic of SCOT alignment of single-cell multi-omics data.** A population of cells is aliquoted for different single-cell sequencing assays. SCOT constructs k-NN graphs based on sample-wise correlations and finds a probabilistic coupling between the samples of each domain that minimizes the distance between the two intra-domain graph distance matrices. Barycentric projection projects one domain onto another based on this coupling matrix.

from one domain to another. One way to think about it is as the problem of moving a pile of sand to fill in a hole through the least amount of work. An emerging number of applications, including several in biology, are using optimal transport to learn a mapping between data distributions Alvarez-Melis and Jaakkola (2018), Schiebinger et al. (2019), Cang and Nie (2020), Yang et al. (2018), Yang and Uhler (2019). Schiebinger *et al.* Schiebinger et al. (2019) use it to study temporal changes in gene expression by using regularized unbalanced optimal transport to compute expression differences between time points . SpaOTsc Cang and Nie (2020) maps cells with high ligand expression onto cells with high receptor expression to recover cell signaling relationships in spatially resolved single-cell RNA-seq datasets. ImageAEOT Yang et al. (2018) maps single-cell images to a common latent space through an autoencoder and then uses optimal transport to track cell trajectories. In related work, the same authors use autoencoders and optimal transport to learn transport maps among multiple domains Yang and Uhler (2019). However, the application of their method to single-cell datasets requires some form of supervision, like class labels, to be used during transport.

The classic optimal transport problem requires datasets in the same metric

space. Mémoli *et al.* Mémoli (2011) generalized optimal transport to the Gromov-Wasserstein distance, which compares metric spaces directly instead of comparing samples across spaces and makes optimal transport suitable for multi-modal alignment. In natural language processing, Alvarez *et al.* Alvarez-Melis and Jaakkola (2018) used this approach to measure similarities between pairs of words across languages to compute the similarity between languages. As far as we are aware, the only biological application of Gromov-Wasserstein optimal transport comes from Nitzan et al. (2019), which uses it to reconstruct the spatial organization of cells from transcriptional profiles.

We present Single-Cell alignment using Optimal Transport (SCOT), an unsupervised algorithm that uses Gromov-Wasserstein-based optimal transport to align single-cell multi-omics datasets (presented schematically in Figure 6.1). Like UnionCom, SCOT aims to preserve local geometry when aligning single-cell data. SCOT achieves this by constructing a k-nearest neighbor ($k-$NN) graph for each dataset. SCOT then finds a probabilistic coupling between the samples of each dataset that minimizes the distance between the graph distance matrices produced from the k-NN graph. Finally, it uses the coupling matrix to project one single-cell dataset onto another through barycentric projection, thus aligning them. Unlike MMD-MA and UnionCom, SCOT requires tuning only two hyperparameters and is robust to the choice of one. We compare the alignment performance of SCOT with MMD-MA and UnionCom on four simulated and two real-world datasets. We show that SCOT aligns datasets as well as the state-of-the-art methods and scales well with increasing numbers of samples. Moreover, we demonstrate that, the Gromov-Wasserstein distance can guide SCOT's hyperparameter tuning in a fully unsupervised setting when no orthogonal alignment information is available. Thus, unlike other methods, SCOT provides a heuristic for hyperparameter selection without validation data.

Our source code and scripts for replicating the results are available at https://github.com/rsinghlab/SCOT. Documentation and tutorial for using SCOT is available at https://rsinghlab.github.io/SCOT.

6.2 Methods

SCOT uses Gromov-Wasserstein optimal transport, which preserves local geometry when moving data points form one domain to another. The output of this transport problem is a matrix of probabilities that represent how likely it is that data points from one modality correspond to data points in the other. In the next section, we introduce optimal transport followed by its extension to the Gromov-Wasserstein distance. Finally, we present the details of our SCOT algorithm. We present the case for two datasets: $X^1 = (x_1^1, x_2^1, \ldots, x_{n_1}^1)$ from $\mathcal{X}^1$ and $X^2 = (x_1^2, x_2^2, \ldots, x_{n_2}^2)$ from $\mathcal{X}^2$. The datasets have n_1 and n_2 points, respectively. We do not require any correspondence information or assume that there is any ground truth correspondence, but we do assume there is some underlying shared structure so that the datasets can be meaningfully aligned.

6.2.1 Gromov-Wasserstein distance

Classic optimal transport, introduced in Section 5.2.2, requires defining a cost function across domains, which can be difficult to implement when the domains are in different metric spaces. Gromov-Wasserstein distance extends optimal transport by comparing distances between samples rather than directly comparing the samples themselves Alvarez-Melis and Jaakkola (2018). We assume that we have metric

measure spaces $(\mathcal{X}^1, d_1, \mu)$ and $(\mathcal{X}^2, d_2, \nu)$, where d_1 and d_2 are distances on $\mathcal{X}^1$ and $\mathcal{X}^2$, respectively Mémoli (2011). Instead of defining a cost function between spaces, Gromov-Wasserstein uses the difference between pairwise distances. Given a cost function $L : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$, the Gromov-Wasserstein distance between μ and ν is defined by

$$GW(\mu, \nu) := \min_{\pi \in \Pi(\mu,\nu)} \int_{\mathcal{X}^1 \times \mathcal{X}^2} \int_{\mathcal{X}^1 \times \mathcal{X}^2} L(d_1(x_i^1, x_j^2), d_2(x_k^2, x_l^2)) d\pi(x_i^1, x_k^2) d\pi(x_j^1, x_l^2). \tag{6.2.1}$$

In contrast to basic optimal transport (Equation 5.2.1), Gromov-Wasserstein (Equation 6.2.1) optimal transport considers the effect of transporting pairs of points rather than single points. Intuitively, $L(d_1(x_i^1, x_j^2), d_2(x_k^2, x_l^2))$ captures how transporting x_i^1 to x_k^2 and x_j^1 to x_l^2 would distort the original distances between x_i^1 and x_j^1 and between x_k^2 and x_l^2. This change ensures that the optimal transport plan π will preserve some local geometry. In the case of $L(x, y) = L_2(x, y) = \frac{1}{2}(x - y)^2$, Gromov-Wasserstein is a distance on the space of metric measure spaces Mémoli (2011).

For the discrete case, we compute pairwise distance matrices D^1 and D^2 and the fourth order tensor $\mathbf{L} \in \mathbb{R}^{n_1 \times n_1 \times n_2 \times n_2}$, where $\mathbf{L}_{ijkl} = L(D^1_{ik}, D^2_{jl})$. The discrete Gromov-Wasserstein problem is

$$GW(p^1, p^2) = \min_{\Gamma \in \Pi(p^1, p^2)} \sum_{i,j,k,l} \mathbf{L}_{ijkl} \Gamma_{ij} \Gamma_{kl}. \tag{6.2.2}$$

The summation can also be expressed as the inner product $\langle \mathbf{L}(D^1, D^2) \otimes \Gamma, \Gamma \rangle$. Equation 6.2.2 is now both non-linear and non-convex and involves operations on a fourth-order tensor, including the $\mathcal{O}(n_1^2 n_2^2)$ operation tensor product $L(D^1, D^2) \otimes \Gamma$ for a naive implementation. Peyré *et al.* show that for some choices of loss function this product can be computed in $\mathcal{O}(n_1^2 n_2 + n_1 n_2^2)$ cost Peyré et al. (2016). In particular,

for the case $L = L_2$, the inner product can be computed by

$$\mathbf{L}(D^1, D^2) \otimes \Gamma = (D^1)^2 p^1 \mathbb{1}_{n_2}^T + \mathbb{1}_{n_1} (p^2)^T ((D^2)^2)^T - D^1 \Gamma (D^2)^T. \tag{6.2.3}$$

As in the classic optimal transport case, the coupling matrix can be efficiently computed for an entropically regularized optimization problem:

$$GW_\epsilon(p^1, p^2) = \min_{\Gamma \in \Pi(p^1, p^2)} \langle \mathbf{L}(D^1, D^2) \otimes \Gamma, \Gamma \rangle - \epsilon H(\Gamma). \tag{6.2.4}$$

Larger values of ϵ lead to an easier optimization problem but also a denser coupling matrix, meaning that solutions will indicate significant correspondences between more data points. Smaller values of ϵ lead to sparser solutions, meaning that the coupling matrix is more likely to find the correct one-to-one correspondences for datasets where there are one-to-one correspondences. However, it also yields a harder (more non-convex) optimization problem Alvarez-Melis and Jaakkola (2018).

Peyré *et al.* Peyré et al. (2016) propose using a projected gradient descent approach for optimization, where both the projection and the gradient are taken with respect to Kullback-Leibler divergence. These projections are computed via Sinkhorn iterations. Algorithm 1 in presents the algorithm for $L = L_2$.

6.2.2 Single-Cell alignment using Optimal Transport (SCOT)

Our method, SCOT, works as follows - first, we compute the pairwise distances on our data in a way similar to Nitzan et al. (2019). To do this, we use the correlations

Algorithm 1: SCOT Alignment with Gromov-Wasserstein OT

Inputs: Datasets $X^1,\ X^2$. Regularization coefficient ϵ. Number of neighbors k.

// Compute graph distances $D_1,\ D_2$; Set $p^1 \leftarrow Uniform(X^1), p^2 \leftarrow Uniform(X^2)$;

$D_{12} \leftarrow D_1^2 \mathbb{1}_{n_2}^T + \mathbb{1}_{n_1} p^2 (D_1^2)^T$;

while *not converged* **do**

  $\hat{D}_\Gamma \leftarrow D_{12} - 2 D_1 \Gamma D_2^T$;

  // Compute cost matrix

  $u \leftarrow \mathbb{1}_1,\ K \leftarrow \exp\{-\hat{D}_\Gamma / \epsilon\}$; // Perform Sinkhorn iterations

  while *not converged* **do**

    $u \leftarrow p^1 \oslash Kv,\ v \leftarrow p^2 \oslash K^T u$;

  end

  $\Gamma \leftarrow \mathrm{diag}(u) K \mathrm{diag}(v)$;

end

Return: $n_1 \Gamma X^2$

between data points

$$r(x, y) = 1 - \frac{(x - \bar{x})\dot{(}y - \bar{y})}{\|(x - \bar{x})\|_2 \, \|(y - \bar{y})\|_2}, \tag{6.2.5}$$

within each dataset to construct k-NN connectivity graphs. We find that connectivity graphs, which connect nodes with binary edges, empirically perform better than weighted edges, potentially because connectivity graphs can denoise the data. Next, we compute the shortest path distance on the graph between each pair of nodes via Dijkstra's algorithm. We set the distance of any unconnected nodes to be the maximum finite distance in the graph and normalize the matrix by dividing the elements by this maximum distance. If k is the number of samples, then the k-NN graph is the complete graph, so the corresponding distance matrix is a matrix of all ones.In this case, the distance matrix does not provide information about the local geometry, so we recommend keeping k small relative to the number of samples to avoid this scenario. We find that our approach is robust to the choice of k (Supplementary Figure 7.6B).

Since we do not know the true distribution of the original datasets, we follow Alvarez-Melis and Jaakkola (2018) and set μ and ν to be the uniform distributions on the data points. Then, we solve for the optimal coupling Γ which minimizes Equation 6.2.4. To implement this method, we use the Python Optimal Transport toolbox (`https://pot.readthedocs.io/en/stable/`) Flamary and Courty (2017).

One of the advantages of using optimal transport is the probabilistic interpretation of the resulting coupling matrix Γ, where the entries of the normalized row $\frac{1}{p_i^1}\Gamma_i$ are the probabilities that the fixed data point x_i^1 corresponds to each x_j^2. However, to use the evaluation metrics previously used in the field and to visualize alignment, we need to project the two datasets into the same space. The Procrustes approach proposed in Alvarez-Melis and Jaakkola (2018) does not generalize to datasets with different feature and sample dimensions, so we use a barycentric projection:

$$x_i^1 \mapsto \frac{1}{p_i^1} \sum_{j=1}^{n_2} \Gamma_{ij} x_j^2. \tag{6.2.6}$$

6.2.3 Alternative Unsupervised Alignment Procedure

In the description of SCOT, the number k for nearest neighbors and the entropy weight ϵ are hyperparameters. One way to set these hyperparameters for optimal alignment is to use some orthogonal correspondence information to select the best alignment either directly Cao et al. (2020), Liu et al. (2019) or by performing cross-validation Singh et al. (2020). This selection strategy is problematic for truly unsupervised setting, where no correspondence information is available *a priori* upon sequencing separate cell cultures. As a solution, we provide an alternative procedure to learn reasonable alignments based on tracking the Gromov-Wasserstein distance (Equation 6.2.2). This procedure is based on our observation that the Gromov-

Wasserstein distance serves as a proxy for measuring alignment quality (Supplementary Figure 7.6A). In this procedure, we alternate between optimizing ϵ and k to minimize the Gromov-Wasserstein distance between the domains (Algorithm 2). Although the lowest Gromov-Wasserstein distance is not always the best alignment, it consistently appears to be one of the better alignments.

Algorithm 2: Unsupervised hyperparameter search procedure

Input: Datasets $X^1,\ X^2$.
$n \leftarrow \min(n_1, n_2),\ k_1 \leftarrow \min(0.2n, 50)$
$\epsilon_1 \leftarrow \arg\min_{\epsilon \in [10^{-3}, 10^{-2}]} \text{SCOT}(X^1, X^2, k_1, \epsilon)$ // Fix k_1 and vary ϵ
// Fix ϵ_1 and vary k
if $n > 250$ **then**
  $k_2 \leftarrow \arg\min_{k \in [20, 100]} \text{SCOT}(X^1, X^2, k, \epsilon_1)$
end
else
  $k_2 \leftarrow \arg\min_{k \in [0.05n, 0.2n]} \text{SCOT}(X^1, X^2, k, \epsilon_1)$
end
// Do a more refined search around k_2 and ϵ_1
$k_{\text{best}}, \epsilon_{\text{best}} \leftarrow \arg\min_{k \in [k_2 - 5, k_2 + 5],\ \epsilon \in [10^{-0.25}\epsilon_1, 10^{0.25}\epsilon_1]} \text{SCOT}(X^1 X^2, k, \epsilon)$
Return: $k_{\text{best}}, \epsilon_{\text{best}}$

6.3 Experimental Setup

6.3.1 Simulated datasets

We follow Liu *et al.* Liu et al. (2019) and benchmark SCOT on three different simulations[1]. All three simulations contain two domains with 300 samples that have been non-linearly projected to 1000- and 2000-dimensional feature spaces, respectively. The three simulations are a bifurcation, a Swiss roll, and a circular frustum (Appendix Figure 7.2) with points belonging to three different groups. The first simulation is a bifurcated tree in two-dimensional space, simulating a branching

[1] https://noble.gs.washington.edu/proj/mmd-ma/

structure one might observe in cell differentiation. The second simulation maps the branching structure onto a Swiss roll in three-dimensional space through a non-linear projection. The third simulation is a circular frustum in three-dimensional space, which simulates the cell cycle superimposed on a linear differentiation process (Supplementary Figure 7.2). In addition to these three previously existing simulations, we use Splatter Zappia et al. (2017) to create simulated single-cell RNA sequencing count data, which we call synthetic RNA-seq. We generate 5000 cells with 1000 genes from three cell groups and reduce the count matrix to the five genes with the highest variances. This count matrix is mapped into two new domains with dimensions $p_1 = 50$ and $p_2 = 500$ by multiplying it with two randomly generated matrices, resulting in data with dimensions 5000×50 and 5000×500.

All four datasets were simulated with 1—1 sample-wise correspondences, which are solely used for evaluating model performance. Since each domain is projected to a different dimension, there is no feature-wise correspondence. In all simulations, we Z-score normalize the features before alignment as in Liu et al. (2019).

6.3.2 Single-cell multi-omics datasets

We use two sets of single-cell multi-omics data to demonstrate SCOT's performance on real datasets. Both datasets are generated by co-assays; thus, we have known cell-level correspondence information for benchmarking. The first dataset is generated using the scGEM assay Cheow et al. (2016), which simultaneously profiles gene expression and DNA methylation. The dataset (Sequence Read Archive accession SRP077853) is derived from human somatic cell samples undergoing conversion to induced pluripotent stem cells (iPSCs) and show a continuous trajectory (Appendix Figure 7.4). This dataset was also used by Cao *et al.* Cao et al. (2020) to demon-

strate the performance of their UnionCom algorithm. We pre-process the data as described in the original publications Cheow et al. (2016), Cao et al. (2020), resulting in data with dimensions 177×34 for the gene expression data and 177×27 for the chromatin accessibility data.

The second dataset is generated by the SNAREseq assay Chen et al. (2019), which links chromatin accessibility with gene expression. The data (Gene Expression Omnibus accession GSE126074) is derived from a mixture of human cell lines: BJ, H1, K562, and GM12878 and show distinct cell type clusters (Appendix Figure 7.4). We pre-process the datasets following Chen *et al.* Chen et al. (2019). The resulting data matrices are of size 1047×19 for ATAC-seq and 1047×10 for RNA-seq. We unit normalize all real datasets as done in Singh et al. (2020).

6.3.3 Evaluation metrics

We compare SCOT with the two state-of-the-art unsupervised single-cell alignment methods MMD-MA Liu et al. (2019) and UnionCom Cao et al. (2020). None of these methods use any correspondence information for aligning the datasets. However, all datasets have 1–1 sample-level correspondence information, which we use to quantify the alignment performance through the "fraction of samples closer than the true match" (FOSCTTM) metric introduced by Liu *et al.* Liu et al. (2019). For each domain, we compute the Euclidean distances between a fixed sample point and all the data points in the other domain. Next, we use these distances to compute the fraction of samples that are closer to the fixed sample than its true match. Finally, we average these values for all the samples in both domains. For perfect alignment, all samples would be closest to their true match, yielding an average FOSCTTM of zero. Therefore, a lower average FOSCTTM corresponds to a better alignment.

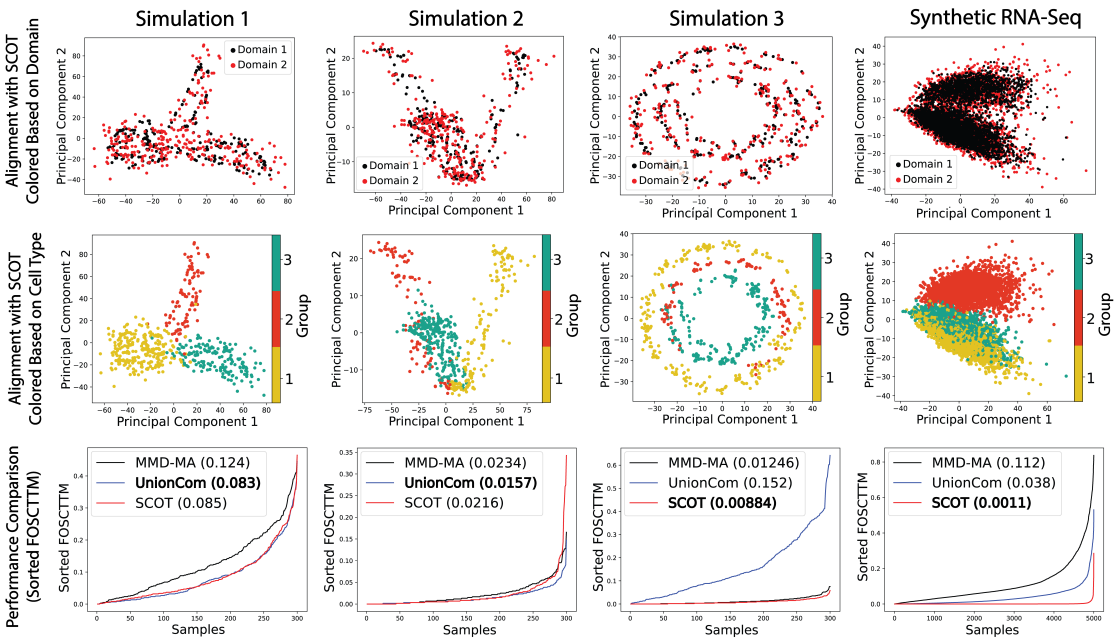

Figure 6.2: **Aligning simulated datasets.** Each column presents a different simulation. **Top:** our alignment colored by domain (plotted in 2D using PCA). **Middle:** our alignment colored by group. **Bottom:** sorted "fraction of samples closer than the true match" (FOSCTTM) for MMD-MA, UnionCom, and SCOT to visualize the distribution across samples with the average FOSCTTM values in the legend.

Since all the datasets have group-specific (simulations) or cell-type-specific (real experiments) labels, we also adopt the metric used by Cao *et al.* Cao et al. (2020) called "label transfer accuracy" to assess the quality of the cell label assignment. This metric measures the ability to correctly transfer sample labels from one domain to another based on their neighborhood in the aligned domain. As described in Cao et al. (2020), we train a k-nearest neighbor classifier (for $k = 5$, default parameter used in Cao et al. (2020)) on one of the domains and predict the sample labels in the other domain. The label transfer accuracy is the proportion of correctly predicted labels, so it ranges from 0 to 1, and higher values indicate good performance. We apply this metric to alignments selected by the FOSCTTM measure.

6.3.4 Hyperparameter tuning

We run each method over a grid of hyperparameters and select the setting that yields the lowest average FOSCTTM. For SCOT, the grid covers the regularization weight $\epsilon \in \{0.0001, 0.0005, 0.001, 0.005, ..., 0.1\}$ and number of neighbors $k \in \{10, 15, 20, \ldots 100, \frac{1}{6}n_x\}$. We observe empirically that going above $\frac{1}{6}n$ for k does not yield any improvement in alignment.

We pick the hyperparameters for MMD-MA and UnionCom based on the default values and recommended ranges. MMD-MA has three hyperparameters: weights $\lambda_1, \lambda_2 \in \{10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}, 10^{-7}\}$ for the terms in the optimization problem and the dimensionality $p \in \{4, 5, 6, 16, 32, 64\}$ of the embedding space. UnionCom requires the user to specify four hyperparameters: the number $k_{max} \in \{40, 100\}$ of maximum number of neighbors in the graph,the dimensionality $p \in \{4, 5, 6, 16, 32, 64\}$ of the embedding space, the trade-off parameter $\beta \in \{0.1, 1, 10, 15, 20\}$ for the embedding, and a regularization coefficient $\rho \in \{0, 5, 10, 15, 20\}$. We select the embedding dimension $p \in \{16, 32, 64\}$ around the default value of 32 set by UnionCom but also add $p \in \{4, 5, 6\}$ to match the recommended values for MMD-MA. We keep the hyperparameter search space size approximately consistent across the three methods. For each dataset, we present alignment and runtime results for the best performing hyperparameters. Considering that MMD-MA and UnionCom both have more hyperparameters to tune, we later approximately doubled the search space by choosing intermediary values but did not observe any significant improvement in their alignment quality (up to three significant figures).

Furthermore, we consider the scenario where correspondence information is unavailable to pick the optimal hyperparameters. For SCOT, we apply the alternative unsupervised alignment procedure (Algorithm 2 in Appendix) to align all the

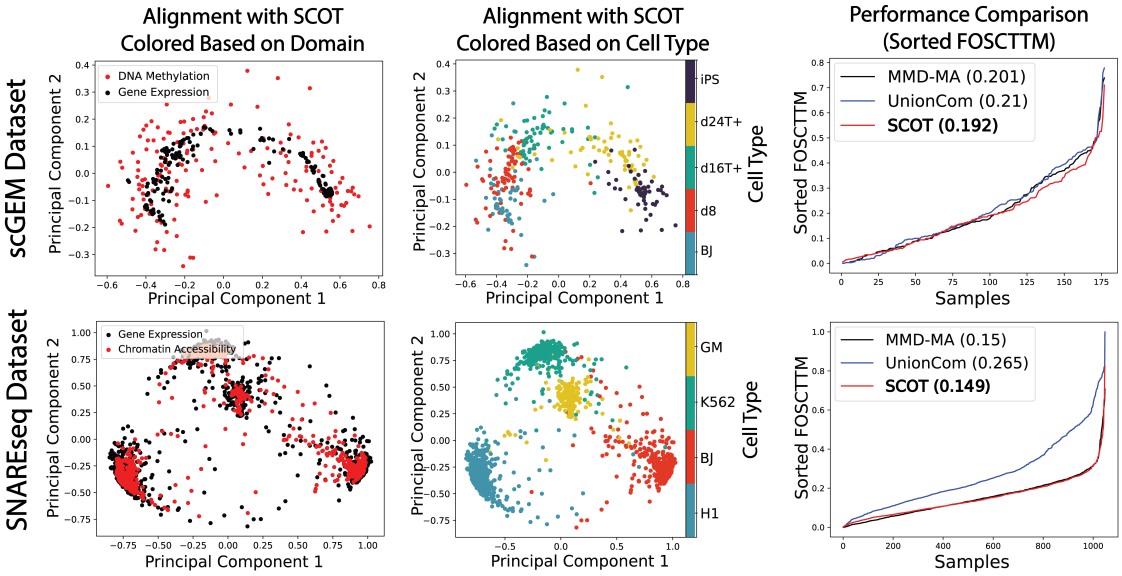

Figure 6.3: **Aligning real world single-cell sequencing dataset.** Each row presents a different real-word single-cell sequencing dataset. **Left:** our alignment colored based by domain (plotted in 2D using PCA). **Middle:** our alignment colored by cell-type. **Right:** sorted "fraction of samples closer than the true match" (FOSCTTM) for MMD-MA, UnionCom, and SCOT order to visualize the distribution across samples with the average FOSCTTM values in the legend.

datasets. Since MMD-MA and UnionCom do not provide a hyperparameter selection strategy, we rely on the default hyperparameters; we use UnionCom's provided default parameters of $k_{max} = 40, p = 32, \rho = 10$, and $\beta = 1$, and the center values of MMD-MA's recommended range: $p = 5, \lambda_1 = 10^{-5}$, and $\lambda_2 = 10^{-5}$. We also present the alignment results for all three methods in this fully unsupervised setting.

6.4 Results

6.4.1 SCOT successfully aligns the simulated datasets

We first compare SCOT's performance with MMD-MA and UnionCom for the four simulation datasets. In this experiment, we select the best performing hyperparam-

eters for each method using the tuning process described in the previous section. In Figure 6.2, we sort and plot the FOSCTTM score for each sample for the simulations from Liu et al. (2019), as well as the synthetic RNA-seq count data from Splatter Zappia et al. (2017). Overall, we observe that SCOT consistently achieves one of the lowest average FOSCTTM scores, thereby demonstrating its ability to recover the correct correspondences. We also report the label transfer accuracy results (Table 6.1) when the first domain is used to train a classifier to predict the labels in the second domain. We observe that SCOT consistently yields high label transfer accuracy scores, indicating that samples are correctly mapped to their assigned groups.

6.4.2 SCOT gives state-of-the-art single-cell multi-omics alignment

Next, we apply our method to real single-cell sequencing data. Overall, SCOT gives the lowest average FOSCTTM measure in comparison to MMD-MA and UnionCom (Figure 6.3, last column) and recovers accurate 1–1 correspondences in single-cell datasets. For the scGEM data, we report label transfer accuracy using the DNA methylation domain for predicting the cell-type labels in the gene expression domain. For the SNARE-seq dataset, we use the gene expression domain for predicting cell labels in the chromatin accessibility domain. SCOT yields the best label transfer

	Sim. 1	**Sim. 2**	**Sim. 3**	**Syn. RNA-seq**	**scGEM**	**SNARE seq**
SCOT	0.950	0.977	**0.957**	**0.998**	**0.593**	**0.982**
MMD-MA	0.890	0.913	0.947	0.706	0.588	0.942
UnionCom	**0.960**	**0.986**	0.613	0.997	0.582	0.423

Table 6.1: Alignment performance by label transfer accuracy ($k = 5$) with the classifier trained on the first domain of each dataset.

accuracy result on both datasets (Table 6.1).

While MMD-MA and UnionCom project both datasets to a shared low-dimensional space, SCOT projects one dataset onto the other. We project SCOT in both directions for all datasets, but we do not observe a significant difference in performance between the two directions.

6.4.3 Unsupervised hyperparameter tuning aligns well

We compare the alignment performances in Table 6.2 given by SCOT's alternative tuning procedure guided by the Gromov-Wasserstein distance against MMA-MA's and UnionCom's default parameters. SCOT returns nearly the same alignments for simulated data and only marginally worse alignments for real data. In contrast, MMD-MA and UnionCom fail to align some of the simulated and all real datasets with the default parameter values. In particular, MMD-MA and UnionCom perform very badly on the simulation 3, scGEM, and SNAREseq datasets, and MMD-MA performs worse than its optimal alignment on the synthetic RNA-Seq dataset. Therefore, the proposed procedure could guide a user to an alignment close to the optimal result when no orthogonal information is available for hyperparameter tuning.

	Sim. 1	**Sim. 2**	**Sim. 3**	**Syn. RNA-seq**	**scGEM**	**SNARE seq**
SCOT (GW)	**0.088**	0.025	**0.009**	**0.001**	**0.209**	**0.218**
MMD-MA	0.125	**0.023**	0.739	0.384	0.437	0.473
UnionCom	0.091	0.028	0.684	0.028	0.691	0.510

Table 6.2: Alignment performance by FOSCTTM scores for SCOT chosen by lowest Gromov-Wasserstein distance, default MMD-MA, and default UnionCom for simulated and real datasets.

6.4.4 SCOT computationally scales well

We compare SCOT's running times with the baseline methods for the best performing hyperparameters on the synthetic RNA-seq dataset by varying the number of cells. We run CPU computations on an Intel Xeon e5-2670 with 16GB memory and GPU computations on a single NVIDIA GTX 1080ti with VRAM of 11GB. SCOT's running time scales similarly to that of MMD-MA, even though SCOT runs on a CPU and MMD-MA runs on a GPU (Figure 6.4). Both methods scale better than the GPU-based UnionCom implementation.

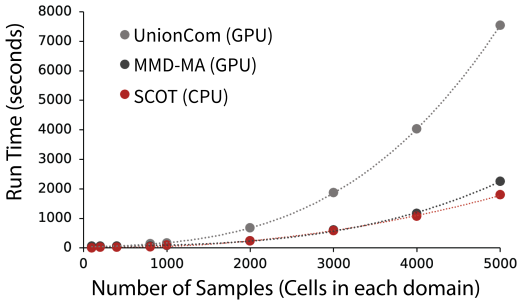

Figure 6.4: **Runtime comparisons with growing sample size**

6.5 Discussion

We have demonstrated that SCOT, which uses Gromov Wasserstein optimal transport for unsupervised single-cell multi-omics data integration, yields alignments on par with the state-of-the-art methods, UnionCom and MMD-MA, and has advantages over them when no orthogonal information exists for hyperparameter tuning. Our formulation of a coupling matrix based on matching graph distances is somewhat similar to UnionCom's initial step; however, UnionCom only matches sample-to-sample distances, while Gromov-Wasserstein distance considers the cost of moving pairs of points, enabling our method to better preserve local geometry. Additionally, SCOT performs global alignment of the marginal distributions, which is similar to how MMD-MA uses the MMD term to ensure that the two distributions agree

globally in the latent space. We hypothesize that these properties result in SCOT's state-of-the-art performance. Furthermore, SCOT's optimization runs in less time and with fewer hyperparameters, and the Gromov-Wasserstein distance can guide the user to choose an alignment when no validation information exists. Therefore, unlike other methods, SCOT easily yields high quality alignments in the realistic fully unsupervised setting.

While barycentric projection provides a way to visualize the alignment, it assumes that cells in one dataset should be mapped to the convex hull of the other dataset. Future work will develop unbalanced optimal transport, which would take care of outliers as well as under- or over-represented groups. There are also other ways to use the coupling matrix to infer alignment such as using it with other dimension reduction methods like t-SNE (as in UnionCom) to align the manifolds while embedding them both into a new space. Alternatively, depending on the application, a projection may not be required; it may be sufficient to have probabilities relating the samples to one another. Future work will develop effective ways to utilize the coupling matrix and extend our framework to handle more than two alignments at a time.

Chapter Seven

Integration of Datasets with Cell-Type Representation Disparities

This chapter is based on work presented in Demetci et al. (2022).

7.1 Introduction

The ability to measure multiple aspects of the single-cell offers the opportunity to gain critical biological insights about cell development and diseases. However, many existing single-cell sequencing technologies cannot be simultaneously applied to the same cell, resulting in multi-omics datasets sampled from distinct cell populations. While these measurements can be analyzed separately, integrating them prior to analysis can help explain how different molecular views interact and regulate cellular functions. Unfortunately, single-cell assays that measure different molecular aspects in separately sampled cell populations lack direct sample–sample and feature–feature correspondences across these measurements. This lack of correspondences makes it hard to use integration methods that require some shared information to perform single-cell alignment Cao et al. (2020). Therefore, *unsupervised* single-cell multi-omics data alignment methods are crucial for integrative single-cell data analysis.

Several unsupervised methods Liu et al. (2019), Cao et al. (2020), Dou et al. (2020), Stuart et al. (2019), including our previous work, SCOT Demetci et al. (2021), have shown state-of-the-art performance for integrating different single-cell measurement domains. Since these methods were mainly evaluated on real-world co-assay datasets (with 1–1 correspondence between cells across domains), our understanding of their performance on datasets obtained from experiments that are not co-assays is limited. Such experiments perform separate sampling to measure distinct genomic features, like gene expression and 3D chromatin conformation. Therefore, their datasets can consist of varying proportions of cell-types across different mea-

surements, creating cell-type imbalance and lacking 1–1 cell correspondences. We hypothesize that alignment methods that perform well on co-assay datasets may not effectively handle the differences in cell-type proportions of the commonly available non-co-assay datasets. Indeed, a recent method, Pamona Cao et al. (2021), extended our SCOT framework and used partial Gromov-Wasserstein (GW) optimal transport to allow for missing or underrepresented cell-types in one domain when performing alignment. The paper showed that current integration methods Demetci et al. (2021), Liu et al. (2019), Cao et al. (2020), Stuart et al. (2019) tend to perform worse under such settings.

We present SCOTv2, a novel extension of SCOT that can effectively align both co-assay and non-co-assay datasets using a single framework. It uses *unbalanced* GW optimal transport to align datasets with disproportionate cell-types while only introducing one additional hyperparameter. This unbalanced framework relaxes the constraint that each point must be mapped with its original mass during the optimal transport. Specifically, an underrepresented cell-type in one domain can be transported with more mass to match the proportion of that cell-type in the other domain and vice-versa. The SCOTv2 framework is summarized in Figure 7.1. We demonstrate that SCOTv2 aligns datasets with imbalance in cell-type representations better than state-of-the-art baselines and computationally scales as well as the fastest methods. Furthermore, we extend SCOTv2 to integrate single-cell datasets with more than two measurements, making it a multi-omics alignment tool. We perform alignments of five real-world single-cell datasets, with both simulated and natural cell-type imbalance as well as two and more than two domains ($M \geq 2$), demonstrating SCOTv2's applicability across a wide range of scenarios. Finally, similar to the previous version, we present a self-tuning heuristic process to select hyperparameters for SCOTv2 without any corresponding information like cell-type

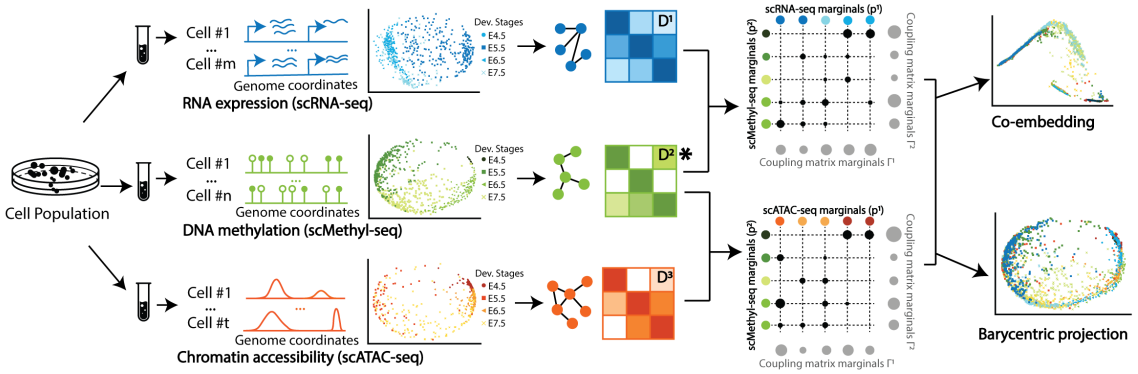

Figure 7.1: **Overview of SCOTv2 on scNMT-seq dataset** Clark et al. (2018), which contains unbalanced cell-type representation across three domains - RNA expression, chromatin accessibility, and DNA methylation. SCOTv2 selects an anchor domain (denoted with *) and aligns other measurements to it. First, it computes intra-domain distances matrices D^m for $m = 1, 2, 3$, which are used to solve for correspondence matrices between the anchor and other domains. The circle sizes in the matrices depict the magnitude of the correspondence probabilities or how much mass to transport. Unbalanced GW relaxes the mass conservation constraint, so the transport map does not need to move each point with its original mass. Finally, it either co-embeds the domains into a common space or uses barycentric projections to project them onto the anchor domain.

annotations or matching cells or features in truly unsupervised settings.

7.2 Method

Here, we reintroduce the SCOT formulation to account for unbalanced datasets and integrate M domains (or single-cell measurements) $X^m = (x_1^m, x_2^m, \ldots x_{n_m}^m) \in \mathbb{R}^{d_m}$ for $m = 1, \ldots M$ with n_m data points (or cells) each.

7.2.1 Unbalanced Optimal Transport of SCOTv2

In Section 6.2.1, we introduced Gromov-Wasserstein optimal transport as the basis for SCOT. Our proposed improvement to align datasets with different numbers of samples or proportions of cell-types is to instead use unbalanced optimal transport, which adds divergence terms to allow for mass variations in the marginals Liero et al. (2018), Séjourné et al. (2021). We follow Séjourné *et al* Séjourné et al. (2021), and use the Kullback-Leibler divergence ,

$$\mathrm{KL}(p||q) = \sum_x p(x) \log\left(\frac{p(x)}{q(x)}\right), \tag{7.2.1}$$

to measure the difference between the marginals of the coupling Γ and the input marginals p^1 and p^2. Thus, we solve the unbalanced GW problem:

$$GW_{\epsilon,\rho}(p^1, p^2) = \min_{\Gamma \geq 0} \langle \mathbf{L}(D^1, D^2) \otimes \Gamma, \Gamma \rangle - \epsilon H(\Gamma) + \rho \mathrm{KL}(\Gamma \mathbb{1}_{n_2} || p^1) + \rho \mathrm{KL}(\Gamma^T \mathbb{1}_{n_1} || p^2), \tag{7.2.2}$$

where $\rho > 0$ is a hyperparameter that controls the marginal relaxation. When ρ is large, the marginals of Γ should be close to p^1 and p^2, and when ρ is small, the marginals of Γ may differ more, allowing each point to transport with more or less mass than it originally had. See Algorithm 3 for implementation details.

7.2.2 Extending SCOTv2 for Multi-Domain Alignment

To align more than two datasets ($M > 2$), we use one domain as an anchor to align the other domains. The anchor should be the domain with the clearest biological structures, for example, a dataset with the best-defined cell-type clusters. We propose selecting the anchor via the kNN graph used to compute D^m. For every node

Algorithm 3: Pseudocode for Unbalanced GW Optimal Transport (UGWOT)

Input: Marginal probabilities p^1 and p^2, intra-domain distance matrices D^1 and D^2, relaxation coefficient ρ, regularization coefficient ϵ

Initialize the coupling matrix: $\Gamma = \pi = p^1 \otimes p^2$

while Γ *not converged* **do**

$\Gamma \leftarrow \pi$

$\Gamma_{(mass)} \leftarrow \sum_{i,j} \Gamma_{i,j}$ $\quad \tilde{\epsilon} \leftarrow \Gamma_{(mass)}\epsilon, \quad \tilde{\rho} \leftarrow \Gamma_{(mass)}\rho$

// Compute cost C:

$\Gamma^1 \leftarrow \Gamma \mathbb{1}_{n_2}$, $\Gamma^2 \leftarrow \Gamma^T \mathbb{1}_{n_1}$

$A \leftarrow (D^1)^{\circ 2}\Gamma^1$, $B \leftarrow (D^2)^{\circ 2}\Gamma^2$

$D \leftarrow D^1 \Gamma D^2$

$E \leftarrow \epsilon \sum_{ij} \log\left(\frac{\Gamma_{i,j}}{p_i^1 p_j^2}\right)\Gamma_{i,j} + \rho\left(\sum_i \log\left(\frac{\Gamma_i^1}{p_i^1}\right)\Gamma_i^1 + \sum_j \log\left(\frac{\Gamma_j^2}{p_j^2}\right)\Gamma_j^2\right)$

$C \leftarrow A + B - 2D + E$

// Perform Sinkhorn iterations

while (u, v) *not converged* **do**

$u \leftarrow -\frac{\tilde{\epsilon}\tilde{\rho}}{\tilde{\epsilon}+\tilde{\rho}} \log\left[\sum_{i,j} \exp(v_j - C_{ij})/\tilde{\epsilon} + \log p^2\right]$

$v \leftarrow -\frac{\tilde{\epsilon}\tilde{\rho}}{\tilde{\epsilon}+\tilde{\rho}} \log\left[\sum_{i,j} \exp(u_i - C_{ij})/\tilde{\epsilon} + \log p^1\right]$

// Update: $\pi_{ij} \leftarrow \exp\left[u_i + v_j - C_{ij}\right] p_i^1 p_j^2$

// Rescale: $\pi \leftarrow \sqrt{\Gamma_{(mass)}/\pi_{(mass)}}\pi$

Return: Γ

x_i^m in the graph, we calculate the average of the k neighboring node values $\mathcal{N}_k(x_i^m)$. Next, we measure the difference between this average and the true value of the node. This difference reflects how well the averaged neighborhood represents the given node. We then average these differences across the graph and select the domain with the lowest averaged difference as the anchor. Intuitively, we select the anchor whose kNN graph best reflects its dataset. Suppose X^1 is the anchor dataset. Then, for $m = 2, 3, \ldots, N$, we compute the coupling matrix Γ^m according to Equation 6.2.4.

To have all of the datasets aligned in the same domain, we can either use barycentric projection to project each X^m for $m = 2, 3, \ldots, M$ onto X^1 or find a shared embedding space as described in Section 7.2.3. In the first iteration of SCOT, we used a barycentric projection to align and project one dataset onto the other. Due

to the marginal relaxation, we now search for a non-negative $n_1 \times n_m$ dimensional matrix Γ instead of $\Gamma \in \Pi(p^1, p^m)$. Because of this change, the adjusted barycentric projection is:

$$x_i^m \mapsto \frac{\sum_{j=1}^{n_1} \Gamma_{ij}^m x_j^1}{\sum_{j=1}^{n_1} \Gamma_{ij}^m}. \tag{7.2.3}$$

7.2.3 Embedding with the Coupling Matrix

Other methods such as MMD-MA and UnionCom align datasets by embedding them into a common latent space of dimension $p \leq \min_{m=1,\ldots,M} d_m$. Here d_m represents the original dimension size of measurement (or domain) m. Embedding the datasets in a new space often leads to a better alignment as it introduces the additional benefits of dimension reduction, allowing more meaningful structures in the datasets such as cell-types to be more prevalent. Due to these benefits, we also enable the embedding option through a modification of the t-SNE method proposed by UnionCom Cao et al. (2020). The full details of t-SNE can be found in Van der Maaten and Hinton (2008).

For each domain m, we compute P^m, an $n_m \times n_m$ cell-to-cell transition matrix; each entry $P_{j|i}^m$ is the conditional probability that a data point x_i^m would pick x_j^m as its neighbor when chosen according a Gaussian distribution centered at x_i^m:

$$P_{j|i}^m = \frac{exp(-||x_i^m - x_j^m||^2/2\sigma_i^2)}{\sum_{k \neq i} exp(-||x_i^m - x_k^m||^2/2\sigma_i^2)}. \tag{7.2.4}$$

The bandwidth σ_i is chosen according to the density of the data points through a binary search for the value of σ_i that achieves the user-supplied perplexity value.

P^m is computed by averaging $P^m_{i|j}$ and $P^m_{j|i}$ to give more weight to outlier points:

$$P^m_{ij} = \frac{P^m_{i|j} + P^m_{j|i}}{2n_m} \tag{7.2.5}$$

Then, to jointly embed all domains through the anchor domain X^1, the optimization problem is:

$$\min_{X^{1'},\ldots,X^{M'}} \sum_{m=1}^{M} \mathrm{KL}(P^m||Q^{m'}) + \beta \sum_{m=2}^{M} ||X^{1'} - X^{m'}(\Gamma^m)^T||^2_F, \tag{7.2.6}$$

where $X^{m'}$ is the lower dimensional embedding of X^m, P^m is defined as in Equation 7.2.4, and Γ^m is the coupling matrix from solving Equation 7.2.2 for $m = 1, 2, \ldots, M$, $X^{m'}$. The probability matrix Q^m is computed through a Student-t distribution with one degree of freedom:

$$Q^{m'}_{ij} = \frac{(1 + ||x^{m'}_i - x^{m'}_j||)^{-1}}{\sum_{k \neq l} 1 + (||x^{m'}_k - x^{m'}_l||)^{-1}}. \tag{7.2.7}$$

The intuition behind the term $\mathrm{KL}(P^m||Q^{m'})$ is very similar to that of GW; if two points have a high transition probability in the original space, then they should also have a high transition probability in the latent space. The term $||X^{1'} - X^{m'}(\Gamma^m)^T||^2_F$ measures how well aligned the new embeddings $X^{1'}$ and $X^{m'}$ are according to the prescribed coupling matrix Γ^m. Finally, $\beta > 0$ controls the trade-off between preserving the original geometry with the KL term and enforcing the alignment found with GW. We solve this optimization problem using gradient descent from UnionCom with a default latent space dimension size $p = 3$ Cao et al. (2020). The overall SCOTv2 method is presented as Algorithm 4.

Algorithm 4: Pseudocode for SCOTv2 Algorithm

Input: Datasets $X^1, \ldots, X^M$, number of neighbors in nearest neighbor graphs k, entropic regularization coefficient ϵ, mass conservation relaxation coefficient ρ.

for $m = 1, \ldots, M$ **do**

  // Initialize marginal probabilities: $p^m \leftarrow \text{Uniform}(X^m)$;

  //Construct G^m, a $k-$NN graph based on pairwise correlations

  // Compute intra-domain distance matrix D^m on G^m with Dijsktra's algorithm.

  // Compute a "neighborbood correlation" score, c^m:

  $c^m = \frac{1}{n_m} \sum_{i=1}^{n_m} \frac{1}{k} \sum_{x_j^m \in \mathcal{N}_k(x_i^m)} \text{corr}(x_j^m, x_i^m)$

end

// Select an anchor domain X^{m*}: $m^* = \arg\max_{m=1,\ldots M} c^m$

for $m = 1, \ldots, M$ $(m \neq m^*)$ **do**

  // Compute pairwise coupling matrices between the anchor domain X^{m*} and all other domains:

  $\Gamma^m \leftarrow GW_{\epsilon,\rho}(p^m, p^{m*})$

  if *Barycentric projection* **then**

    $x_i^{m'} \leftarrow \frac{\sum_{j=1}^{n_1} \Gamma_{ij}^m x_j^{m*}}{\sum_{j=1}^{n_1} \Gamma_{ij}^m}$

  end

  else

    // Find shared embedding

    $X^{1'} \ldots X^{M'} \leftarrow$

    $\min_{X^{m'}, \ldots, X^{M'}} \sum_{m=1}^{M} \text{KL}(P^m || Q^{m'}) + \beta \sum_{m \neq m*} ||X^{m*'} - X^{m'}(\Gamma^m)^T||_F^2$

  end

end

Return: Aligned datasets, $X^{1'} \ldots X^{M'}$.

7.2.4 Heuristic process for self-tuning hyperparameters

SCOTv2 has three hyperparameters: (1) k for the number of neighbors to consider in nearest neighbor graphs, (2) the weight of the entropic regularization term, ϵ, and (3) the coefficient of the mass relaxation constraint, ρ. The barycentric projection of one domain onto another does not require any hyperparameters. However, jointly embedding the domains in a latent space requires selecting the dimension p.

Ideally, orthogonal correspondence information such as 1–1 correspondences and

cell-type labels can guide hyperparameter tuning as validation. However, such information is hard to obtain in most cases. First, no validation data on cell-to-cell correspondences exists for non-co-assay datasets. Second, it is challenging to infer cell-types for certain sequencing domains such as 3D chromatin conformation. Lastly, the cell-type annotations may not always agree across single-cell domains.

We provide a heuristic to self-tune hyperparameters in the completely unsupervised setting. We first choose a k for the neighborhood graphs that yields a high average correlation value between the neighborhood predicted values and measured genomic values of the graph nodes. This step is the same as the one used to select the anchor domain for multi-omics alignment in Section 7.2.2. Next, we choose ϵ and ρ values that minimize the Gromov-Wasserstein distance between the aligned datasets. Algorithm 5 gives the details of this procedure.

Algorithm 5: Unsupervised hyperparameter search procedure

Input: Datasets $X^1, \ldots, X^M$.
// Find k for each domain
for $m = 1, \ldots, M$ **do**
  $k^m = \underset{k \in \{10,20,\ldots,150\}}{\text{argmax}} \frac{1}{n_m} \sum_{i=1}^{n_m} \frac{1}{k} \sum_{x_j^m \in \mathcal{N}_k(x_i^m)} \text{corr}(x_j^m, x_i^m)$
  // Use k^m to compute D^m
end
// Use the GW distance to pick ρ and ϵ
for $m = 2, \ldots, M$ **do**
  $\epsilon^m, \rho^m = \arg\min_{\epsilon,\rho} GW_{\epsilon,\rho}(\mathbb{1}_{n_1}, \mathbb{1}_{n_m})$
end
Return: k^m, ϵ^m, ρ^m.

7.3 Experimental Setup

7.3.1 Datasets

We evaluate SCOTv2 on single-cell datasets with disproportionate cell-types using two schemes. (1) We subsample different cell-types in co-assay datasets to simulate cell-type representation disparities between sequencing modalities. (2) We select real-world separately sequenced single-cell multi-omics datasets, which lack 1–1 cell correspondences and have different cell-type proportions across modalities due to the sampling procedure. Additionally, we present results on the original co-assay datasets with 1–1 cell correspondence to demonstrate the flexibility of SCOTv2 across balanced and unbalanced single-cell datasets.

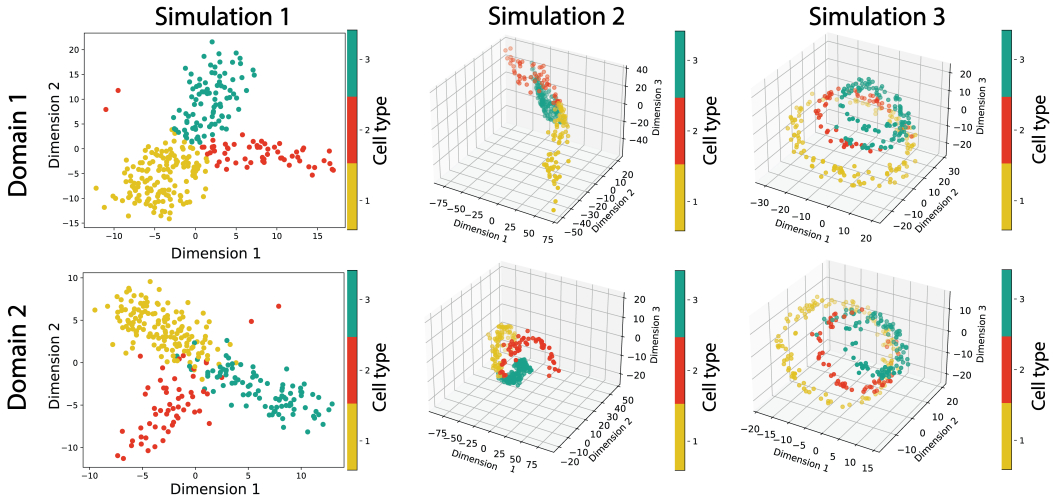

Figure 7.2: **Visualization of original simulation data before alignment** by MDS projections, as displayed in Liu et al. (2019). Data was generated by Liu *et al* Liu et al. (2019) and retrieved from https://noble.gs.washington.edu/proj/mmd-ma/.

7.3.1.1 Co-assay single-cell datasets with 1–1 cell correspondence

We use three co-assay datasets to validate our model, sequenced by SNARE-seq, scGEM, and scNMT technologies. SNARE-seq is a two-modality sequencing technology that simultaneously captures the chromatin accessibility and transcriptional profiles of cells Chen et al. (2019). This dataset contains a total of 1047 cells from four cell lines: BJ (human fibroblast cells), H1 (human embryonic cells), K562 (human erythroleukemia cells), and GM12878 (human lymphoblastoid cells) (Gene Expression Omnibus access code: GSE126074). We follow the same data preprocessing steps outlined by Chen *et al.* Chen et al. (2019). The scGEM technology is a three-modality sequencing technology that profiles the genetic sequence, gene expression, and DNA methylation states in the same cell Cheow et al. (2016). The dataset we use is derived from human somatic cell samples undergoing conversion to induced pluripotent stem cells (Sequence Read Archive accession code SRP077853) Cheow et al. (2016). We access the preprocessed data provided by Welch *et al.* Welch et al. (2017), which only contains the gene expression and DNA methylation modalities [1]. The dataset sequenced by scNMT-seq method Argelaguet et al. (2019) contains three modalities of genomic data: gene expression, DNA methylation, and chromatin accessibility, from mouse gastrulation samples, going through the Carnegie stages of vertebrate development (Gene Expression Omnibus access code: GSE109262). We access the preprocessed data through the scripts[2] provided by the authors. While the SNARE-seq and scGEM datasets contain the same number of cells across measurements, scNMT-seq modalities contain different cell-type proportions after preprocessing due to varying noise levels in measurements. Supplementary Table 7.1 lists the number of cells belonging to different cell-types in each domain for scNMT-seq

[1]Preprocessed data for the scGEM dataset accessed here: https://github.com/jw156605/MATCHER

[2]Preprocessing scripts for the scNMT-seq data accessed here: https://github.com/PMBio/scNMT-seq/

dataset.

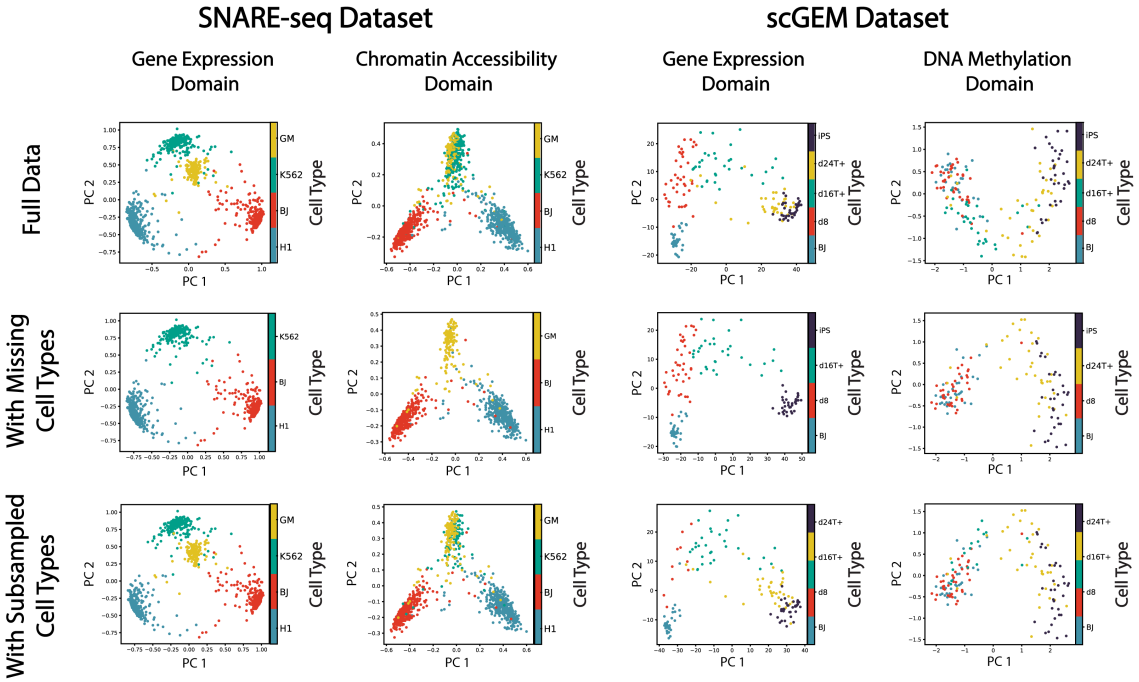

Figure 7.3: SNAREseq and scGEM datasets prior to alignment for all three simulation scenarios.

7.3.1.2 Single-cell datasets with simulated cell-type imbalance.

To test alignment performance sensitivity to different levels and types of cell-type proportion disparities across modalities, we generate simulation datasets by subsampling SNARE-seq and scGEM co-sequencing datasets in two ways. (1) We remove a cell-type from one modality. (2) We reduce the proportion of a cell-type in one modality by subsampling it at 50% and another cell-type in the other modality by subsampling it at 75%. We simulate this setting to test how the alignment methods will behave when multiple cell-types have disproportionate representation at different levels (for example, half or quarter percentage of cell-types missing) across modalities.

For these cases, we uniformly pick at random which cell-type to subsample or re-

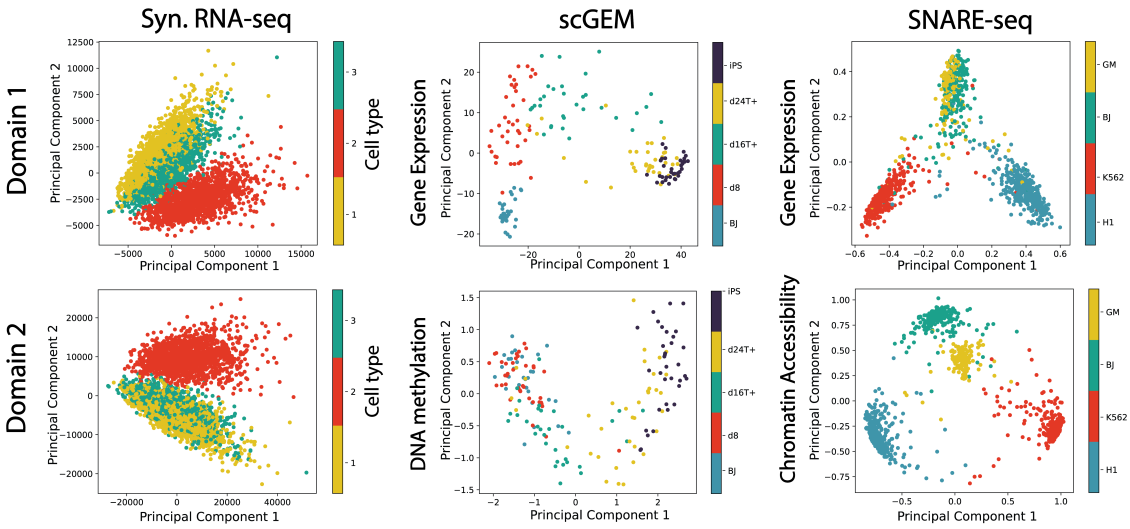

Figure 7.4: **Original synthetic RNA-seq and real world single-cell data visualized before alignment** by PCA projections. The synthetic RNA dataset is genrated using Splatter Zappia et al. (2017), and scGEM and SNARE-eq come from Cheow et al. (2016) Chen et al. (2019), respectively.

move. Specifically, for scGEM in simulation case (1), we remove "d16T+" cells in the DNA methylation domain while retaining the original gene expression domain, and remove the "d24T+" cells in the gene expression domain while retaining the original DNA methylation domain. For the SNARE-seq dataset, we remove "GM" cells in the gene expression domain and "K562" in the chromatin accessibility domain. In simulation case (2), we subsample the "d8" cluster of the scGEM dataset at 75% in the gene expression modality and the "d16T+" cluster at 50% in the DNA methylation modality. For SNARE-seq, we subsample the "H1" cluster at 75% and the "K562" cluster at 50% in the gene expression and chromatin accessibility domains, respectively.

7.3.1.3 Single-cell datasets without 1—1 correspondences

We also align non-co-assay datasets, containing separately sequenced single-cell -omic measurements. Bonora *et al.* generated the first dataset we use, "sciOmics" Bonora

et al. (2021). This dataset consists of sciRNA-seq, sciATAC-seq, and sciHiC measurements, capturing gene expression, chromatin accessibility, and 3D chromosomal conformation profiles of mouse embryonic stem cells undergoing differentiation. The measurements were taken at five stages: days 0, 3, 7, 11, and as fully differentiated neural progenitor cells (NPCs). The second non-co-assay dataset, "MEC," contains gene expression and chromatin accessibility measurements taken using the 10X Chromium scRNA-seq and scATAC-seq technologies on mouse mammary epithelial cells (MEC). Since each modality consists of separately sampled cell populations, these contain disparate cell-type proportions across modalities. Table 7.1 lists the number of cells belonging to different cell-types in each domain for sciOmics and MEC datasets.

	Modality 1 (Gene Expression)	**Modality 2** (Chromatin Accessibility)	**Modality 3** (DNA Methyl. or Chrom. Conform.)
scNMT	$n = 579$	$n = 647$	$n = 725$
E4.5	76 (12.73%)	63 (9.73%)	65 (8.96%)
E5.5	104 (17.42%)	89 (13.76%)	91 (12.55%)
E6.5	146 (24.46%)	220 (34%)	278 (38.34%)
E7.5	271 (45.39%)	175 (42.5%)	291 (40.14%)
sciOmics	$n = 1058$	$n = 1296$	$n = 2154$
Day0	489 (46.22%)	164 (12.65%)	987 (45.82%)
Day3	127 (12%)	702 (54.17%)	435 (20.19%)
Day7	78 (7.37%)	77 (5.94%)	243 (11.28%)
Day11	145 (13.71%)	175 (13.5%)	164 (7.61%)
NPC	219 (20.7%)	178 (13.73%)	325 (15.09%)
MEC	$n = 26273$	$n = 21262$	
Basal	11138 (42.39%)	13353 (62.8%)	
L-Sec (Prog)	7683 (29.24%)	3343 (15.72%)	
L-HR	3439 (13.09%)	2624 (12.34%)	NA
L-Sec (Mat)	2869 (10.92%)	1165 (5.48%)	
L-Sec (Prolif)	758 (2.89%)	7 (0.033%)	
Stroma	386 (1.47%)	770 (3.62%)	

Table 7.1: Number of cells per (and percentages of) cell-type across different modalities in the scNMT-seq, sciOmics, and MEC datasets after quality control procedures and the non-coassay datasets.

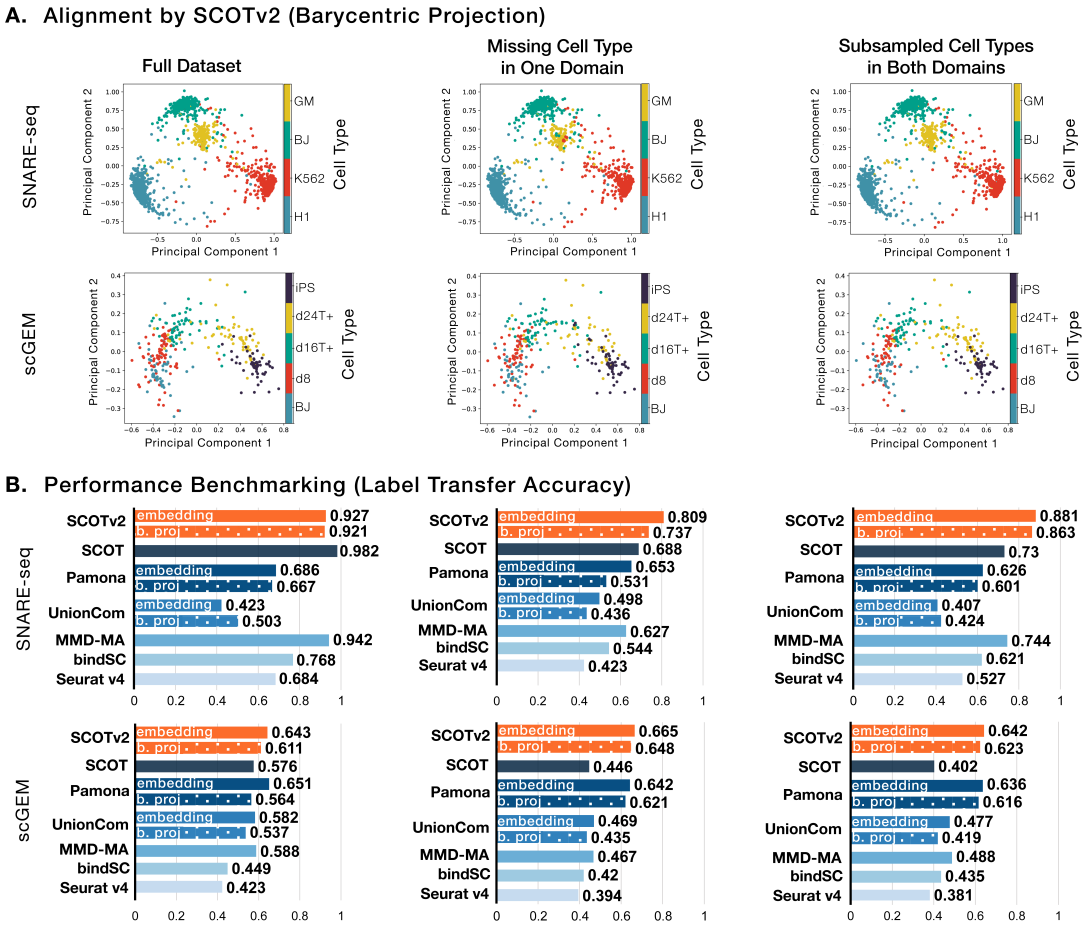

Figure 7.5: **Alignment results for simulations and balanced co-assay datasets.** **A** visualizes the barycentric projection alignment on SNARE-seq and scGEM for the full co-assay datasets, simulations with a missing cell-type in the epigenomic domain, and subsampled cell-types in both domains. **B** compares the alignment performance of SCOTv2 to the benchmarks through LTA. For SCOTvs, Pamona, and UnionCom, we report results on both embedding into a shared space (solid bars) and the barycentric projection (dotted bars).

7.3.2 Evaluation metrics and baseline methods

Although most of the datasets lack 1–1 cell correspondences, we can evaluate alignment using cell-type labels through label transfer accuracy (LTA) as in Cao et al. (2021, 2020), Demetci et al. (2021). This metric assesses the clustering of cell-types after alignment by training a kNN classifier on a training set (50% of the aligned data) and then evaluates its predictive accuracy on a test dataset (the other 50%

of the aligned data). Higher values correspond to better alignments, indicating that cells that belong to the same cell-type are aligned close together after integration. We benchmark our method against the current unsupervised single-cell multi-omic alignment methods, Pamona Cao et al. (2021), UnionCom Cao et al. (2020), MMD-MA Singh et al. (2020), bindSC Dou et al. (2020), Seuratv4 Stuart et al. (2019), and the previous version of SCOT, which performs alignment without the KL term Demetci et al. (2021). Pamona Cao et al. (2021), as previously discussed, uses partial Gromov-Wasserstein (GW) optimal transport to align single-cell datasets. UnionCom Cao et al. (2020) performs unsupervised topological alignment through a two-step procedure that first finds a correspondence between the domains, considering both global and local geometries with a hyperparameter to control the trade-off between them, and then embeds them in a new shared space. MMD-MA Singh et al. (2020) uses the maximum mean discrepancy (MMD) measure to align and embed two datasets in a new space. BindSC Dou et al. (2020) requires the users to bring input datasets to the gene expression feature space by constructing a gene activity score matrix for the epigenomic domains, then finds a correspondence matrix between samples through bi-order canonical correspondence analysis (bi-CCA), and jointly embeds the domains into a new space. Finally, Seuratv4 Stuart et al. (2019) also requires gene activity score matrices for epigenomic domains and then identifies correspondence anchors via CCA. Based on these anchors, it imputes one genomic domain based from the other domain and co-embeds them into a shared space using UMAP.

Since bindSC and Seurat v4 require the creation of gene activity score matrices for epigenomic datasets, they might be more difficult to use with certain sequencing domains. For instance, gene activity scoring is challenging for 3D chromosomal conformation. Of all the selected baselines, only Pamona and UnionCom can align

more than two domains, so we only use them as baselines for experiments with multiple domains ($M > 2$).For each benchmark, we define a hyperparameter grid of similar granularity and perform extensive tuning (see Figure 7.6). We report the alignment results with the best performing hyperparameter combinations in Section 7.4.1.

7.3.3 Hyperparameter Tuning Procedure Details

For each alignment method, we define a grid of hyperparameters and choose the best performing combination for each experiment. If methods share similar hyperparameters in their formulation, we keep the range defined for these consistent across all algorithms. Examples for such hyperparameters are dimensionality of the latent space, p, for the algorithms that commonly embed datasets; entropic regularization constant, ϵ, for methods that employ optimal transport; number of neighbors, k, for methods that model single-cell datasets with nearest neighbor graphs. Otherwise, we refer to the publication and the code repository for each method to choose a hyperparameter range.

For Pamona, we tune four hyperparameters: $k \in \{20, 30, \ldots, 150\}$, the number of neighbors in the cell neighborhood graphs, $\epsilon \in \{5e-4, 3e-4, 1e-4, 7e-3, 5e-3, \ldots, 1e-2\}$, the entropic regularization coefficient for the optimal transport formulation, $\lambda \in \{0.1, 0.5, 1, 5, 10\}$, the coefficient for the trade-off between aligning corresponding cells and preserving local geometries, and lastly, $p \in \{3, 4, 5, 10, 30, 32\}$, the output dimension for embedding . We choose the ranges for ϵ and k to be consistent with the corresponding hyperparameters in SCOT and SCOTv2 algorithms and the ranges for the embedding dimensions to be consistent with the recommended values in MMD-MA and UnionCom embeddings.

For UnionCom, we tune the trade-off parameter $\beta \in \{0.1, 1, 5, 10, 15, 20\}$ and the regularization coefficient $\rho \in \{0, 0.1, 1, 5, 10, 15, 20\}$ based on the ranges reported by Cao *et al.* in the publication Cao et al. (2020). We additionally tune the maximum neighborhood size permitted in the neighborhood graphs, $k_{max} \in \{40, 100, 150\}$, as well as the embedding dimensionality $p \in \{3, 4, 5, 10, 30, 32\}$. The sweep range for hyperparameter k_{max} is smaller than the other hyperparameters because UnionCom automatically starts from $k = 2$ and goes up to k_{max} to find the lowest k that returns a connected graph to use in the algorithm. Therefore, more refined search is not needed.

For MMD-MA, we choose the weights λ_1 and $\lambda_2 \in \{1e-2, 5e-3, 1e-3, 5e-4, \ldots, 1e-9\}$. This range includes the hyperparameter range suggested by Singh *et al* ($\lambda_1,\ \lambda_2 \in \{1e-3, 1e-4, 1e-5, 1e-6, 1e-7\}$) but extends it further to increase the granularity for the sake of more fair comparison against methods that require a higher number of hyperparameters to test, such as Pamona and UnionCom. Similarly to other methods, we also select the embedding dimensionality from $p \in \{3, 4, 5, 10, 30, 32\}$.

For bindSC, we choose the couple coefficient that assigns weight to the initial gene activity matrix $\alpha \in \{0, 0.1, 0.2, \ldots 0.9\}$ and the couple coefficient that assigns weight factor to multi-objective function $\lambda \in \{0.1, 0.2, \ldots, 0.9\}$. Additionally, we choose the number of canonical vectors for the embdedding space $K \in \{3, 4, 5, 10, 30, 32\}$.

Lastly, for Seurat v4, we tune the number of neighbors to consider when finding anchors, $k \in \{5, 10, 15, 20\}$, dimensions of the final co-embedding space, $p \in \{3, 4, 5, 10, 30, 32\}$ and the choice of the reference and anchor domains when finding anchors.

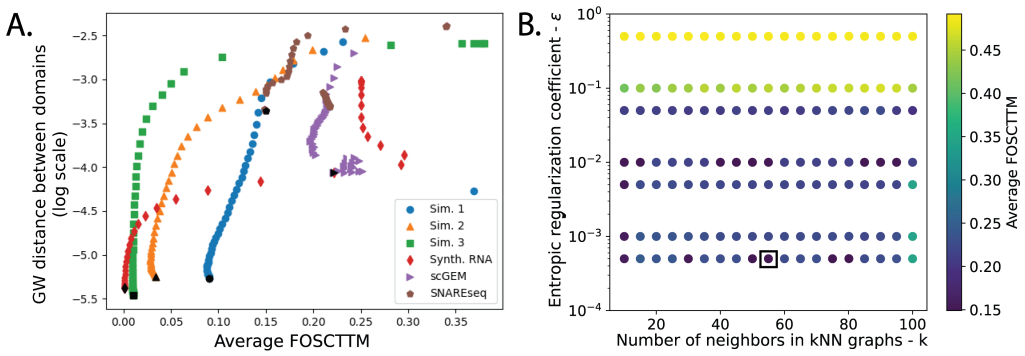

Figure 7.6: **A.Gromov-Wasserstein distance vs alignment quality** for $\epsilon \in [1e-4, 1e-1]$ and $k = \min(50, 0.2n_x)$. Alignment quality is shown by average FOSCTTM values on the x-axis. On the y-axis, we plot Gromov-Wasserstein distance between the aligned domains (log scale). **B. Average FOSCTTM (indicated by color) with different hyperparameter combinations** for SNAREseq data (a representative example). Alignment is mostly robust to the choice of k. Black box indicates the best performing hyperparameter setting($k = 50$, $\epsilon = 5e-4$ for mean FOSCTTM of 0.149).

7.4 Results

7.4.1 SCOTv2 consistently yields high-quality alignments

We first present the alignment results for real-world co-assay datasets with simulated cell-type imbalance. Figure 7.5 (A) visualizes the barycentric projection alignments performed by SCOTv2 plotted as 2D PCA for SNARE-seq and scGEM datasets, respectively. We use barycentric projection for visualization purposes for the ease of comparison with the original domains, plotted in Supplementary Figure 7.3. Here, we integrate datasets under three different settings described in the previous section: (1) Balanced datasets (or "full datasets" with no subsampling), (2) Missing cell-type in the epigenomic domains, and (3) Subsampled cells in both domains (one cell-type at 50% in the epigenomic domains and another cell-type at 75% in the gene expression domains). We include alignment results on the full datasets with 1–1 sample correspondences to ensure that SCOTv2 performs well for balanced cases as

well.

Qualitatively, we see that SCOTv2 preserves the cell-type annotations after alignment for all three settings. In Figure 7.5 (B), we report the quantitative performance of SCOTv2 and all the other state-of-the-art baselines using the Label Transfer Accuracy (LTA) scores. MMD-MA, UnionCom, Seurat, and bindSC fail to reliably align datasets with disproportionate cell-type representation across modalities. While Pamona tends to yield high-quality alignments for cases with cell-type disproportion, it fails to perform well on the SNARE-seq balanced dataset as well as its subsampling simulation. We additionally apply Pamona to randomly downsampled co-assays (Figure 7.8). We show that while Pamona's partial optimal transport framework handles cell-type disproportion better than the balanced optimal transport formulation (demonstrated by SCOT), SCOTv2 still shows an advantage in all SNARE-seq simulations ($\sim 20\%$ increase in LTA), as well as the smaller downsampling schemes ($\sim 10\%$).

Among all methods tested, SCOTv2 consistently gives more high-quality alignments across different scenarios of cell-type representation. It also demonstrates a $\sim 22\%$ average increase in LTA over the previous version of the algorithm (SCOT) when comparing the barycentric projection results and $\sim 27\%$ for the embedding results. Supplementary Figure 7.8 presents similar results (SCOTv2 attains an LTA of 0.786 followed by Pamona at 0.62 on SNAREseq and 0.542 followed by Pamona at 0.538 on scGEM) for missing cell-types in the other (gene expression) domain, suggesting that our choice of domain with missing cell-type does not affect the performance comparison results. UnionCom, Pamona, and SCOTv2 allow us to perform both barycentric projections and embed the single-cell domains in a lower-dimensional space. Overall, we observe that embedding yields higher LTA values than barycentric projection. Since the barycentric projection projects one domain

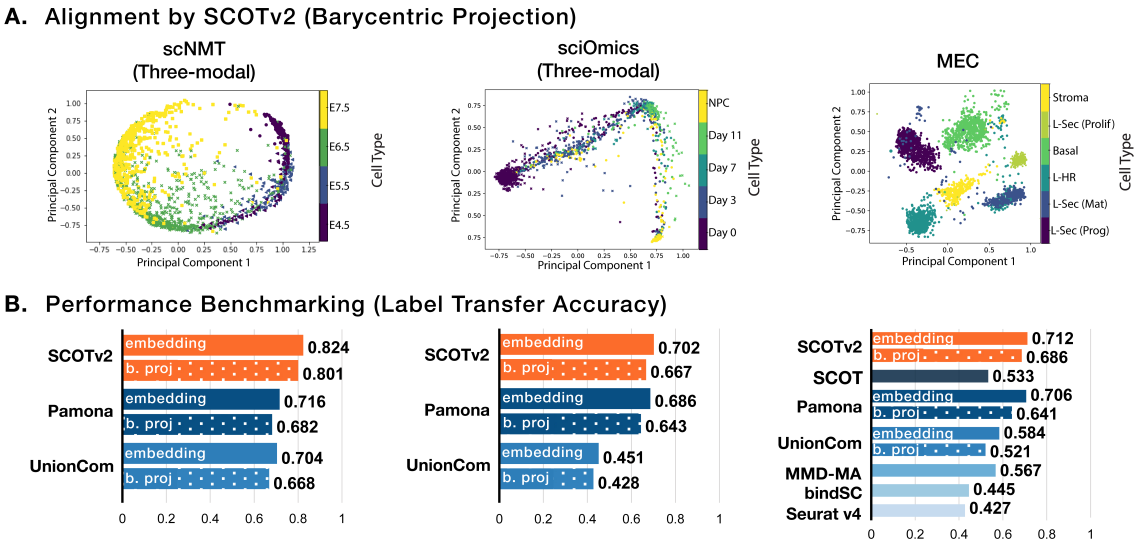

Figure 7.7: **Alignment results for multi-modal ($M > 2$) and separately sequenced datasets. A** visualizes the alignment of scNMT-seq, sciOmics, and MEC. All datasets have unequal sample sizes and cell-type proportions across domains. **B** benchmarks alignment performance through LTA. As in Figure 7.5, we report results both by embedding (solid bars) and barycentric projection (dotted bars) for the methods that allow for both. For scNMT-seq and sciOmics, which are three-modal datasets, we only demonstrate results for SCOTv2, Pamona, and UnionCom, which can handle more than two modalities.

onto another, the separation of the domain being projected onto (or anchor domain) limits the clustering separation after alignment. In contrast, the embedding utilizes t-SNE to enhance cell-type separation, allowing for better-separated clusters after alignment.

Next, we report the alignment performance of SCOTv2 on single-cell datasets with disparities in cell-type representation due to sampling during experiments. We include scNMT, a co-assay with varying levels of cells across domains due to quality control procedures, along with sciOmics and MEC for this experiment. Note that scNMT and sciOmics have three different modalities, and hence, we can only report the baselines for methods that can align datasets with $M > 2$. Figure 7.7(A) presents the qualitative alignment results for SCOTv2 with PCA. SCOTv2 performs well on all three datasets, including the ones with three modalities. The LTA scores in

	SNARE (full dataset)	**SNARE** (missing cell-type)	**SNARE** (subsampled)	**scGEM** (full dataset)	**scGEM** (missing cell-type)
SCOTv2	0.826	**0.653**	**0.751**	**0.509**	**0.521**
SCOT	**0.852**	0.572	0.588	0.423	0.323
Pamona	0.554	0.423	0.419	0.385	0.414
MMD-MA	0.523	0.407	0.431	0.360	0.296
UnionCom	0.411	0.406	0.422	0.332	0.315
bindSC	0.713	0.584	0.475	0.387	0.254
Seurat	0.428	0.517	0.503	0.408	0.377
	scGEM (subsampled)	**scNMT**	**sciOmics**	**MEC**	
SCOTv2	**0.415**	**0.727**	**0.537**	**0.584**	
SCOT	0.314	N/A	N/A	0.466	
Pamona	0.308	0.588	0.329	0.417	
MMD-MA	0.287	N/A	N/A	0.233	
UnionCom	0.276	0.474	0.306	0.349	
bindSC	0.262	N/A	N/A	0.412	
Seurat	0.329	N/A	N/A	0.387	

Table 7.2: **Alignment performance benchmarking in the fully unsupervised setting.** We run SCOTv2 and SCOT using their heuristics to approximately self-tune hyperparameters. We use default parameters for other methods due to a lack of similar procedures for unsupervised self-tuning.

Figure 7.7(B) demonstrate that SCOTv2 consistently yields the best alignments on the three real-world datasets. These results highlight its ability to reliably integrate separately sampled with disproportionate cell-type representation and multiple ($M > 2$) modalities simultaneously.

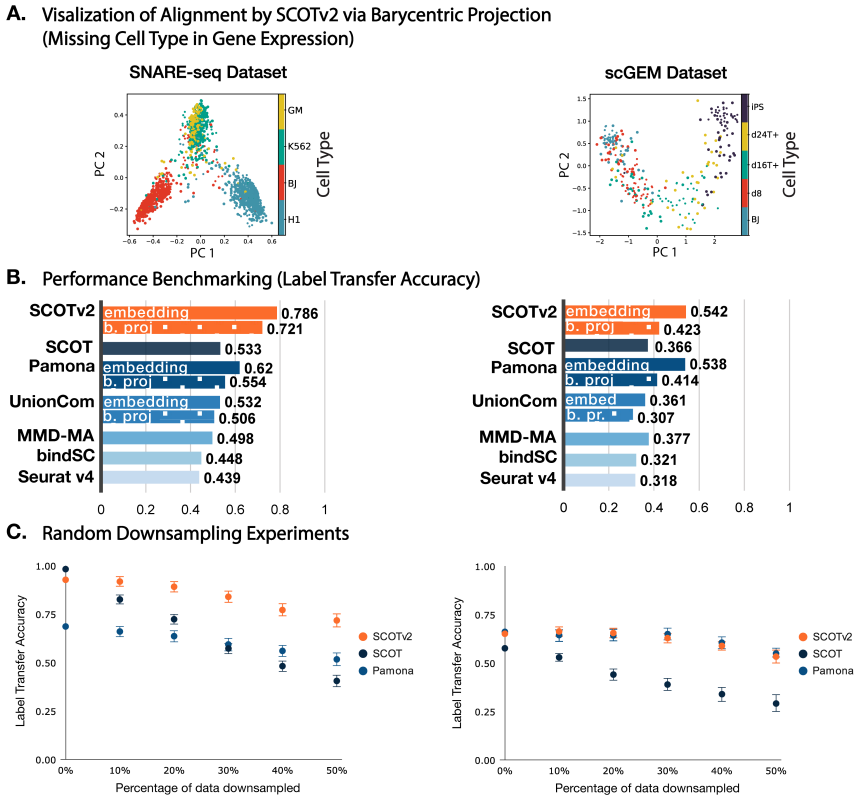

Figure 7.8: **Alignment results on simulations with co-assay datasets. Panel A** visualizes the alignment results by SCOTv2, using barycentric projection, on co-assay datasets SNARE-seq and scGEM when a cell-type is missing in the gene expression domain. **Panel B** quantifies the alignment quality in this experiment with label transfer accuracy and compares to baseline methods. **Panel C** plots the average label transfer accuracy results obtained from SCOTv2, SCOT, and Pamona algorithms when aligning randomly downsampled datasets. These experiments are repeated five times and the standard deviation is shown with error bars.

7.4.2 Hyperparameter self-tuning aligns well

The benchmarking results above present the alignment performance of each algorithm at its best hyperparameter setting; however, users may not have 1—1 correspondences to validate alignments, for the purpose of hyperparameter selection, in real-world applications. While users may have access to cell-type labels, inferring cell-types is highly difficult in specific modalities of single-cell sequencing, such as 3D chromatin conformation. Additionally, different sequencing modalities might dis-

agree on cell-type clustering (as is often the case with scRNA-seq and scATAC-seq datasets). In these situations, users might not have sufficient validation data for tuning hyperparameters.

We design a heuristic process (described in Section 7.2.4), as done previously for SCOT, that allows SCOTv2 to select hyperparameters in a completely unsupervised manner. Other alignment methods do not provide an unsupervised hyperparameter tuning procedure. Therefore, without validation data, a user would have to use the default parameters. In Table 7.2, we compare alignment performance for our heuristic against the default parameters of other methods. While our heuristic does not always yield the optimal hyperparameter combination, it does give more favorable results over the default settings of the other methods. Thus, we recommend using it in cases that lack orthogonal information for hyperparameter tuning.

7.4.3 SCOTv2 scales well with increasing number of samples

We compare the runtime of SCOTv2 with the top performing methods: Pamona, MMD-MA, UnionCom, and the previous version of SCOT by subsampling various numbers of cells from the MEC dataset. MMD-MA, UnionCom, and SCOTv2 have GPU versions, while Pamona and SCOT only have CPU versions. We run MMD-MA and UnionCom on a single NVIDIA GTX 1080ti GPU with VRAM of 11GB and Pamona and SCOT on Intel Xeon e5-2670 CPU with 16GB memory. We also run SCOTv2 on the same CPU to give comparable results to Pamona's runtimes. Figure 7.9 depicts that SCOT, MMD-MA, Pamona, and SCOTv2 show similar computational scaling.

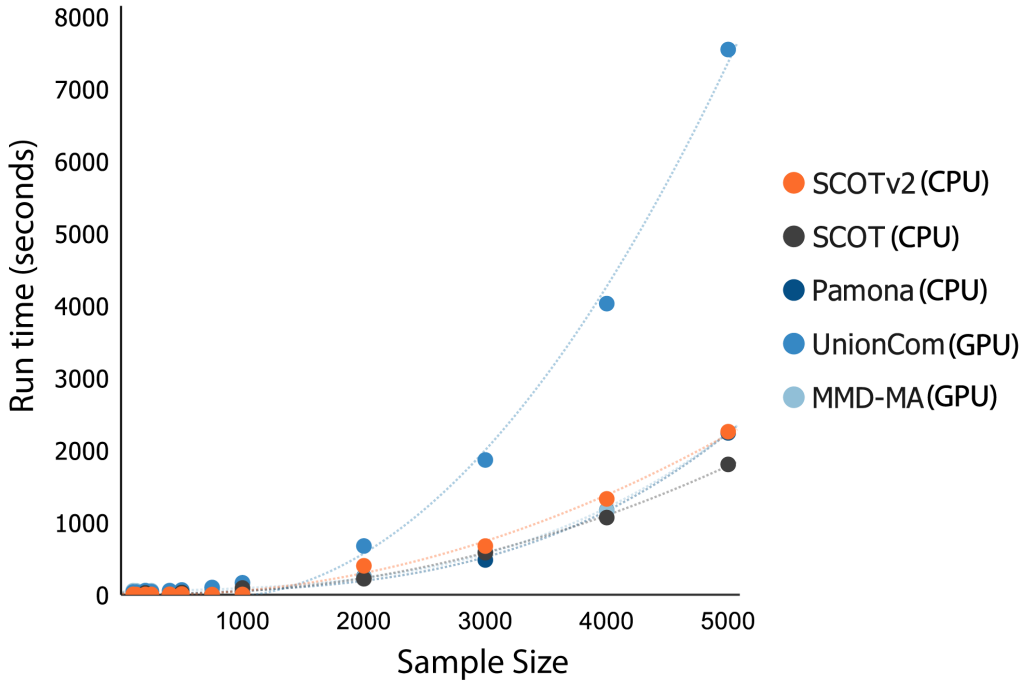

Figure 7.9: Runtimes for SCOTv2, SCOT, Pamona, UnionCom, and MMD-MA as the number of samples increases.

7.5 Discussion

We present SCOTv2, an improved unsupervised alignment algorithm for multi-omics single-cell alignment. It extends the alignment capabilities of SCOT to datasets with cell-type representation disproportions across different sequencing measurements. It also performs alignment for single-cell datasets with more than two measurements ($M > 2$). Experiments on real-world subsampled co-assay datasets and separately sampled and sequenced single-cell datasets demonstrate that SCOTv2 reliably yields high-quality alignments for a wide range of cell-type disproportions without compromising its computational scalability. Furthermore, SCOTv2's flexible marginal constraints enable it to consistently give good alignments results for both balanced and unbalanced single-cell datasets. In addition to effectively handling cell-type imbalances and multi-omics alignment, SCOTv2 can self-tune its hyperparameters making it applicable in complete unsupervised settings. Therefore, SCOTv2 offers

a convenient way to align multiple single-cell measurements without requiring any orthogonal correspondence information.

In this second iteration of SCOT, we have utilized the coupling matrix in a new way to find a latent embedding space. While this dimension reduction improves cell-type separation, using the coupling matrix directly may offer even more insights into interactions between the aligned domains. Future work will consider how to use the probabilities in the coupling matrix directly for downstream analysis like improved clustering and pseudo-time inference. Though SCOTv2 has runtimes that scale with other methods, it requires $O(n^2)$ memory storage for the distance matrices, which may be an issue for especially large datasets. One way to address this limitation would be to develop a procedure to align a representative subset of each domain that can be extended to the entire dataset. Therefore, we will explore this direction to further improve the scalability of SCOTv2.

Chapter Eight

Conclusion

8.1 Summary of Contributions

In this thesis, we have presented various data-driven projects with applications from dynamical systems to computational biology. Below, we briefly summarize these contributions.

8.1.1 Enabling Equation-Free Modeling via Diffusion Maps

We have presented a new framework for conducting unsupervised equation-free modeling on high-dimensional dynamical systems with a slow-fast time-scale separation. Equation-free modeling reduces the dynamics of the high-dimensional, or microsystem, to a low-dimensional representation called the macrosystem. Lifting and restriction operators allow the user to conduct numerical analysis in the macrosystem. However, finding the macrosystem representation and defining lifting and restriction operators is a challenging task. Here, we introduce an application-independent way to use diffusion maps to find the macrosystem and define the operators.

This data-driven approach applies diffusion maps to simulations of the microsystem to define a low-dimensional representation. Additionally, we describe an application-independent restriction operator based on the Nystrom method and a lifting operator defined through an optimization problem. We apply these tools to an optimal velocity traffic system on a ring road. We can compute traffic jam patterns as both steady states and periodic orbits through the macrosystem.

8.1.2 JUSTFAIR

We have presented JUSTFAIR, the first large-scale, public database linking federal criminal sentencing information with both the defendant demographics and sentencing judge identity. JUSTFAIR contains nearly 600,000 records, merged using data from the United States Sentencing Commission, the Federal Judicial Center, the Public Access to Court Electronic Records system, and Wikipedia. We manually validated that the information in the dataset is correct with an estimated 99% accuracy. While the original dataset contains cases from 2002-2018, we have also repeated this merging process for 2019 data.

We then used the JUSTFAIR database with a hierarchical linear model to analyze racial disparities in sentencing among judges. Our model uses log-transformed sentence length with cases nested within federal judges and districts and controls for many factors, including recommended sentence, crime type, and presence of a plea agreement. We find that judges sentence Black and Hispanic defendants for considerably longer than white defendants. Specifically, Black and Hispanic defendants on average receive 13% and 19% longer sentences than white defendants, respectfully.

Though we are confident that our results demonstrate considerable sentencing disparity, we present these results with some caution. First, the JUSTFAIR dataset is a sample of all federal sentencing rather than the population. Though we have no reason to believe that cases are systematically missing, this sample may not represent the entire population well. In this case, our results would be imperfect approximations of the actual disparity. Second, our analysis could not prove that cases were randomly assigned to judges in the majority of districts. Thus, some judges may see more of a particular type of case that we cannot control for. As a result, they could appear to be more or less biased than they truly are. Thus, while

we advise caution while interpreting our results, especially for individual judges.

8.1.3 Auditing with Optimal Transport

As automated decision-making systems become more common, it is even more imperative that we develop thorough audits to interrogate them. Singular notions of fairness that may not apply in every situation limit current methods. We present a flexible auditing framework through optimal transport. In particular, our approach compares outcomes distributions to detect disparities between different populations. By allowing the user to define a reference population, our auditing technique can adapt to various scenarios and definitions of fairness. Furthermore, we use the coupling matrix with measures of distance between features to investigate how people are treated.

We have applied this framework to a simulated dataset and two real datasets. We show that the Wasserstein distance can capture bias when other fairness metrics fail through a simulation of school admissions. Next, we offer that the COMPAS dataset unfairly treats Black people as significantly more likely to recidivate than similar white people. Specifically, the algorithm treats Black people with no priors like white people with priors. Finally, we audit the German credit dataset to use the coupling matrix to explore the feature space associated with outcomes. Through all of these examples, we have demonstrated the broad applicability of our optimal transport framework.

8.1.4 SCOT

Integration of single-cell data offers critical insights into interactions between different molecular views of the cell. Though important, multi-omics alignment is a challenging task due to a lack of correspondence information between modalities. We have presented Single Cell alignment using Optimal Transport (SCOT), a Gromov-Wasserstein-based method for integrating multi-omics datasets. In our first iteration of SCOT, we showed that we could align both simulated and real-world datasets and state-of-the-art methods with fewer hyperparameters. Additionally, SCOT computationally scales well as sample size increases. Finally, we introduce an unsupervised hyperparameter tuning method that uses Gromov-Wasserstein distance as a proxy for alignment quality.

While SCOT performed on par with other alignment methods and was easier to use, it shared a common drawback with other techniques: we benchmarked it on balanced datasets with 1–1 correspondences. Real-world datasets often contain cell-type imbalances across modalities, so we introduced SCOTv2 to handle this case. SCOTv2 uses unbalanced Gromov-Wasserstein, which relaxes the conservation of mass constraint in standard optimal transport. Additionally, SCOTv2 can manage more than two domains and embed the aligned datasets into a new space. We demonstrate that SCOTv2 retains its alignment quality on balanced datasets while outperforming other methods on unbalanced datasets.

8.2 Future Directions

In this section, we discuss possible future directions for all of these projects.

8.2.1 Enabling Equation-Free Modeling via Diffusion Maps

We can either extend our equation-free modeling methodology or apply it in new areas. The trickiest part of using this diffusion-map-based approach is sampling the microsystem representatively. Ensuring that we sample points from the entire region of interest is challenging without extensive prior system knowledge. Since all of the macrosystem analysis relies on accurately representing the microsystem, this process is critical. Developing methodologies to ensure coverage of larger areas of phase space would drastically improve using this framework.

While we presented a novel way to conduct bifurcation analysis of periodic orbits using diffusion maps and equation-free modeling, it would be interesting to see if this approach could be applied to more complex patterns. In this case, the translational symmetry of the traffic jams was easy for us to pick up on, but diffusion maps could better parametrize other spatially chaotic structures. There are many other high-dimensional systems where we cannot characterize a macrovariable directly where it would be interesting to apply this framework. For example, diffusion maps could detect patches of turbulent fluid surrounded by laminar flow.

8.2.2 JUSTFAIR

There are two main future directions for work with the JUSTFAIR dataset: improving the data itself and conducting further analysis. First, the dataset is not representative of all federal criminal cases in the United States. Of the original 1,265,688 records in the USSC dataset, we managed to merge 595,850 (or 47.1%) cases with their corresponding judge information. Finding better data sources to integrate with and improving the merging pipeline could improve this match rate

dramatically. While we added 2019 data to the original dataset covering 2002-2018, there will always be more recent data to add to the JUSTFAIR dataset. Thus, one additional point of future work includes automating the merging process to proceed annually.

The quality of data included in the JUSTFAIR dataset severely limited our analyses. Our original goal included naming the most disparate federal judges, but we did not feel confident enough to share this information due to the missing cases. While improving the data quality could improve this setback, we could also work on better ways to control for missing data. With more precise controls of cases, we could better interrogate interjudge variations.

Finally, JUSTFAIR only includes cases about federal criminal sentencing. Since state courts hear most criminal trials, scraping data from state courts offers even more insight into criminal sentencing in the United States. Unfortunately, each state has its own sentencing framework, record-keeping practice, and level of information availability, so grabbing state-level records will be a challenging endeavor. Nevertheless, making such data available and analyzing it is worthwhile to improve the equity of the justice system.

8.2.3 Auditing with Optimal Transport

We can improve our proposed optimal transport-based auditing framework in many ways. First, we can formalize our notions of auditing in the context of other fairness metrics. For example, individual fairness depends on the earth mover's distance, which is a special case of optimal transport. Future work can develop mathematical theory to connect the two concepts more clearly. Second, we use the 2-Wasserstein

distance to audit, which may be computationally intractable for large datasets. Instead, we could take advantage of entropic regularization. Furthermore, we use balanced optimal transport, which requires that all mass must be mapped from one distribution onto another. To relax this constraint, we could use partial optimal transport or unbalanced optimal transport, as in SCOTv2.

8.2.4 SCOT

Between our two iterations of SCOT, we have introduced an unsupervised single-cell multi-omics alignment method that can handle both balanced and unbalanced cell-type representations. In both versions, we use the coupling matrix for either projecting or embedding the datasets. However, this use of the coupling matrix does not take full advantage of its probabilistic nature. Future work will consider using the coupling matrix directly for downstream analysis, such as improved clustering or pseudo-time inference. Furthermore, the coupling matrix is limited to the existing samples in each domain, so it cannot generalize to new points. Another improvement would be to learn a map between the domains to allow adding new samples.

While we have demonstrated SCOTv2's applicability to datasets with up to 10,000 samples, single-cell datasets are growing larger as technologies improve. Large memory storage for the distance matrices limits SCOT, so future work could develop techniques to align representative subsets of each domain. Rather than aligning the full datasets, we could work on smaller pieces and then generalize. This approach, combined with the future direction of learning a map between the domains, would allow SCOT to scale better. The main challenge in this improvement would be to take representative samples for each mode.

Chapter Nine

Bibliography

Bibliography

R. Abebe, S. Barocas, J. Kleinberg, K. Levy, M. Raghavan, and D. G. Robinson. Roles for computing in social change. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 252–260, 2020.

D. S. Abrams, M. Bertrand, and S. Mullainathan. Do judges vary in their treatment of race? *J. Legal Stud.*, 41(2):347–383, 2012.

Administrative Office of the United States Courts. The Selection, Appointment, and Reappointment of United States Magistrate Judges, 2010. URL http://www.nced.uscourts.gov/pdfs/Selection-Appointment-Reappointment-of-Magistrate-Judges.pdf.

Administrative Office of the United States Courts. U.S. Courts FAQs: Filing a Case. https://www.uscourts.gov/faqs-filing-case#faq-How-are-judges-assigned-to-cases?, 2020. Accessed: 2020-12-22.

Administrative Office of the U.S. Courts. The Federal Court System in the United States, 4th ed. Technical report, United States Federal Court System, 2016. URL https://www.uscourts.gov/sites/default/files/federalcourtssystemintheus.pdf.

Administrative Office of the U.S. Courts. Public Access to Court Electronic Records, 2020a. URL https://www.pacer.gov.

Administrative Office of the U.S. Courts. A Journalist's Guide to the Federal Courts. Technical report, United States Federal Court System, 2020b. URL https://www.uscourts.gov/sites/default/files/journalists_guide_to_the_federal_courts.pdf.

A. Agarwal, A. Beygelzimer, M. Dudík, J. Langford, and H. Wallach. A reductions approach to fair classification. In *International Conference on Machine Learning*, pages 60–69. PMLR, 2018.

D. Alvarez-Melis and T. S. Jaakkola. Gromov-wasserstein alignment of word embedding spaces. *arXiv preprint arXiv:1809.00013*, 2018.

M. Amodio and S. Krishnaswamy. MAGAN: Aligning biological manifolds. In *International Conference on Machine Learning*, pages 215–223. PMLR, 2018.

J. M. Anderson, J. R. Kling, and K. Stith. Measuring interjudge sentencing disparity: Before and after the federal sentencing guidelines. *The Journal of Law and Economics*, 42(S1):271–308, 1999.

J. Angwin, J. Larson, S. Mattu, and L. Kirchner. Machine bias. *ProPublica, May*, 23:2016, 2016.

D. S. Ardia. Court Transparency and the First Amendment. *Cardozo L. Rev.*, 38: 835–919, 2016.

R. Argelaguet, S. J. Clark, H. Mohammed, L. C. Stapel, C. Krueger, C.-A. Kapourani, I. Imaz-Rosshandler, T. Lohoff, Y. Xiang, C. W. Hanna, S. Smallwood, X. Ibarra-Soria, F. Buettner, G. Sanguinetti, W. Xie, F. Krueger, B. Göttgens, P. J. Rugg-Gunn, G. Kelsey, W. Dean, J. Nichols, O. Stegle, J. C. Marioni, and W. Reik. Multi-omics profiling of mouse gastrulation at single-cell resolution. *Nature*, 576(7787):487–491, 2019. doi: 10.1038/s41586-019-1825-8. URL https://doi.org/10.1038/s41586-019-1825-8.

M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein generative adversarial networks. In *International conference on machine learning*, pages 214–223. PMLR, 2017.

D. Arnold, W. Dobbie, and C. S. Yang. Racial bias in bail decisions. *Q. J. Econ.*, 133(4):1885–1932, 2018.

M. Bando, K. Hasebe, A. Nakayama, A. Shibata, and Y. Sugiyama. Dynamical model of traffic congestion and numerical simulation. *Physical review E*, 51(2): 1035–1042, 1995.

N. Barkas, V. Petukhov, D. Nikolaeva, Y. Lozinsky, S. Demharter, K. Khodosevich, and P. V. Kharchenko. Joint analysis of heterogeneous single-cell RNA-seq dataset collections. *Nature methods*, 16(8):695–698, 2019.

W.-J. Beyn. The Numerical Computation of Connecting Orbits in Dynamical Systems. *IMA Journal of Numerical Analysis*, 10(3):379–405, 1990. ISSN 0272-4979. doi: 10.1093/imanum/10.3.379.

E. Black, S. Yeom, and M. Fredrikson. FlipTest: fairness testing via optimal transport. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 111–121, 2020.

G. Bonora, V. Ramani, R. Singh, H. Fang, D. L. Jackson, S. Srivatsan, R. Qiu, C. Lee, C. Trapnell, J. Shendure, Z. Duan, X. Deng, W. S. Noble, and C. M. Disteche. Single-cell landscape of nuclear configuration and gene expression during stem cell differentiation and X inactivation. *Genome Biology*, 22(1): 279, 2021. doi: 10.1186/s13059-021-02432-w. URL https://doi.org/10.1186/s13059-021-02432-w.

P. Brunovský. Tracking invariant manifolds without differential forms. *Acta Math. Univ. Comenian. (N.S.)*, 65(1):23–32, 1996. ISSN 0862-9544.

J. Buolamwini and T. Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91. PMLR, 2018.

J. Burrell. How the machine 'thinks: Understanding opacity in machine learning algorithms. *Big Data & Society*, 3(1):1–12, 2016. ISSN 2053-9517. doi: 10.1177/2053951715622512. URL http://bds.sagepub.com/lookup/doi/10.1177/2053951715622512.

S. D. Bushway and A. M. Piehl. Judging judicial discretion: Legal factors and racial discrimination in sentencing. *Law and Society Review*, pages 733–764, 2001.

T. Calders, F. Kamiran, and M. Pechenizkiy. Building Classifiers with Independency Constraints. In *2009 IEEE International Conference on Data Mining Workshops*, pages 13–18, 2009. doi: 10.1109/ICDMW.2009.83.

Z. Cang and Q. Nie. Inferring spatial and signaling relationships between cells from single cell transcriptomic data. *Nature communications*, 11(1):1–13, 2020.

K. Cao, X. Bai, Y. Hong, and L. Wan. Unsupervised topological alignment for single-cell multi-omics integration. *Bioinformatics*, 36(Supplement_1):i48–i56, 2020.

K. Cao, Y. Hong, and L. Wan. Manifold alignment for heterogeneous single-cell multi-omics data integration using Pamona. *Bioinformatics*, 08 2021. ISSN 1367-4803. doi: 10.1093/bioinformatics/btab594. URL https://doi.org/10.1093/bioinformatics/btab594. btab594.

J. Cart. JUSTFAIR Explorer, 2020. URL https://www.qsideinstitute.org/justfair.

S. Chen, B. B. Lake, and K. Zhang. High-throughput sequencing of transcriptome and chromatin accessibility in the same cell. *Nature Biotechnology*, 37(12):1452–1457, 2019.

Y. Chen, J. Wang, and Y. Liu. Strategic Recourse in Linear Classification. *arXiv preprint arXiv:2011.00355*, 2020.

L. F. Cheow, E. T. Courtois, Y. Tan, R. Viswanathan, Q. Xing, R. Z. Tan, D. S. Q. Tan, P. Robson, L. Yuin-Han, S. R. Quake, and W. F. Burkholder. Single-cell multimodal profiling reveals cellular epigenetic heterogeneity. *Nature Methods*, 13 (10):833–836, 2016.

S. Chiappa and A. Pacchiano. Fairness with Continuous Optimal Transport, 2021.

T. Chin, J. Ruth, C. Sanford, R. Santorella, P. Carter, and B. Sandstede. Enabling equation-free modeling via diffusion maps. *Journal of Dynamics and Differential Equations*, 2022.

A. Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments, 2016.

A. Chouldechova and A. Roth. The Frontiers of Fairness in Machine Learning. *CoRR*, abs/1810.08810, 2018. URL http://arxiv.org/abs/1810.08810.

M.-V. Ciocanel, C. Topaz, R. Santorella, S. Sen, C. Smith, and A. Hufstetler. JUSTFAIR Database, 2020a. URL https://www.qsideinstitute.org/justfair.

M.-V. Ciocanel, C. M. Topaz, R. Santorella, S. Sen, C. M. Smith, and A. Hufstetler. JUSTFAIR: Judicial System Transparency through Federal Archive Inferred Records. *PLOS One*, 15(10):e0241381, 2020b.

S. J. Clark, R. Argelaguet, C.-A. Kapourani, T. M. Stubbs, H. J. Lee, C. Alda-Catalinas, F. Krueger, G. Sanguinetti, G. Kelsey, J. C. Marioni, et al. scNMT-seq enables joint profiling of chromatin accessibility DNA methylation and transcription in single cells. *Nature communications*, 9(1):1–9, 2018.

A. Cohen and C. S. Yang. Judicial Politics and Sentencing Decisions. *Am. Econ. J. Econ. Pol.*, 11(1):160–191, 2019.

R. R. Coifman and S. Lafon. Diffusion maps. *Applied and computational harmonic analysis*, 21(1):5–30, 2006.

R. R. Coifman, S. Lafon, A. B. Lee, M. Maggioni, B. Nadler, F. Warner, and S. W. Zucker. Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps. *Proceedings of the National Academy of Sciences of the United States of America*, 102(21):7426–7431, 2005.

R. R. Coifman, I. G. Kevrekidis, S. Lafon, M. Maggioni, and B. Nadler. Diffusion maps, reduction coordinates, and low dimensional representation of stochastic systems. *Multiscale Modeling & Simulation*, 7(2):842–864, 2008.

U. S. S. Commission. Demographic differences in sentencing: An update to the 2012 Booker Report. *Federal Sentencing Reporter*, 30(3):212–229, 2018.

S. Corbett-Davies and S. Goel. The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning, 2018.

Z. Cui, H. Chang, S. Shan, and X. Chen. Generalized unsupervised manifold alignment. In *Advances in Neural Information Processing Systems*, pages 2429–2437, 2014.

M. Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in neural information processing systems*, pages 2292–2300, 2013.

P. Demetci, R. Santorella, B. Sandstede, W. S. Noble, and R. Singh. Gromov-Wasserstein optimal transport to align single-cell multi-omics data. *International Conference on Research in Computational Molecular Biology (RECOMB)*, 2021.

P. Demetci, R. Santorella, B. Sandstede, and R. Singh. Unsupervised integration of single-cell multi-omics datasets with disparities in cell-type representation. *International Conference on Research in Computational Molecular Biology (RECOMB)*, 2022.

J. K. Doerner and S. Demuth. The Independent and Joint Effects of Race/Ethnicity, Gender, and Age on Sentencing Outcomes in U.S. Federal Courts. *Justice Quarterly*, 27(1):1–27, 2010.

J. Dou, S. Liang, V. Mohanty, X. Cheng, S. Kim, J. Choi, Y. Li, K. Rezvani, R. Chen, and K. Chen. Unbiased integration of single cell multi-omics data. *bioRxiv*, 2020. doi: 10.1101/2020.12.11.422014. URL https://www.biorxiv.org/content/early/2020/12/11/2020.12.11.422014.

T. Droske. Correcting Native American Sentencing Disparity Post-Booker. *Marquette Law Review*, 91(3):723–813, 2008.

C. J. Dsilva, R. Talmon, R. R. Coifman, and I. G. Kevrekidis. Parsimonious Representation of Nonlinear Dynamical Systems Through Manifold Learning: A Chemotaxis Case Study. *Applied and Computational Harmonic Analysis*, 2015.

C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226, 2012.

R. Erban, T. A. Frewen, X. Wang, T. C. Elston, R. Coifman, B. Nadler, and I. G. Kevrekidis. Variable-free exploration of stochastic models: A gene regulatory network example. *The Journal of Chemical Physics*, 126(15), 2007.

Federal Judicial Center. History of the Federal Judiciary, 2013. URL https://www.fjc.gov/history/judges/biographical-directory-article-iii-federal-judges-export.

Federal Judicial Center. Criminal Integrated Database (IDB): 1996 to Present, 2016. URL https://www.fjc.gov/sites/default/files/idb/codebooks/Criminal%20Code%20Book%201996%20Forward.pdf.

Federal Judicial Center. Criminal Defendants Filed, Terminated, and Pending from FY 1996 to present, 2019. URL https://www.fjc.gov/research/idb.

Federal Judicial Center. The Integrated Database: A Research Guide, 2020. URL https://www.fjc.gov/sites/default/files/IDB-Research-Guide.pdf.

M. Feldman, S. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian. Certifying and removing disparate impact, 2015.

B. Feldmeyer and J. T. Ulmer. Racial/ethnic threat and federal sentencing. *J. Res. Crime Delinq.*, 48(2):238–270, 2011.

G. M. Fenner and J. L. Koley. Access to Judicial Proceedings: To Richmond Newspapers and Beyond. *Harv. CR-CLL Rev.*, 16:459, 1981.

J. B. Fischman and M. M. Schanzenbach. Racial disparities under the federal sentencing guidelines: The role of judicial discretion and mandatory minimums. *J. Empir. Legal Stud.*, 9(4):729–764, 2012.

R. Flamary and N. Courty. POT Python Optimal Transport library, 2017. URL https://github.com/rflamary/POT.

T. W. Franklin. Sentencing Native Americans in US Federal Courts: An Examination of Disparity. *Justice Quarterly*, 30(2):310–339, 2013.

Free Law Project. Juriscraper Legal Scraping Toolkit, 2020. URL https://free.law/juriscraper/.

P. Gordaliza, E. Del Barrio, G. Fabrice, and J.-M. Loubes. Obtaining fairness using optimal transport theory. In *International Conference on Machine Learning*, pages 2357–2365, 2019.

V. Gupta, P. Nokhiz, C. D. Roy, and S. Venkatasubramanian. Equalizing recourse across groups. *arXiv preprint arXiv:1909.03166*, 2019.

A. Hanna, E. Denton, A. Smart, and J. Smith-Loud. Towards a Critical Race Methodology in Algorithmic Fairness. *CoRR*, abs/1912.03593, 2019. URL http://arxiv.org/abs/1912.03593.

M. Hardt, E. Price, and N. Srebro. Equality of Opportunity in Supervised Learning. *CoRR*, abs/1610.02413, 2016. URL http://arxiv.org/abs/1610.02413.

R. D. Hartley and R. Tillyer. Defending the Homeland: Judicial Sentencing Practices for Federal Immigration Offenses. *Justice Quarterly*, 29(1):76–104, 2012.

A. L. Hoffmann. Where fairness fails: data, algorithms, and the limits of antidiscrimination discourse. *Information, Communication & Society*, 22(7):900–915, 2019. doi: 10.1080/1369118X.2019.1573912. URL https://doi.org/10.1080/1369118X.2019.1573912.

H. Hofmann. https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data), 2000.

L. Hu and Y. Chen. Fair Classification and Social Welfare. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 535–545, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450369367. doi: 10.1145/3351095.3372857. URL https://doi.org/10.1145/3351095.3372857.

C. Ilvento. Metric Learning for Individual Fairness. *CoRR*, abs/1906.00250, 2019. URL http://arxiv.org/abs/1906.00250.

R. Jiang, A. Pacchiano, T. Stepleton, H. Jiang, and S. Chiappa. Wasserstein Fair Classification, 2019.

J. E. Johndrow, K. Lum, et al. An algorithm for removing sensitive information: application to race-independent recidivism prediction. *The Annals of Applied Statistics*, 13(1):189–220, 2019.

B. Johnson. RECAP: Cracking open US courtrooms. *The Guardian*, 2009.

B. D. Johnson and S. Betsinger. Punishing the "model minority": Asian-American criminal sentencing outcomes in federal district courts. *Criminology*, 47(4):1045–1090, 2009.

C. K. R. T. Jones. Geometric singular perturbation theory. In *Dynamical systems (Montecatini Terme, 1994)*, volume 1609 of *Lecture Notes in Math.*, pages 44–118. Springer, Berlin, 1995. doi: 10.1007/BFb0095239. URL https://doi.org/10.1007/BFb0095239.

C. Jung, M. Kearns, S. Neel, A. Roth, L. Stapleton, and Z. S. Wu. Eliciting and enforcing subjective individual fairness. *arXiv preprint arXiv:1905.10660*, 2019.

A.-H. Karimi, G. Barthe, B. Schölkopf, and I. Valera. A survey of algorithmic recourse: definitions, formulations, solutions, and prospects. *arXiv preprint arXiv:2010.04050*, 2020a.

A.-H. Karimi, B. Schölkopf, and I. Valera. Algorithmic recourse: from counterfactual explanations to interventions. *arXiv preprint arXiv:2002.06278*, 2020b.

I. G. Kevrekidis and G. Samaey. Equation-Free Multiscale Computation: Algorithms and Applications. *Annual Review of Physical Chemistry*, 60:321–344, 2009.

I. G. Kevrekidis, C. W. Gear, J. M. Hyman, P. G. Kevrekidis, O. Runborg, C. Theodoropoulos, et al. Equation-free, coarse-grained multiscale computation: Enabling microscopic simulators to perform system-level analysis. *Communications in Mathematical Sciences*, 1(4):715–762, 2003.

I. G. Kevrekidis, C. W. Gear, and G. Hummer. Equation-Free: The Computer-Aided Analysis of Complex Multiscale Systems. *American Institute of Chemical Engineers Journal*, 50(7):1346–1355, 2004.

C. Kitchens. Research Notes Issue 1: Collection of Individual Offender Data. Technical report, United States Sentencing Commission, 2019a. URL https://www.ussc.gov/sites/default/files/pdf/research-and-publications/research-notes/20190719_Research-Notes-Issue1.pdf.

C. Kitchens. Research Notes Issue 2: Analyzing Federal Sentencing Length & Type. Technical report, United States Sentencing Commission, 2019b. URL https://www.ussc.gov/sites/default/files/pdf/research-and-publications/research-notes/20190926_Research-Notes-Issue2.pdf.

C. Kuehn. *Multiple time scale dynamics*, volume 191 of *Applied Mathematical Sciences*. Springer, Cham, 2015. URL https://doi.org/10.1007/978-3-319-12316-5.

K. Kwegyir-Aggrey, R. Santorella, and S. M. Brown. Everything is Relative: Understanding Fairness with Optimal Transport. *arXiv preprint arXiv:2102.10349*, 2021.

T. Kyckelhahn. Research Notes Issue 3: Offense Type & Sentencing Guideline Analysis Issues Prior to Fiscal Year 2018. Technical report, United States Sentencing Commission, 2019. URL https://www.ussc.gov/sites/default/files/pdf/research-and-publications/research-notes/20191114_Research-Notes-Issue3.pdf.

S. S. Lafon. *Diffusion maps and geometric harmonics*. PhD thesis, Yale University, 2004.

C. R. Laing, T. A. Frewen, and I. G. Kevrekidis. Coarse-grained dynamics of an activity bump in a neural field model. *Nonlinearity*, 20(9):2127, 2006.

J. Larson, S. Mattu, L. Kirchner, and J. Angwin. https://github.com/propublica/compas-analysis, 2016a.

J. Larson, S. Mattu, L. Kirchner, and J. Angwin. How we analyzed the COMPAS recidivism algorithm. *ProPublica (5 2016)*, 9(1), 2016b.

M. Liero, A. Mielke, and G. Savaré. Optimal entropy-transport problems and a new Hellinger–Kantorovich distance between positive measures. *Inventiones mathematicae*, 211(3):969–1117, 2018.

M. T. Light. The new face of legal inequality: Noncitizens and the long-term trends in sentencing disparities across US district courts, 1992–2009. *Law Soc. Rev.*, 48 (2):447–478, 2014.

M. T. Light. The Declining Significance of Race in Criminal Sentencing: Evidence from US Federal Courts. *Social Forces*, 2021.

M. T. Light, M. Massoglia, and R. D. King. Citizenship and punishment: The salience of national membership in US criminal courts. *Am. Sociol. Rev.*, 79(5): 825–847, 2014.

J. Liu, Y. Huang, R. Singh, J.-P. Vert, and W. S. Noble. Jointly embedding multiple single-cell omics measurements. *BioRxiv*, page 644310, 2019.

P. Liu, H. R. Safford, I. D. Couzin, and I. G. Kevrekidis. Coarse-grained variables for particle-based models: diffusion maps and animal swarming simulations. *Computational Particle Mechanics*, 1(4):425–440, 2014.

M. Lohaus, M. Perrot, and U. V. Luxburg. Too Relaxed to Be Fair. In H. D. III and A. Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 6360–6369. PMLR, 13–18 Jul 2020.

C. Marschler, J. Sieber, R. Berkemer, A. Kawamoto, and J. Starke. Implicit Methods for Equation-Free Analysis: Convergence Results and Analysis of Emergent Waves in Microscopic Traffic Models. *SIAM Journal on Applied Dynamical Systems*, 13 (3):1202–1238, 2014a.

C. Marschler, J. Starke, P. Liu, and I. G. Kevrekidis. Coarse-grained particle model for pedestrian flow using diffusion maps. *Physical Review E*, 89(1), 2014b.

F. Mémoli. Gromov–Wasserstein distances and the metric approach to object matching. *Foundations of computational mathematics*, 11(4):417–487, 2011.

G. Morina, V. Oliinyk, J. Waton, I. Marusic, and K. Georgatzis. Auditing and Achieving Intersectional Fairness in Classification Problems. *arXiv preprint arXiv:1911.01468*, 2019.

D. B. Mustard. Racial, Ethnic, and Gender Disparities in Sentencing: Evidence from the US Federal Courts. *J. Law Econ.*, 44(1):285–314, 2001.

M. Nitzan, N. Karaiskos, N. Friedman, and N. Rajewsky. Gene expression cartography. *Nature*, 576(7785):132–137, 2019.

Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464): 447–453, 2019.

G. Peyré, M. Cuturi, and J. Solomon. Gromov-wasserstein averaging of kernel and distance matrices. In *International Conference on Machine Learning*, pages 2664–2672, 2016.

G. Peyré, M. Cuturi, et al. Computational optimal transport. *Foundations and Trends in Machine Learning*, 11(5-6):355–607, 2019.

N. Quadrianto and V. Sharmanska. Recycling privileged learning and distribution matching for fairness. *Advances in Neural Information Processing Systems*, 30: 677–688, 2017.

N. Quadrianto, J. Petterson, and A. Smola. Distribution Matching for Transduction. In Y. Bengio, D. Schuurmans, J. Lafferty, C. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems*, volume 22. Curran Associates, Inc., 2009. URL https://proceedings.neurips.cc/paper/2009/file/2bb232c0b13c774965ef8558f0fbd615-Paper.pdf.

J. J. Rachlinski and A. J. Wistrich. Judging the judiciary by the numbers: Empirical research on judges. *Ann. Rev. Law Soc. Sci.*, 13:203–229, 2017.

I. D. Raji and J. Buolamwini. Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial ai products. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 429–435, 2019.

I. D. Raji, A. Smart, R. N. White, M. Mitchell, T. Gebru, B. Hutchinson, J. Smith-Loud, D. Theron, and P. Barnes. Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 33–44, 2020.

K. Rawal and H. Lakkaraju. Interpretable and Interactive Summaries ofActionable Recourses. *arXiv preprint arXiv:2009.07165*, 2020.

L. Reedt, C. Semisch, and K. Blackwell. Effective Use of Federal Sentencing Data. Technical report, United States Sentencing Commission, 2013. URL https://www.ussc.gov/sites/default/files/pdf/research-and-publications/datafiles/20131122-ACS-Presentation.pdf.

M. M. Rehavi and S. B. Starr. Racial Disparity in Federal Criminal Sentences. *J. Pol. Econ.*, 122(6):1320–1354, 2014.

M. M. Schanzenbach and E. H. Tiller. Reviewing the Sentencing Guidelines: Judicial Politics, Empirical Evidence, and Reform. *U. Chi. Law Rev.*, 75(2):715–760, 2008.

G. Schiebinger, J. Shu, M. Tabaka, B. Cleary, V. Subramanian, A. Solomon, J. Gould, S. Liu, S. Lin, P. Berube, et al. Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming. *Cell*, 176 (4):928–943, 2019.

R. Schunck. Within and between estimates in random-effects models: Advantages and drawbacks of correlated random effects and hybrid models. *The Stata Journal*, 13(1):65–76, 2013.

R. W. Scott. Inter-judge sentencing disparity after Booker: A first look. *Stanford Law Rev.*, 63(1):1–66, 2010.

A. D. Selbst, D. Boyd, S. A. Friedler, S. Venkatasubramanian, and J. Vertesi. Fairness and Abstraction in Sociotechnical Systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 59–68, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450361255. doi: 10.1145/3287560.3287598. URL https://doi.org/10.1145/3287560.3287598.

C. Silvia, J. Ray, S. Tom, P. Aldo, J. Heinrich, and A. John. A General Approach to Fairness with Optimal Transport. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 3633–3640, 2020.

R. Singh, P. Demetci, G. Bonora, V. Ramani, C. Lee, H. Fang, Z. Duan, X. Deng, J. Shendure, C. Disteche, et al. Unsupervised manifold alignment for single-cell multi-omics data. In *Proceedings of the 11th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, pages 1–10, 2020.

C. M. Smith, N. Goldrosen, M.-V. Ciocanel, R. Santorella, C. M. Topaz, and S. Sen. Racial Disparities in Criminal Sentencing Vary Considerably across Federal Judges, Jul 2021. URL osf.io/preprints/socarxiv/j2gbn.

B. E. Sonday, M. Haataja, and I. G. Kevrekidis. Coarse-graining the dynamics of a driven interface in the presence of mobile impurities: Effective description via diffusion maps. *Physical Review E*, 80(3), 2009.

S. Starr. Did Booker Increase Disparity? Why the Evidence is Unpersuasive. *Fed. Sent. Rep.*, 25(5):323–326, 2013.

S. B. Starr. Estimating Gender Disparities in Federal Criminal Cases. *Am. Law Econ. Rev.*, 17(1):127–159, 2015.

T. Stuart, A. Butler, P. Hoffman, C. Hafemeister, E. Papalexi, W. M. M. III, Y. Hao, M. Stoeckius, P. Smibert, and R. Satija. Comprehensive integration of single-cell data. *Cell*, 77(7):1888–1902, 2019.

C. Syckes. Issue 4: The Significance of [SOURCES]: Resolving Guideline Application Discrepancies in Federal Sentencing Documents. Technical report, United States Sentencing Commission, 2020. URL https://www.ussc.gov/sites/default/files/pdf/research-and-publications/research-notes/20200331_Research-Notes-Issue4.pdf.

T. Séjourné, F.-X. Vialard, and G. Peyré. The Unbalanced Gromov Wasserstein Distance: Conic Formulation and Relaxation. *arXiv*, 2021.

B. Taskesen, J. Blanchet, D. Kuhn, and V. A. Nguyen. A Statistical Test for Probabilistic Fairness, 2020.

Transactional Records Access Clearninghouse. TRAC Terms of Usage, 2002. URL https://tracfed.syr.edu/notices/terms.html.

J. T. Ulmer and M. S. Bradley. Punishment in Indian Country: Ironies of Federal Punishment of Native Americans. *Justice Quarterly*, 35(5):751–781, 2018.

United States Sentencing Commission. An Overview of the U.S. Sentencing Commission, 2011. URL https://www.ussc.gov/sites/default/files/pdf/about/overview/USSC_Overview.pdf.

United States Sentencing Commission. Federal Sentencing: The Basics, 2018. URL https://www.ussc.gov/sites/default/files/pdf/research-and-publications/research-projects-and-surveys/miscellaneous/201811_fed-sentencing-basics.pdf.

United States Sentencing Commission. Individual Offender Files, 2019. URL https://www.ussc.gov/research/datafiles/commission-datafiles.

United States Sentencing Commission. Variable Codebook for Individual Offenders, 2020. URL https://www.ussc.gov/sites/default/files/pdf/research-and-publications/datafiles/Variable_Codebook_for_Individual_Offenders.pdf.

L. Van der Maaten and G. Hinton. Visualizing data using t-SNE. *Journal of machine learning research*, 9(11), 2008.

C. Villani. *Optimal transport – Old and new*, volume 338, pages xxii+973. Springer, 01 2008. doi: 10.1007/978-3-540-71050-9.

C. Wang and S. Mahadevan. Manifold alignment without correspondence. In *Twenty-First International Joint Conference on Artificial Intelligence*, 2009.

H. Wang, B. Ustun, and F. Calmon. Repairing without Retraining: Avoiding Disparate Impact with Counterfactual Distributions. In K. Chaudhuri and R. Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 6618–6627. PMLR, 09–15 Jun 2019. URL http://proceedings.mlr.press/v97/wang19l.html.

J. D. Welch, A. J. Hartemink, and J. F. Prins. MATCHER: manifold alignment reveals correspondence between single cell transcriptome and epigenome dynamics. *Genome biology*, 18(1):138, 2017.

J. D. Welch, V. Kozareva, A. Ferreira, C. Vanderburg, C. Martin, and E. Z. Macosko. Single-cell multi-omic integration compares and contrasts features of brain cell identity. *Cell*, 177(7):1873–1887, 2019.

Wikipedia contributors. List of United States district and territorial courts — Wikipedia, The Free Encyclopedia, 2020. URL https://en.wikipedia.org/wiki/List_of_United_States_district_and_territorial_courts.

C. Wilson, A. Ghosh, S. Jiang, A. Mislove, L. Baker, J. Szary, K. Trindel, and F. Polli. Building and Auditing Fair Algorithms: A Case Study in Candidate Screening. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 666–677, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383097. doi: 10.1145/3442188.3445928. URL https://doi-org.uri.idm.oclc.org/10.1145/3442188.3445928.

Y. Wu, L. Zhang, and X. Wu. On Convexity and Bounds of Fairness-Aware Classification. In *The World Wide Web Conference*, WWW '19, page 3356–3362, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450366748. doi: 10.1145/3308558.3313723. URL https://doi.org/10.1145/3308558.3313723.

S. Xue, M. Yurochkin, and Y. Sun. Auditing ML Models for Individual Bias and Unfairness. In *International Conference on Artificial Intelligence and Statistics*, pages 4552–4562. PMLR, 2020.

C. S. Yang. Have Interjudge Sentencing Disparities Increased in an Advisory Guidelines Regime? Evidence from Booker. *NYUL Rev.*, 89:1268–1342, 2014.

C. S. Yang. Free at Last? Judicial Discretion and Racial Disparities in Federal Sentencing. *J. Leg. Stud.*, 44(1):75–111, 2015.

K. D. Yang and C. Uhler. Multi-Domain Translation by Learning Uncoupled Autoencoders. *arXiv preprint arXiv:1902.03515*, 2019.

K. D. Yang, K. Damodaran, S. Venkatchalapathy, A. C. Soylemezoglu, G. Shivashankar, and C. Uhler. Autoencoder and Optimal Transport to Infer Single-Cell Trajectories of Biological Processes. *bioRxiv*, page 455469, 2018.

M. B. Zafar, I. Valera, M. G. Rodriguez, and K. P. Gummadi. Fairness Constraints: Mechanisms for Fair Classification, 2017.

L. Zappia, B. Phipson, and A. Oshlack. Splatter: simulation of single-cell RNA sequencing data. *Genome biology*, 18(1):1–15, 2017.

M. Zehlike, P. Hacker, and E. Wiedemann. Matching code and law: achieving algorithmic fairness with optimal transport. *Data Mining and Knowledge Discovery*, 34(1):163–200, 2020.