Manifold Interpolating Optimal-Transport Flows for Trajectory Inference

Anonymous Author(s)
Affiliation
Address
email

Abstract

Here, we present a method called Manifold Interpolating Optimal-Transport Flow (MIOFlow) that learns stochastic, continuous population dynamics from static snapshot samples taken at sporadic timepoints. MIOFlow combines dynamic models, manifold learning, and optimal transport by training neural ordinary differential equations (Neural ODE) to interpolate between static population snapshots as penalized by optimal transport with manifold ground distance. Further, we ensure that the flow follows the geometry by operating in the latent space of an autoencoder that we call a geodesic autoencoder (GAE). In GAE the latent space distance between points is regularized to match a novel multiscale geodesic distance on the data manifold that we define. We show that this method is superior to normalizing flows, Schrödinger bridges and other generative models that are designed to flow from noise to data in terms of interpolating between populations. Theoretically, we link these trajectories with dynamic optimal transport. We evaluate our method on simulated data with bifurcations and merges, as well as scRNA-seq data from embryoid body differentiation, and acute myeloid leukemia treatment.

1 Introduction

Recently several neural networks that perform flows, or transports between probability distributions, have been proposed in the literature. However, most have these works have focused on generative modeling: i.e., flowing from a noise distribution such as a Gaussian distribution to a data distribution in order to generate data. Examples include score-based generative matching [32, 33], diffusion models [17], Schrödinger bridges [26, 9], and continuous normalizing flows (CNF) [6, 14]. However, here we focus on a different problem—learning the continuous dynamics of probability distributions that are measured as static snapshots over time. Such distributions occur in naturalistic systems where populations are measured at sporadic intervals but individuals are not (or cannot be followed) over time. An example in modern biology is single-cell data that is measured over a time course of weeks measuring cellular development, where the measurements are destructive to the cell at each timepoint but the entire population is followed over time. To continuously interpolate such populations, we propose Manifold Interpolation Optimal-transport Flow (MIOFlow)[1], a new framework for flows based on optimal transport penalties applied to dynamics learned in a manifold embedding. MIOFlow is a neural ordinary differential equations (Neural ODE) [6] network to transport a sampling of high dimensional data points between timepoints such that 1. the transport occurs on a manifold defined

[1]The code will be publicly available

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

by samples, 2. the transport is penalized to agree with measured timepoints using Wasserstein, 3. the transport is inherently stochastic.

The key works tackling a similar problem include TrajectoryNet [35], dynamics modeled by a recurrent neural network [16], and Waddington-OT [30]. TrajectoryNet is based on a continuous normalizing flow where the magnitude of the derivative is penalized to create a more optimal flow. However, this approach suffers from several drawbacks. The first drawback is the requirement of starting from a Gaussian distribution. It can be difficult to match distributions from the real world to Gaussian distributions in an interpolating sense, all intermediate distributions must be penalized to start from a Gaussian rather than being trained to flow from one to the next. Second, continuous normalizing flows are deterministic. Hence, to model the intrinsic stochasticity in biology, we have to force chaotic dynamics starting from slight noise added to any initial distributions. Third, continuous normalizing flow models require calculating the trace of the Jacobian, which requires $O(d^2)$ operations to compute [6], making our method d times faster per function evaluation [6].

Additionally, unlike normalizing flows which operate in ambient data dimensions, we focus on flows on a data manifold. We feature a two-pronged approach with which to enforce this. First, propose the *Geodesic Autoencoder* to embed the data such that distances in the latent space match a new manifold distance called diffusion geodesic distance. Second, we penalize the transport using a manifold optimal transport method [27, 40]. These two steps enforce flows on the manifold, whereas in earlier works such as TrajectoryNet KL-divergence is used to match distributions, and in [22] access to the metric tensor of a Riemannian manifold is required.

The main contributions of our work include:

- The MIOFlow framework for efficiently learning continuous stochastic dynamics of static snapshot data based on Neural ODEs that implement optimal transport flows on a data manifold.
- The Geodesic Autoencoder to create a manifold embedding of high dimensional data in the latent space of the autoencoder.
- A new multiscale manifold distance called diffusion geodesic distance, and theoretical results showing its convergence to a geodesic on the manifold in the limit of infinitely many points.
- Theoretical results relating MIOFlow trajectories to dynamic optimal transport.
- Empirical results showing that our flows can model divergent trajectories on toy data.
- Empirical results on single-cell developmental and treatment response data.

2 Related Work

Static methods such as pseudotime assign an abstract measure of time for all cells depending on an initial state [15, 38] in can be combined to infer the cell lineage, i.e. the principal curves or trajectories [34]. When estimates of cell growth rates are available, Zhang et al. [41] proposed a method based on optimal transport. Another line of research is on the direct rate of differentiation of cell or RNA velocity [18, 3]. This method does not provide trajectories, but it can be combined to define a transition matrix to infer trajectories [42]. Saelens et al. [29] provide a rigorous comparison of 45 static methods.

The task of recovering dynamics from multiple snapshots has gained a lot of attention, especially in biology with scRNA-seq data. Similar methods used optimal transport [30] either with a recurrent neural network [16], generative adversarial networks [28], flows defined by the gradient of an input convex neural network [4], or a CNF [35]. Here, we circumvent the need for a CNF by modeling the trajectories directly on the support of the distributions, hence maintaining the computational cost low even in high dimension. Some generative methods that interpolate between a prior to a target can also be used to recover trajectories. For example [9], solve the Schrödinger bridge (SB) between

two distributions. SB methods however use a series of sinkhorn iterations and do not explicitly learn dynamics, making their interpolations inaccurate.

3 Preliminaries and Background

3.1 Problem Formulation and Notation

We consider a distribution μ_t over $\mathbb{R}^d$ evolving over time $t \in \mathbb{R}$, from which we only observe samples from a finite set of T distributions $\{\mu_i\}_{i=0}^{T-1}$. For each time t, we observe a sample $\mathsf{X}_t \sim \mu_t$ of size n_t. We note the set of all observations X of size $n := \sum_i n_i$. We also want to characterize the evolution of the support of μ_t. We aim to define a trajectory from an initial points $X_0 \sim \mu_0$ to $X_{T-1} \sim \mu_{T-1}$, given the intermediate conditions $X_2 \sim \mu_2, \dots, X_{T-1} \sim \mu_{T-1}$. We are thus interested in matching distributions given a set of T samples $\{\mathsf{X}_i\}_{i=1}^{T-1}$, and initial condition X_0.

In the following, we assume that the distributions are absolutely continuous with respect to the Lebesgue measure, and we use the same notation for the distribution and its density. We note the equivalence between two distances $d_1 \simeq d_2$, meaning that there exist. We assume that all Stochastic Differential Equations (SDE) admit a solution [12]. All proofs are presented in the supplementary material.

3.2 Optimal Transport

In this section, we provide a brief overview of optimal transport, our primary approach for interpolating between distributions using a neural ODE. We also review its dynamic formulation based on two monographs [27, 40]. The Monge problem, the initial optimal transport formulation, is to find a transport plan $T : \mathcal{X} \to \mathcal{Y}$ that minimizes the total cost $\sum d(x, T(x))$ of moving elements in $\mathcal{X}$ to $\mathcal{Y}$ for a given cost function $d : \mathcal{X} \times \mathcal{Y} \to [0, \infty)$. With this formulation, it is not possible to divide the mass of an element in $\mathcal{X}$ to multiple elements in $\mathcal{Y}$. The Kantorovich formulation solves this issue by formulating the problem with distributions. We consider two distributions μ and ν defined on $\mathcal{X}$ and $\mathcal{Y}$, and $\Pi(\mu, \nu)$ the set of joint distributions on $\mathcal{X} \times \mathcal{Y}$ with marginals μ and ν, i.e. $\pi(dx, \mathcal{Y}) = \mu(dx)$ and $\pi(\mathcal{X}, dy) = \nu(dy)$ for $\pi \in \Pi(\mu, \nu)$. The transport plan π moves the mass from μ to ν, where the cost of moving a unit mass from the initial $x \in \mathcal{X}$ to the final $y \in \mathcal{Y}$ is $d(x, y)$. This formulation gives rise to the p-Wasserstein distance

$$W_p(\mu, \nu) := \left(\inf_{\pi \in \Pi(\mu,\nu)} \int_{\mathcal{X} \times \mathcal{Y}} d(x, y)^p \pi(dx, dy) \right)^{1/p},$$

where $p \in [1, \infty)$. For the rest of the paper, we suppose $\mathcal{X} = \mathcal{Y} = \mathbb{R}^d$, and $d(x, y) := ||x - y||_2$, the Euclidean distance. With these assumptions, Benamou and Brenier [2] provide a *dynamic formulation* of the optimal transport problem. For simplicity, we assume a fix time interval $[0, 1]$, but note that the following holds for any time interval $[t_0, t_1]$, and we let $\mathcal{C} := C([0, 1], \mathbb{R}^d)$ the space of $\mathbb{R}^d$ valued continuous function. The transport plan is replaced by a time-evolving distribution ρ_t, such that $\rho_0 = \mu$ and $\rho_1 = \nu$. The movement of mass is described by a time-evolving vector field $v(x, t)$. It is also required that the total mass of ρ_t must be the same for all $t \in [0, 1]$ (see [7] for the unbalanced case). That is ρ_t must satisfy the conservation of mass (or continuity equation)

$$\partial_t \rho_t + \nabla \cdot (\rho_t v) = 0, \tag{1}$$

where ∇ is the divergence operator. When ρ_t satisfies these two conditions, Benamou and Brenier [2] show that

$$W_2(\mu, \nu)^2 = \inf_{(\rho_t, v)} \int_0^1 \int_{\mathbb{R}^d} \|v(x, t)\|^2 \rho_t(dx) dt. \tag{2}$$

This formulation arises from the field of fluid mechanics; the optimal vector field is the divergence of a pressureless potential flow. It can also be formulated on the path space of $\mathbb{R}^d$

$$W_2(\mu, \nu)^2 = \inf_{X_t} \mathbb{E}\left[\int_0^1 \|f(X_t, t)\|_2^2 dt \right] \text{ s.t. } dX_t = f(X_t, t)dt,\ X_0 \sim \mu,\ X_1 \sim \nu, \tag{3}$$

where the infimum is over all absolutely continuous stochastic path X_t on $\mathcal{C}$ (see [23]).

Adding Diffusion For various applications, it can be interesting to incorporate a diffusion term in the paths or trajectories. These types of flows are often used in control dynamics [1, 24, 13], and mean field games [19]. We consider an SDE $dX_t = f(X_t, t)dt + \sqrt{\sigma}dB_t$, where B_t is a standard Brownian motion. To model this new dynamic, one can replace (1) with the Fokker–Planck equation of this SDE with diffusion, i.e. $\partial_t \rho_t + \nabla \cdot (\rho_t v) = \Delta(\sigma \rho_t/2)$. From this formulation, one can also retrieve the Benamou-Brenier optimal transport. Indeed, Mikami [23] established the convergence to the W_2 when the diffusion term goes to zero. Similar to (3), we can phrase the problem using an SDE

$$\inf_f \mathbb{E}\left[\int_0^1 \|f\|_2^2 dt\right] \text{ s.t. } dX_t = f(X_t, t)dt + \sqrt{\sigma}dB_t,\ X_0 \sim \mu,\ X_1 \sim \nu. \tag{4}$$

The two previous formulations admit the same entropic interpolation ρ_t [26, 23]. In this paper, we utilize such a diffusion term, with the knowledge that it still converges to the transport between distributions. Moreover, as the diffusion goes to zero, the infimum converges to the W_2.

The dynamic formulations (2) and (3) are equivalent with W_2 holds when the ground distance is the Euclidean. In practice, we might have observations sampled from or generated by a lower dimensional manifold. Thus using the Euclidean distance in the ambient space ignores the underlying geometry. In the following section, we present theoretical results that will justify our approximation of a distance equivalent to the geodesic distance on a closed Riemannian manifold.

3.3 Manifold Geodesics

Part of MIOFlow is a Geodesic Autoencoder to learn a representation $\mathcal{Z}$, such that the L^2 norm in $\mathcal{Z}$ is equivalent to the geodesic on a manifold $\mathcal{M}$. To motivate this, we use results from Riemannian geometry.

We consider a closed Riemannian manifold $(\mathcal{M}, d_{\mathcal{M}})$, where $d_{\mathcal{M}}$ is the geodesic distance, and we note h_t the heat kernel on $\mathcal{M}$. For a point $x \in \mathcal{M}$, the heat kernel $h_t(x, \cdot)$ induces a measure, i.e. how the heat has propagated on the manifold at time t given an arbitrary initial distribution. The diffusion ground distance between $x, y \in \mathcal{M}$ is based on the L^1 norm between the measures induced by the heat kernel given the initial conditions δ_x and δ_y.

Definition 1. The diffusion ground distance between $x, y \in \mathcal{M}$ is

$$D_\alpha(x, y) := \sum_{k \geq 0} 2^{-k\alpha} ||h_{2^{-k}}(x, \cdot) - h_{2^{-k}}(y, \cdot)||_1,$$

for $\alpha \in (0, 1/2)$, the scale parameter $k \geq 0$, and h_t the heat kernel on $\mathcal{M}$.

Next we state an important result from Leeb and Coifman [20] (Thm. 2), this theorem links the diffusion ground distance and the geodesic on a Riemannian manifold.

Theorem 1. *Let $(\mathcal{M}, d_{\mathcal{M}})$ a closed Riemannian manifold, with geodesic $d_{\mathcal{M}}$, for $\alpha \in (0, 1/2)$, the distance D_α is equivalent to $d_{\mathcal{M}}^{2\alpha}$.*

In practice, we cannot always evaluate the heat kernel for an arbitrary manifold. However, convergence results from Coifman and Lafon [8] provide a natural approximation. For a rotation invariant kernel $k_\epsilon : \mathcal{M} \times \mathcal{M} \to \mathbb{R}^+$, where $\epsilon > 0$ is a bandwidth parameter, we define $k_{\epsilon,\beta}(x, y) := k_\epsilon(x, y)/q^\beta(x)q^\beta(y)$ where $q(x) := \int k_\epsilon(x, y)d\mu(y)$, and μ is the underlying distribution on $\mathcal{M}$. The kernel $k_{\epsilon,\beta}$ aims to capture the local geometry of the manifold. To mimic the propagation of the heat, we define the transition kernel $p_{\epsilon,\beta}(x, y) := k_{\epsilon,\beta}(x, y)/q_\beta(x)$ where $q_\beta(x) := \int k_{\epsilon,\beta}(x, y)d\mu(y)$. We note its resulting operator $P_{\epsilon,\beta}$. For $\beta = 1$, it is shown that the operator $P_{\epsilon,1}^{t/\epsilon}$ will converge to the heat operator as ϵ goes to zero (Prop. 3 [8]). For the rest of the paper we assume $\beta = 1$, and we define $P_\epsilon := P_{\epsilon,1}$.

4 Manifold Interpolating Optimal-Transport Flow

To learn individual trajectories from multiple cross-sectional samples of a population, we propose MIOFlow shown in Fig. 1. Our method consists of two steps. We first learn an embedding $\mathcal{Z}$ which

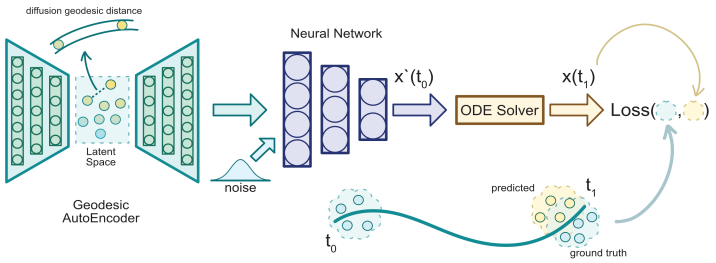

Figure 1: Overview of the MIOFlow pipeline. Geodesic Autoencoder learns the latent space of representation of data is learnt such that is preserves a diffusion geodesic distance. With the combination of noise a neural network learns to predict the derivative $x'(t_0)$ and an ODE Solver produces the predicted $x(t_1)$. At a population level, predictions of points at time t_1 are penalized by the Wasserstein distance during training.

preserves the diffusion geodesic distance—which we define using our Geodesic Autoencoder. Then, we learn continuous trajectories based on an optimal transport loss. We model the trajectories with a Neural ODE [6], allowing us to learn non-linear paths and interpolate between timepoints.

4.1 Geodesic Autoencoder Embedding

In light of Thm. 1, define an approximation of the diffusion ground distance D_α (Def. 1) from a collection of samples X that we call *diffusion geodesic distance*.

Our approximation relies on a discrete version of the operator P_ϵ. We build a graph with the affinity matrix defined by $(\mathbf{K}_\epsilon)_{ij} := k_\epsilon(x_i, x_j)$, and we consider the density normalized matrix $\mathbf{M}_\epsilon := \mathbf{Q}^{-1}\mathbf{K}_\epsilon\mathbf{Q}^{-1}$, where $\mathbf{Q}$ is a diagonal matrix with $\mathbf{Q}_{ii} := \sum_j (\mathbf{K}_\epsilon)_{ij}$. In practice, we choose k_ϵ to be a Gaussian kernel or the α-decay kernel [25]. Lastly, a Markov diffusion operator is defined by $\mathbf{P}_\epsilon := \mathbf{D}^{-1}\mathbf{M}_\epsilon$, where $\mathbf{D}_{ii} := \sum_{j=1}^n (\mathbf{M}_\epsilon)_{ij}$ is a diagonal matrix. The stationary distribution associated to $\mathbf{P}_\epsilon$ is $\boldsymbol{\pi}$, where $\boldsymbol{\pi}_i = \mathbf{D}_{ii}/\sum_j \mathbf{D}_{jj}$, since $\mathbf{P}_\epsilon$ is $\boldsymbol{\pi}$-reversible. Finally, we note that the matrix $\mathbf{P}_\epsilon$ follows a similar construction as the kernel p_ϵ; the integrals are approximated by sums. We refer to [8] for more details about the convergence of the matrix operator $\mathbf{P}_\epsilon$ to the operator P_ϵ. Until now, we have defined a matrix $\mathbf{P}_\epsilon$ that approximates the operator P_ϵ, which in turn converges to the heat operator. Now, we define an approximation of the diffusion ground distance, based on the matrix $\mathbf{P}_\epsilon$. We use the notation $(\mathbf{P}_\epsilon)^t_{i:}$ to represent the i-th row of $\mathbf{P}^t_\epsilon$, it represents the transition probabilities of a t-steps random walk started at x_i.

Definition 2. We define the *diffusion geodesic distance* between $x_i, x_j \in \mathsf{X}$ as

$$G(x_i, x_j) := \sum_{k=0}^{K} 2^{-(K-k)/2} ||(\mathbf{P}_\epsilon)^{2^k}_{i:} - (\mathbf{P}_\epsilon)^{2^k}_{j:}||_1 + 2^{-(K+1)/2}||\boldsymbol{\pi}_i - \boldsymbol{\pi}_j||_1.$$

The diffusion geodesic compares the transitions probabilities of a random walk at various scales given two different initial states x_i, x_j. With a slight abuse of notation, a similar distance between distributions (rather than points) is constructed in [37] where it was applied to approximate the Earth Mover's Distance with a geodesic ground cost.

Training To use the diffusion geodesic distance, we train a *Geodesic Autoencoder*, with encoder outputing $\phi : \mathbb{R}^d \to \mathcal{Z}$, such that $\|\phi(x_i) - \phi(x_j)\|_2^2 = G(x_i, x_j)$. We draw a subsample of size N, evaluate G and minimize the mean square error

$$L(\phi) := \frac{2}{N} \sum_{i=1}^{N} \sum_{j>i} \left(||\phi(x_i) - \phi(x_j)||_2 - G(x_i, x_j)\right)^2.$$

Learning the embedding $\mathcal{Z}$ over using G as the advantage of being inductive, it is useful since we use it to compute distances between predicted and ground truth observations. Moreover, the

encoder can be trained to denoise, hence becoming robust to predicted values that are not close to the manifold. Empirically, we observed that G fails to capture meaningful distances for a large number of observations. In practice, we choose N to have few observations in most regions of the manifold, which also has the benefit of reducing the size of the matrix $\mathbf{P}_\epsilon$, thus making the computation of G very efficient.

Theorem 2. *Assuming X is sampled from a closed Riemannian manifold $\mathcal{M}$ with geodesic $d_{\mathcal{M}}$, then for a sufficiently large K, N and small $\epsilon > 0$, we have $G(x_i, x_j) \simeq d^{2\alpha}_{\mathcal{M}}(x_i, x_j)$, for all $\alpha \in (0, 1/2)$.*

Corollary 1. If the encoder is such that $L(\phi) = 0$, then $\|\phi(x_i) - \phi(x_j)\|_2 \simeq d^{2\alpha}_{\mathcal{M}}(x_i, x_j)$ for all $x_i, x_j \in \mathsf{X}$.

From that embedding, we can also train a *decoder* $\phi^{-1} : \mathcal{Z} \to \mathbb{R}^d$, with a reconstruction loss $L_r := \sum_x \|\phi^{-1} \circ \phi(x) - x\|_2$. This is particularly interesting, since we can learn the trajectories in $\mathcal{Z}$, then decode them in the ambient space. For the task of inferring cells trajectories, this enables us to understand how specific genes evolved over time, by decoding in the gene space.

4.2 Inferring Trajectories

Given T distributions $\{\mu_i\}_{i=0}^{T-1}$, we want to model the trajectories respecting the conditions $X_i \sim \mu_i$ for fix timepoints $i \in \{0, \dots, T-1\}$. Formally, we want to learn a parametrized function $f_\theta(x, t)$ such that

$$X_t = X_0 + \int_0^t f_\theta(X_u, u)du, \text{ with } X_0 \sim \mu_0, \dots, X_{T-1} \sim \mu_{T-1}. \tag{5}$$

We adapt the main theorem from [35], to consider the path space, and other types of dissimilarities between distributions.

Theorem 3. *We consider a time-varying vector field $f(x, t)$ defining the trajectories $dX_t = f(X_t, t)dt$ with density ρ_t, and a dissimilarity between distributions such that $D(\mu, \nu) = 0$ iff $\mu = \nu$. Given these assumptions, there exist a sufficiently large $\lambda > 0$ such that*

$$W_d(\mu, \nu)^2 = \inf_{X_t} E\left[\int_0^1 \|f(X_t, t)\|_2^2 dt\right] + \lambda D(\rho_1, \nu) \text{ s.t. } X_0 \sim \mu. \tag{6}$$

Moreover, if X_t defined on the embedded space $\mathcal{Z}$, then W_2 is equivalent to the Wasserstein with geodesic distance $W_2(\mu, \nu) \simeq W_{d^{2\alpha}_{\mathcal{M}}}$.

The previous theorem enables us to add the second marginal constraint in the optimization problem. In practice, it justifies the method of matching the marginals and adding a regularization on the vector field $f(x, t)$, that way the optimal (ρ_t, f) corresponds to the one from the W_2. If in addition we model the trajectories in the embed space $\mathcal{Z}$, then the transport is equivalent to the one on the manifold.

Training We observe discrete distributions $\boldsymbol{\mu}_i := (1/n_i) \sum_i \delta_{x_i}$ for $x_i \in \mathsf{X}_i$, and we approximate (5) with a Neural ODE [6], where f_θ is modeled by a neural network with parameters θ. Denote by $\psi_\theta : \mathbb{R}^d \times \mathcal{T} \to \mathbb{R}^{d|\mathcal{T}|}$ the function that represents the Neural ODE, where $\mathcal{T}$ is a set of time indices. We define the predicted sets $\widehat{\mathsf{X}}_1, \dots, \widehat{\mathsf{X}}_{T-1} = \psi_\theta(\mathsf{X}_0, \{1, \dots, T-1\})$; given an initial set X_0 it returns the approximation of (5) for all $t \in \mathcal{T}$. The resulting discrete distributions are $\hat{\boldsymbol{\mu}}_i := (1/n_i) \sum_i \delta_{x_i}$ for $x_i \in \widehat{\mathsf{X}}_i$. To match the marginals, we used two approaches of training; *local* or *global*. For the local method, we only predict the next sample, hence given X_t as initial condition, we predict $\widehat{\mathsf{X}}_{t+1} = \psi_\theta(\mathsf{X}_t, t+1)$. Whereas for the global, we use the entire trajectories given the initial condition X_0. For both cases, we formulate a loss L_m on the marginals, representing the second term in (6). To take into account the first term in (6), we add the loss L_e, where the integral can be approximate with the forward pass of the ODE solver, and $\lambda_e \geq 0$ is a hyperparameter.

$$L_m := \sum_{i=1}^{T-1} W_2(\hat{\boldsymbol{\mu}}_i, \boldsymbol{\mu}_i) \qquad L_e := \lambda_e \sum_{i=1}^{T-2} \int_i^{i+1} \|f(x_t, t)\|_2^2 dt, \text{ where } x_{t_1} \in \mathsf{X}_{t_1} \tag{7}$$

In practice, to compute the Wasserstein between discrete distributions, we use the implementation from the library Python Optimal Transport [11]. This is in contrasts with the maximum-likelihood approach in CNF methods which requires evaluation of the instantaneous change of variables at every integration timestep at $O(d^2)$ additional cost per function evaluation [6].

We add a final density loss L_d inspired by [35] to encourage the trajectories to stay on the manifold. Intuitively, for a predicted point $x \in \widehat{\mathsf{X}}_t$, we minimize the distance to its k-nearest neighbors in X_t given a lower bound $h > 0$

$$L_d := \sum_{t=1}^{T-1} \sum_{x \in \widehat{\mathsf{X}}_t} \ell_d(x,t), \text{ where } \ell_d(x,t) := \sum_{i=1}^{k} \max(0, \text{min-k}(\{\|x-y\| : y \in \mathsf{X}_t\}) - h).$$

Modeling Diffusion We add a diffusion term in the trajectories, resulting in the SDE $dX_t = f(X_t,t)dt + \sqrt{\sigma_t} dB_t$ In practice, we learn T parameters $\{\sigma_t\}_{t=0}^{T-1}$ It has the advantage of mapping the same cell to different trajectories stochastically as in many naturalistic systems. Early in the training process, we notice that it promotes bifurcation, as it helps to spread the initial X_0. As training continues, we empirically observed that $\sigma \to 0$. Hence, as shown in [23], converging to the W_2 formulation. For small σ_t, we train with a regular ODE solver, essentially corresponding to the Euler–Maruyama method. To model complex diffusion, one would need an SDE solver, thus drastically increasing the computational cost.

5 Results

Here, we validate our method empirically on both synthetic test cases and single-cell RNA sequencing datasets chosen to test the capability of MIOFlow to follow divergent trajectories from population snapshot data. Using these test cases we assess whether MIOFlow-derived trajectories qualitatively follow the data manifold efficiently, we also quantitatively assess their accuracy using the Wasserstein W_1, and Maximum Mean Discrepancy (MMD) [31] on held-out timepoints. We compare our results to those other methods that preform population flows including TrajectoryNet [35] which is based on a CNF that is regularized to achieve efficient paths, and Diffusion Schrödinger's Bridge (DSB) [9] which is an optimal transport framework for generative modeling. Additional experiments such as ablation studies, and details are provided in the supplementary material.

5.1 Artificial Data

Petal The petal dataset is a simple yet challenging , as it mimics natural dynamics arising in cellular differentiation, including bifurcations and merges. The dynamics start in the center (purple) and evolves through the end of the leaf (yellow). In Fig. 2, we present the trajectories on the petal dataset for TrajectoryNet, DSB, and MIOFlow. Only our method is able to learn the dynamic, and keeping trajectories on the manifold. We then hold-out timepoint 2 corresponding to the initial bifurcation, and in Tab. 5.1 we compare the interpolation between the three methods. MIOFlow results in a more accurate entropic interpolation for both metrics, while requiring much less computational time.

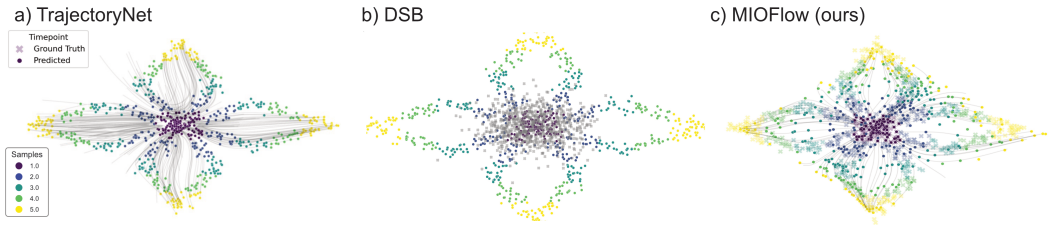

Figure 2: Comparisons between predicted flows of (a) TrajectoryNet, (b) DSB, and (c) MIOFlow on a toy Petal dataset. MIOFlow trajectories both match the data and transition along the data manifold whereas TrajectoryNet takes shortcut paths between the the curved paths and Schroedinger's bridge does not follow the data.

Dyngen We use Dyngen [5] to simulate a scRNA-seq dataset from a dynamical cellular process. We first embed the observations in five dimensions using the non-linear dimensionality reduction method PHATE [25]. We discretize the space in five bins, each representing a differentiation stage. This dataset includes one bifurcation. However, it proves more challenging than the petal dataset. The number of samples per bins is not equal, which tends to give more influence to the lower branch. Moreover, the bifurcation is asymmetric; the curve in the lower branch is less pronounce than the upper one, making it harder to learn a continuous bifurcation. In Fig. 3, we show the trajectories for TrajectoryNet, DSB, and our method. MIOFlow is more accurate as it learns continuous bifurcating trajectories along the two main branches. In Tab. 5.1, we hold-out timepoint 2, and show the interpolation results against the ground truth. MIOFlow is more accurate in terms of EMD, and is similar to TrajectoryNet for the MMD, but requiring much less computational time. However, we see that in timepoint 1, TrajectoryNet deviates far from the trajectoriesm particularly at the branching point where it struggles with induced heterogeneity (see Figure 3). We have further quantifications reflecting this in the supplementary material.

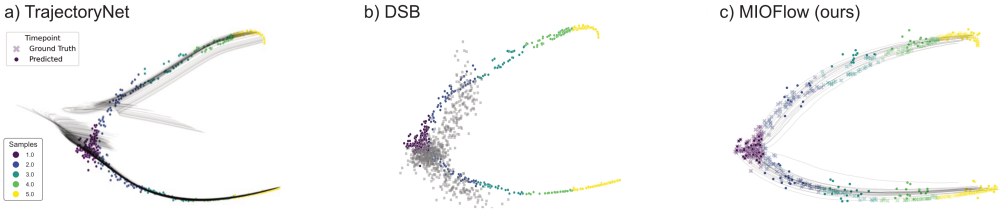

Figure 3: Comparisons between predicted flows of (a) TrajectoryNet, (b) DSB, and (c) MIOFlow on Dyngen [5] simulated dataset. MIOFlow trajectories both match the data and transition along the data manifold throughout the timepoints.

Table 1: Leave-one-out W_1 and MMD between the predicted and ground truth points as well as runtime in (m)inutes and (s)econds on the Petal and Dyngen datasets of TrajectoryNet, DSB. and MIOFlow. Lower is better.

	Petal			Dyngen		
	W_1	MMD	Runtime	W_1	MMD	Runtime
TrajectoryNet	1.697	9.513	64 (m)	1.809	**7.320**	62 (m)
DSB	2.923	9.579	86 (m)	2.251	21.454	81 (m)
MIOFlow (ours)	**1.605**	**9.345**	**284.60 (s)**	**0.756**	7.674	**95.63 (s)**

5.2 Single-Cell Data

Embryoid Body Here we evaluate our method on scRNA-seq data from the dynamic time course of human embryoid body (EB) differentiation over a period of 27 days, measured in 4 timepoints between days 0-3, 6-9, 12-15, 18-21, and 24-27. We reduce the dimensionality of this dataset to 200 dimensions using PCA. See the supplementary material for further details on the preprocessing steps and the parameters.

By combining PCA and the autoencoder with diffusion geodesic distance described in section 4.1, we can decode the trajectories back into the gene space (17846 dimensions), thus describing the evolution of the gene expression of a cell. In Fig. 5 we plot the trajectories, at a gene level of 20 neuronal cells against 20 random one. We generally observe distinct differences between TrajectoryNet and MIOFlow. TrajectoryNet (which is shown in the bottom) shows a tendency to have early divergence in trends artificially showing higher heterogeneity in gene trends then exist in neuronal specification processes. Examples include HAND2, which has a non-monotonic trend, first increasing in expression during development from embryonic stem cells to neural early progenitors and then decreasing as the lineages mature [21]. MIOFlow clearly shows this trend whereas TrajectoryNet trends are more chaotic. Another example is ONECUT2 which is consistently known to be higher in cells that diverge towards the neuronal lineage [39], Other examples are discussed in the supplementary material.

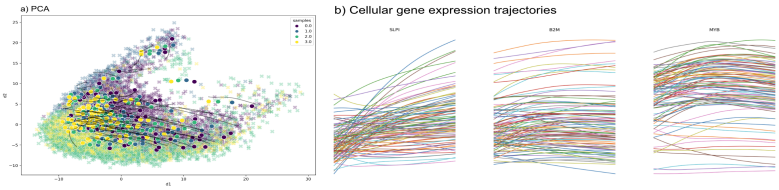

Figure 4: MIOFlow on the AML dataset. a) PCA Embedding comparing ground truth to MIOFlow predicted points and trajectories. b) Cellular gene expression trajectories for *SLPI*, *B2M*, and *MYB* for MIOFlow.

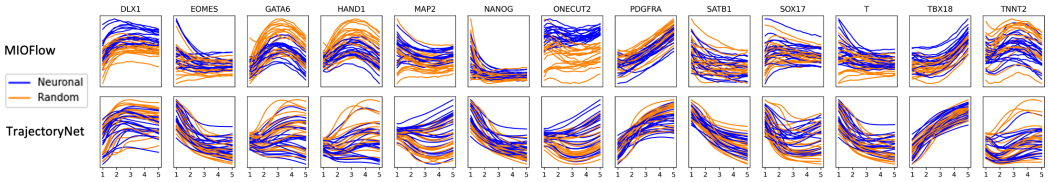

Figure 5: Neuronal gene trajectories vs. random trajectories for MIOFlow (top) vs. TrajectoryNet [36] (bottom). TrajectoryNet consistently shows a tendency for early divergence in trajectories, examples include ONECUT2 which is biologically known to be higher expressed [25] and HAND2 which is known to peak in neural early progenitors (NEPs).

AML Cancer Data with Treatment-based Progression We also assessed whether MIOFlow can give insights on treatment resistant cancer cell populations, rather than differentiation. For this purpose, we used data from Fennell et al. [10] in which AF9 mutant cancer cells implanted into the bone marrow of mice as a model of Acute Myeloid Leukemia (AML). The mice are then treated for 7 days with intensive chemotherapy with single cell measurements taken at day 2, day 5, and day 7. Here we show MIOFlow on this dataset along with signatures of leukemic stem cells (LSCs) which are hypothesized to be dominant in chemotherapy resistant cells. We see heterogeneity in the trajectories, the majority of cells surviving to the last timepoint showing an increase in the markers SLPI, MYB, and B2M—all members of the LSC pathway [10]. This indicates that not all cells start with this signature but are made to transition to this signature to resist the effects of chemotherapy.

6 Conclusion

We proposed MIOFlow a computationally efficient method to recover trajectories from multiple cross-sectional samples of a time-evolving population. Our method includes a novel manifold learning framework, where we learn an embedding preserving the geodesic distance. We show theoretically how MIOFlow is related to dynamical optimal transport on an underlying manifold, and it's capacity to model stochastic trajectories. On artificial datasets, MIOFlow produced continuous trajectories more in line with the manifold compare to DSB and TrajectoryNet. On scRNA-seq data, we leveraged an autoencoder with the geodesic embedding to learn the trajectories in a lower dimensional space while preserving meaningful distances which can then be decoded in gene space. This way we could visualize the evolution of the gene expression of a cell.

7 Limitations and Broader Impact

Our method does have some limitations. Similarly to other works based on continuous time networks, our work makes use of generic ODE solvers and adjoint dynamics which can fail under stiff dynamics. Like other methods that learn dynamics it is possible that the learned dynamics may become stiff through training leading to decreased model accuracy. In addition, our work is sensitive to the time-bins, and it worked best for relatively uniform sampling. This second issue could potentially be alleviated by using unbalanced optimal transport, which allows transporting distributions with unequal masses. Our work can help understand dynamical systems, and in particular cellular differentiation, and thus to the best of our knowledge presents no potential negative societal impacts.

References

[1] Andrei Agrachev and Paul WY Lee. Continuity of optimal control costs and its application to weak kam theory. *Calculus of Variations and Partial Differential Equations*, 39(1):213–232, 2010.

[2] Jean-David Benamou and Yann Brenier. A computational fluid mechanics solution to the Monge-Kantorovich mass transfer problem. *Numerische Mathematik*, 84(3):375–393, 2000.

[3] Volker Bergen, Ruslan A. Soldatov, Peter V. Kharchenko, and Fabian J. Theis. RNA velocity—current challenges and future perspectives. *Molecular Systems Biology*, 17(8):e10282, 2021.

[4] Charlotte Bunne, Laetitia Meng-Papaxanthos, Andreas Krause, and Marco Cuturi. Proximal Optimal Transport Modeling of Population Dynamics. *AISTATS*, 2022.

[5] Robrecht Cannoodt, Wouter Saelens, Louise Deconinck, and Yvan Saeys. Spearheading future omics analyses using dyngen, a multi-modal simulator of single cells. *Nature Communications*, 12(1):1–9, 2021.

[6] Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David Duvenaud. Neural Ordinary Differential Equations. *arXiv:1806.07366 [cs, stat]*, 2019.

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

[8] Ronald R. Coifman and Stéphane Lafon. Diffusion maps. *Applied and Computational Harmonic Analysis*, 21(1):5–30, 2006.

[9] Valentin De Bortoli, James Thornton, Jeremy Heng, and Arnaud Doucet. Diffusion schrödinger bridge with applications to score-based generative modeling. *Advances in Neural Information Processing Systems*, 34, 2021.

[10] Katie A Fennell, Dane Vassiliadis, Enid YN Lam, Luciano G Martelotto, Jesse J Balic, Sebastian Hollizeck, Tom S Weber, Timothy Semple, Qing Wang, Denise C Miles, et al. Non-genetic determinants of malignant clonal fitness at single-cell resolution. *Nature*, 601(7891):125–131, 2022.

[11] Rémi Flamary, Nicolas Courty, Alexandre Gramfort, Mokhtar Z. Alaya, Aurélie Boisbunon, Stanislas Chambon, Laetitia Chapel, Adrien Corenflos, Kilian Fatras, Nemo Fournier, Léo Gautheron, Nathalie T.H. Gayraud, Hicham Janati, Alain Rakotomamonjy, Ievgen Redko, Antoine Rolet, Antony Schutz, Vivien Seguy, Danica J. Sutherland, Romain Tavenard, Alexander Tong, and Titouan Vayer. Pot: Python optimal transport. *Journal of Machine Learning Research*, 22(78):1–8, 2021.

[12] Crispin W Gardiner et al. *Handbook of stochastic methods*, volume 3. springer Berlin, 1985.

[13] Nassif Ghoussoub, Young-Heon Kim, and Aaron Zeff Palmer. A solution to the monge transport problem for brownian martingales. *the Annals of Probability*, 49(2):877–907, 2021.

[14] Will Grathwohl, Ricky T. Q. Chen, Jesse Bettencourt, Ilya Sutskever, and David Duvenaud. FFJORD: Free-form Continuous Dynamics for Scalable Reversible Generative Models. In *7th International Conference on Learning Representations*, 2019.

[15] Laleh Haghverdi, Maren Büttner, F. Alexander Wolf, Florian Buettner, and Fabian J. Theis. Diffusion pseudotime robustly reconstructs lineage branching. *Nat Methods*, 13(10):845–848, 2016.

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

[17] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising Diffusion Probabilistic Models. *NeurIPS*, 2020.

[18] Gioele La Manno, Ruslan Soldatov, Amit Zeisel, Emelie Braun, Hannah Hochgerner, Viktor Petukhov, Katja Lidschreiber, Maria E. Kastriti, Peter Lönnerberg, Alessandro Furlan, Jean Fan, Lars E. Borm, Zehua Liu, David van Bruggen, Jimin Guo, Xiaoling He, Roger Barker, Erik Sundström, Gonçalo Castelo-Branco, Patrick Cramer, Igor Adameyko, Sten Linnarsson, and Peter V. Kharchenko. RNA velocity of single cells. *Nature*, 560(7719):494–498, 2018.

[19] Jean-Michel Lasry and Pierre-Louis Lions. Mean field games. *Japanese journal of mathematics*, 2(1): 229–260, 2007.

[20] William Leeb and Ronald Coifman. Hölder–Lipschitz Norms and Their Duals on Spaces with Semigroups, with Applications to Earth Mover's Distance. *J Fourier Anal Appl*, 22(4):910–953, 2016.

[21] Jun Lei and Marthe J Howard. Targeted deletion of hand2 in enteric neural precursor cells affects its functions in neurogenesis, neurotransmitter specification and gangliogenesis, causing functional aganglionosis. *Development*, 138(21):4789–4800, 2011.

[22] Emile Mathieu and Maximilian Nickel. Riemannian continuous normalizing flows. *Advances in Neural Information Processing Systems*, 33:2503–2515, 2020.

[23] Toshio Mikami. Monge's problem with a quadratic cost by the zero-noise limit of h-path processes. *Probability theory and related fields*, 129(2):245–260, 2004.

[24] Toshio Mikami and Michèle Thieullen. Duality theorem for the stochastic optimal control problem. *Stochastic processes and their applications*, 116(12):1815–1835, 2006.

[25] Kevin R. Moon, David van Dijk, Zheng Wang, Scott Gigante, Daniel B. Burkhardt, William S. Chen, Kristina Yim, Antonia van den Elzen, Matthew J. Hirn, Ronald R. Coifman, Natalia B. Ivanova, Guy Wolf, and Smita Krishnaswamy. Visualizing structure and transitions in high-dimensional biological data. *Nat Biotechnol*, 37(12):1482–1492, 2019.

[26] Michele Pavon, Giulio Trigila, and Esteban G Tabak. The data-driven schrödinger bridge. *Communications on Pure and Applied Mathematics*, 74(7):1545–1573, 2021.

[27] Gabriel Peyré and Marco Cuturi. Computational Optimal Transport. *arXiv:1803.00567 [stat]*, 2020.

[28] Neha Prasad, Karren Yang, and Caroline Uhler. Optimal transport using gans for lineage tracing. *arXiv preprint arXiv:2007.12098*, 2020.

[29] Wouter Saelens, Robrecht Cannoodt, Helena Todorov, and Yvan Saeys. A comparison of single-cell trajectory inference methods. *Nature biotechnology*, 37(5):547–554, 2019.

[30] Geoffrey Schiebinger, Jian Shu, Marcin Tabaka, Brian Cleary, Vidya Subramanian, Aryeh Solomon, Siyan Liu, Stacie Lin, Peter Berube, Lia Lee, Jenny Chen, Justin Brumbaugh, Philippe Rigollet, Konrad Hochedlinger, Rudolf Jaenisch, Aviv Regev, and Eric S. Lander. Reconstruction of developmental landscapes by optimal-transport analysis of single-cell gene expression sheds light on cellular reprogramming. *Cell*, 2019.

[31] Alexander J Smola, A Gretton, and K Borgwardt. Maximum mean discrepancy. In *13th International Conference, ICONIP 2006, Hong Kong, China, October 3-6, 2006: Proceedings*, 2006.

[32] Yang Song and Stefano Ermon. Generative Modeling by Estimating Gradients of the Data Distribution. *NeurIPS*, 2019.

[33] Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-Based Generative Modeling through Stochastic Differential Equations. *arXiv:2011.13456 [cs, stat]*, 2021.

[34] Kelly Street, Davide Risso, Russell B Fletcher, Diya Das, John Ngai, Nir Yosef, Elizabeth Purdom, and Sandrine Dudoit. Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. *BMC genomics*, 19(1):1–16, 2018.

[35] Alexander Tong, Jessie Huang, Guy Wolf, David Van Dijk, and Smita Krishnaswamy. TrajectoryNet: A Dynamic Optimal Transport Network for Modeling Cellular Dynamics. In *Proceedings of the 37th International Conference on Machine Learning*, pages 9526–9536. PMLR, 2020.

[36] Alexander Tong, Guillaume Huguet, Amine Natik, Kincaid MacDonald, Manik Kuchroo, Ronald Coifman, Guy Wolf, and Smita Krishnaswamy. Diffusion Earth Mover's Distance and Distribution Embeddings. *arXiv:2102.12833 [cs]*, 2021.

[37] Alexander Y Tong, Guillaume Huguet, Amine Natik, Kincaid MacDonald, Manik Kuchroo, Ronald Coifman, Guy Wolf, and Smita Krishnaswamy. Diffusion earth mover's distance and distribution embeddings. In *International Conference on Machine Learning*, pages 10336–10346. PMLR, 2021.

[38] Cole Trapnell, Davide Cacchiarelli, Jonna Grimsby, Prapti Pokharel, Shuqiang Li, Michael Morse, Niall J Lennon, Kenneth J Livak, Tarjei S Mikkelsen, and John L Rinn. Pseudo-temporal ordering of individual cells reveals dynamics and regulators of cell fate decisions. *Nature biotechnology*, 32(4):381, 2014.

[39] Jori van der Raadt, Sebastianus HC van Gestel, Nael Nadif Kasri, and Cornelis A Albers. Onecut transcription factors induce neuronal characteristics and remodel chromatin accessibility. *Nucleic acids research*, 47(11):5587–5602, 2019.

[40] Cédric Villani. *Optimal transport: old and new*, volume 338. Springer, 2009.

[41] Stephen Zhang, Anton Afanassiev, Laura Greenstreet, Tetsuya Matsumoto, and Geoffrey Schiebinger. Optimal transport analysis reveals trajectories in steady-state systems. Technical report, 2021.

[42] Ziqi Zhang and Xiuwei Zhang. Inference of high-resolution trajectories in single-cell RNA-seq data by using RNA velocity. *Cell Reports Methods*, 1(6):100095, 2021.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] In the limitations and broader impact section.
   (c) Did you discuss any potential negative societal impacts of your work?
   [Yes] In the limitations and broader impact section. We do not foresee any immediate negative societal impacts of this work.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results?
   [Yes] We stated all assumptions in the main body of the text and provide additional details in the supplementary materials.
   (b) Did you include complete proofs of all theoretical results?
   [Yes] Complete proofs are available in the supplement.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)?
   [Yes] Code and links to publicly available data will be provided in the supplemental material.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)?
   [Yes] All training details are provided in the supplement.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)?
   [Yes] Compute resources are discussed in the supplement.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [Yes] See supplement
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content?
   [Yes] Our data does not contain any personally identifiable information or offensive content. The Embryoid body data is a cell culture, and the AML Cancer dataset is patient derived but contains no identifiable information.

5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]