Optimal Transport using GANs for Lineage Tracing

Neha Prasad [1] [*]   **Karren D. Yang** [1] [*]   **Caroline Uhler** [1]

Abstract

In this paper, we present Super-OT, a novel approach to computational lineage tracing that combines a supervised learning framework with optimal transport based on Generative Adversarial Networks (GANs). Unlike previous approaches to lineage tracing, Super-OT has the flexibility to integrate paired data. We benchmark Super-OT based on single-cell RNA-seq data against Waddington-OT, a popular approach for lineage tracing that also employs optimal transport. We show that Super-OT achieves gains over Waddington-OT in predicting the class outcome of cells during differentiation, since it allows the integration of additional information during training.

1. Introduction

A major goal of developmental biology is understanding the gene programs that drive the differentation of progenitor cells into mature cells. Although single-cell RNA sequencing technologies enable collecting large quantities of single-cell gene expression data, these experiments tend to be destructive to samples: most experiments only take a snapshot of a single cell lineage at one time point. This makes it challenging to trace the lineage of individual cells and necessitates the development of novel computational and experimental strategies to follow individual cell trajectories backwards and forwards in time.

In recent years, various computational methods have been developed to trace pseudo-lineages of cells given single-cell gene expression data (Trapnell et al., 2014; Qiu et al., 2017; Guo et al., 2017; Shin et al., 2015; Ji & Ji, 2016; Marco et al., 2014; Bendall et al., 2014; Setty et al., 2016; Wolf et al., 2019; Weinreb et al., 2018; Schiebinger et al., 2019). All these methods are unsupervised and depend heavily on the assumption that cells close in gene expres-

sion space are more likely to belong to the same lineage, which is often not sufficient for accurately predicting cell fate decisions. For example, Weinreb et al. (2020) recently developed an experimental assay to track families of cells (i.e., cell clones) during hematopoiesis. Their work revealed a considerable gap between the pseudo-lineages that are reconstructed computationally and real cell trajectories. In light of recent experimental methods for providing partial information about single-cell lineages, there is a need for computational methods to integrate these types of data with existing unsupervised lineage tracing frameworks.

In this work, we develop to our knowledge the first model for computational lineage tracing that can integrate additional information such as the clonal data of Weinreb et al. (2020). Since their data does not provide full information about single cell lineages (e.g., due to multiple cells from the same clone that cannot be disambiguated), our work combines a supervised learning framework with the principles of optimal transport to predict individual cell lineages. Optimal transport has emerged as a powerful method for learning couplings between cells and has many advantages in computational lineage tracing compared to other methods (Schiebinger et al., 2019). The best-known computational approach for lineage tracing based on optimal transport, Waddington-OT, uses an iterative scaling algorithm and cannot be adapted easily to integrate additional labeled data (Schiebinger et al., 2019). Recent approaches for performing optimal transport based on GANs have also emerged (Yang & Uhler, 2019), but these have not been rigorously compared against Waddington-OT and were also not adapted to integrate additional labeled data.

Contribution. We propose a new framework for computational lineage tracing, which we name *Super-OT*, that combines a *super*vised learning framework with optimal transport (*OT*) based on GANs. We apply this framework to perform lineage tracing on the dataset of Weinreb et al. (2020). While Waddington-OT marginally outperforms GAN-based optimal transport in the completely unsupervised setting, we find that Super-OT outperforms Waddington-OT in predicting the class outcome of cells during differentiation by integrating additional information during training. We conclude that GAN-based optimal transport is a practical approach for integrating novel types of experimental data in lineage tracing with a principled computational approach.

---

[*]Equal contribution [1]Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, Cambridge, MA, USA. Correspondence to: Caroline Uhler <cuhler@mit.edu>.

2. Method

2.1. Problem Formulation.

Let P_t and P_{t+1} denote cell distributions over the gene expression space $\mathbb{R}^d$ at times t and $t + 1$, respectively. We formulate computational lineage tracing as the problem of learning a transport map $T : \mathbb{R}^d \to \mathbb{R}^d$ that pushes P_t to P_{t+1}, i.e., if $X_t \sim P_t$ and $X_{t+1} \sim P_{t+1}$, then $T(X_t) \sim P_{t+1}$. We assume that we are given samples from P_t, P_{t+1} and their joint distribution $P_{t,t+1}$. We denote the observed empirical distributions by $\hat{P}_t$, $\hat{P}_{t+1}$ and $\hat{P}_{t,t+1}$.

2.2. Optimal Transport using GANs.

There are many possible maps satisfying $T(X_t) \sim P_{t+1}$. One way to constrain the solution space is to solve the Monge optimal transport problem: find the map that minimizes some transportation cost $c : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^+$, i.e.,

$$\min_T \quad \mathbb{E}_{X_t \sim P_t} c(X_{t+1}, T(X_t)) \qquad (1)$$

subject to $T(X_t) \sim P_{t+1}$. In the context of lineage tracing, c is often chosen to be the Euclidean distance $c(x, y) = ||x - y||_2$, which encourages the map to match cells that are closer together in the gene expression space. The intuition is that cells closer in gene expression space should be more likely to belong to the same lineage. Alternatively, the cost can also be computed in a feature space learned by an autoencoder, as proposed by Yang et al. (2020).

In practice, Equation (1) is challenging to optimize due to the constraint $T(X_t) \sim P_{t+1}$. We follow Yang & Uhler (2019) and relax the hard constraint using a divergence that can be optimized using an adversarial approach. Specifically, we instead consider solving

$$\min_T \quad \lambda_1 \mathcal{L}_{TRANS} + \mathcal{L}_{GAN}, \qquad (2)$$

where

$$\mathcal{L}_{TRANS} = \mathbb{E}_{X_t \sim P_t} c(X_{t+1}, T(X_t)), \quad \text{and}$$
$$\mathcal{L}_{GAN} = \max_{D:\ \mathbb{R}^d \to (0,1)} \mathbb{E}_{X_{t+1} \sim P_{t+1}} [\log D(X_{t+1})]$$
$$+ \mathbb{E}_{X_t \sim P_t} [\log(1 - D(T(X_t)))],$$

and $\lambda_1 > 0$ is a hyperparameter. In practice, we can parameterize T, D using neural networks and minimize the loss with respect to the observed empirical distributions.

2.3. Super-OT

Existing methods for lineage tracing based on optimal transport do not leverage labeled information. Our approach extends the GAN-based optimal transport framework to pair cells between two distinct time points that belong to the same clonal family. We consider an additional loss

$$\mathcal{L}_{SUPER} = \mathbb{E}_{(X_t, X_{t+1}) \sim \hat{P}_{t,t+1}} ||X_t - X_{t+1}||_2^2, \quad (3)$$

which ensures that points belonging to the same clonal family are mapped to each other. The final objective then becomes

$$\min_T \quad \lambda_1 \mathcal{L}_{TRANS} + \mathcal{L}_{GAN} + \lambda_2 \mathcal{L}_{SUPER}, \qquad (4)$$

where $\lambda_1, \lambda_2 > 0$ are hyperparameters.

3. Experiments

3.1. Dataset

We use the single-cell gene expression dataset of mouse bone marrow cells from Weinreb et al. (2020). The dataset contains 130,887 samples across 25,289 genes, sampled from three separate time points (days 2, 4, and 6). We let P_t corresponds to undifferentiated day 2 cells, while we let P_{t+1} correspond to differentiated days 4/6 cells. For our experiments, we consider a subset of the data containing neutrophils, monocytes, and their progenitors. This simplified the dataset to contain 1,527 day 2 cells and 39,401 day 4/6 neutrophils and monocytes. Cell family information is provided in the form of a binary matrix in which rows represent cells and columns represent clones, where an entry of 1 indicates that a cell belongs to a particular clone. For the supervised labeling of the data, we consider a day 2 cell to be paired with a day 4/6 cell if they belong to the same cell family and the day 4/6 cell is in the majority class (neutrophil versus monocyte) for that family.

3.2. Model and Training.

We implemented our model in PyTorch (Paszke et al., 2019). The transport T uses four linear layers with hidden dimension of 1000 with ReLU activations. The discriminator D also consists of four linear layers with batch normalization and ReLU activations. The dataset reports the number of transcripts (UMIs) for each gene in each cell, after total counts normalization (i.e., L1 normalization on cells). We add an additional layer of preprocessing by scaling each feature (gene) between the range $[0, 1]$ with L2 normalization. We also apply PCA (Pedregosa et al., 2011) to reduce the dimensionality of the gene expression vectors to 100. Note that this reduced dimensionality justifies using Euclidean distance for $\mathcal{L}_{TRANS}$. The networks are trained jointly using the loss in Equation 4 with different numbers of paired samples. For the transport cost, we used $\lambda_1 = 0.6$, because we found that $\lambda_1 > 0.6$ would cause the model to not converge. Similar to other GANs (Gulrajani et al., 2017), we also include a gradient penalty on the discriminator to improve training stability. Parameters are optimized using Adam (Kingma & Ba, 2014) with learning rates set to 0.0001 for each model. We run the model for a maximum of 100,000 epochs, or until the loss has sufficiently converged. We train each setting with 80% of the cells, and leave the remaining 20% for evaluation.

We evaluate the performance of Super-OT with different numbers of labeled pairs of data: 300, 600, and 900.

3.3. Benchmarks

We compare Super-OT against the following benchmarks:

- Waddington-OT (Schiebinger et al., 2019), which we denote by WOT in Table 1, a lineage tracing method that solves the optimal transport problem using an iterative scaling algorithm.

- Conditional GAN (Mirza & Osindero, 2014): A model that generates the day 4/6 cells from day 2 conditioned on class (monocyte/neutrophil).

- GAN-based Optimal Transport: Conditional GAN combined with transport cost.

- Supervised: A regression model that maps each day 2 cell to a corresponding day 4/6 cell in its clonal family. This method uses all paired information available and should be considered an upper bound on performance.

3.4. Evaluation Criteria.

We evaluate each model on its class prediction accuracy in transporting day 2 cells to day 4/6 monocytes versus neutrophils. Specifically, from the dataset clonal information, we determine whether each day 2 cell is more likely to become a neutrophil or a monocyte and assign this label to the cell. We train a logistic regression model (Pedregosa et al., 2011) to classify between real day 4/6 monocytes and neutrophils, and obtain the predicted class for each transported day 2 cell. We then compare these predicted results with the assigned labels. The accuracy reflects the ability of the models to correctly predict the direction of the trajectory of the day 2 test cells towards their differentiated class.

Waddington-OT differs from the other models in that it produces a probabilistic coupling between cells: each day 2 cell is assigned to a distribution over the day 4/6 cells (Schiebinger et al., 2019). To obtain a class prediction for each day 2 cell, we assign a label depending on whether the majority of the distribution is assigned to day 4/6 monocytes or neutrophils. For fairer comparison with the other models, we alternatively assign a label depending on whether the majority of the distribution is assigned to predicted monocytes or neutrophils using a logistic regression model. This check ensures that differences in performance are not due to the logistic classifier.

3.5. Results

We report the class prediction accuracy of the different models in Table 1. For Super-OT, we report results with different numbers of paired training points (e.g., 300, 600, 900.)

Table 1. Comparison of prediction accuracy of different models. Higher is better. For each deep learning model, we perform three separate runs over different train/test splits and report the average of the best results on the held-out data.

SETTING	ACCURACY
WOT: PREDICTED LABELS	0.6535
WOT: REAL LABELS	0.6531
CONDITIONAL GAN	0.5982
GAN-BASED OT	0.6219
SUPER-OT: 300	0.6731
SUPER-OT: 600	0.6856
SUPER-OT: 900	**0.7188**
SUPERVISED	0.7534

Table 2. Comparison of prediction accuracy for Super-OT: 300, Super-OT: 600, Super-OT: 900 with and without the transport cost.

SETTING	TRANSPORT COST	NO TRANSPORT COST
SUPER-OT: 300	0.6731	0.6663
SUPER-OT: 600	0.6856	0.6810
SUPER-OT: 900	0.7188	0.6997

Out of the unsupervised baseline methods, Waddington-OT outperforms the conditional GAN and GAN-based optimal transport. This is expected because Waddington-OT is based on an iterative scaling algorithm with convergence guarantees, while deep models solve non-convex optimization problems and do not have such guarantees. However, GAN-based optimal transport is flexible and can be adapted to handle additional losses, such as our paired data supervised loss. Super-OT, which is our extension of GAN-based optimal transport with some labeled data, outperforms all the baselines including Waddington-OT. Performance increases as more labeled data is added and is upper bounded by the fully supervised model.

Ablations. Since Super-OT integrates supervised training with optimal transport, we investigate the importance of each component of the loss function.

- Supervised Loss ($\mathcal{L}_{SUPER}$): As shown in Table 1, decreasing the number of labeled points decreases the accuracy of the method.

- Transport cost ($\mathcal{L}_{TRANS}$): As shown in Table 2, removing the transport cost decreases the accuracy of the method for any given number of labeled points.

Visualization. We visualize the day 2, day 4/6 and transported test cells for our Super-OT model using t-SNE (Pedregosa et al., 2011) in Figure 1. We see that the model is separating the Neutrophil/Monocyte distributions correctly,
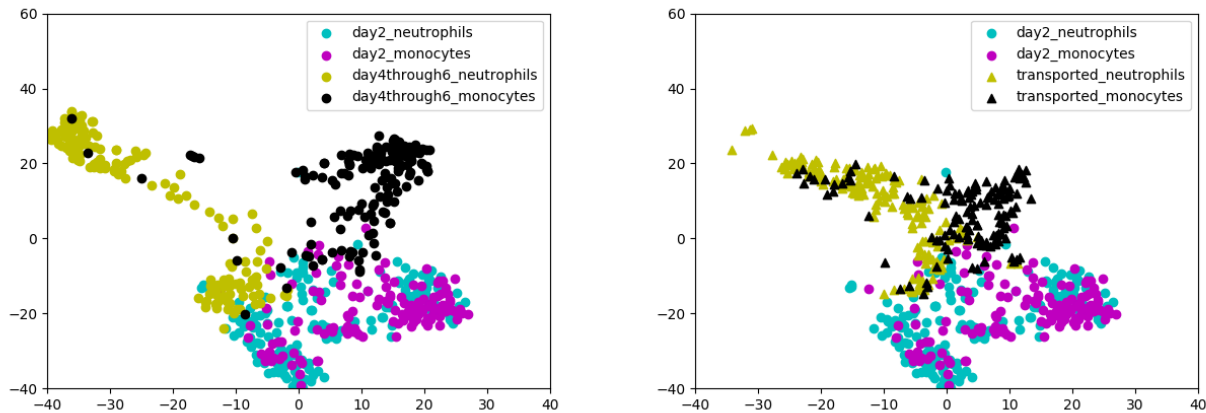
Figure 1. Super-OT: 600 with Transport Cost, Actual Neutrophil/Monocyte Distributions (left), Transported Neutrophil/Monocyte Distributions (right). The plot shows how day 2 neutrophils (blue) and monocytes (pink) transform into differentiated neutrophils (yellow) and monocytes (black).

and that there is overlap between the transported cell distribution and the ground truth. However, the transported cell clusters are closer to each other than compared to the real cell clusters. One explanation is that differentiation is not yet deterministic on day 2: despite overlapping in gene expression space, some day 2 cells become monocytes and others become neutrophils. Therefore the model transports cells to a weighted average between the two clusters. In future work, incorporating stochasticity into the transport map may improve separation of the transported clusters.

3.6. Differential Gene Analysis

To determine how faithful Super-OT is for modeling the molecular programs that drive differentiation to monocytes or neutrophils, we performed a differential gene expression analysis to determine which genes are predicted to be the most differentially expressed between the neutrophil and monocyte distributions. We then compared these gene programs to the actual genes that are differentially expressed between the distributions. Specifically, we used the networks corresponding to the best accuracy values in Table 1. Note that after preprocessing, all of these cells have dimension 100, and so we used the inverse PCA transform to revert cells to their original 25289-dimensional gene space. From here, we iterated through each of the genes and performed a t-test to determine the differentially-expressed genes, setting our p-value threshold at 10^{-6}.

Finally, we selected the genes that were common among all three runs for our analysis. Table 3 lists the precision and recalls for the different settings. We found that there was considerable overlap between the predicted and actual differentially expressed genes. The precision / recall values

Table 3. Differential Gene Expression Analysis: Comparison between the precision / recall values for GAN-Based OT, Super-OT: 300, Super-OT: 600, and Super-OT: 900

SETTING	# OF GENES	PRECISION	RECALL
GAN-BASED OT	12	1.0	0.006726
SUPER-OT: 300	3084	0.4212	0.6969
SUPER-OT: 600	5041	0.3210	0.8662
SUPER-OT: 900	3021	0.3923	0.6096

were robust to different p-values. Importantly, the differentially expressed genes found by Super-OT were much more stable than those found using GAN-based OT.

4. Discussion

In this work, we proposed Super-OT, a novel computational lineage tracing method that combines a supervised learning framework with GAN-based optimal transport. Our framework can easily be extended to map cells between multiple time points, in which we use the timepoint as extra conditioning information. Super-OT achieves gains over Waddington-OT in predicting the class outcome of cells during differentiation by integrating additional information during training. In future work, it would be interesting to analyze how our framework performs when a scaling factor is incorporated to model growth (replication) or shrinkage (death) of cells as proposed by Yang & Uhler (2019). In addition, it would be interesting to investigate how one could incorporate stochasticity into Super-OT, instead of using a deterministic map, to achieve better cluster separation.

Acknowledgments

References

Bendall, S. C., Davis, K. L., Amir, E.-a. D., Tadmor, M. D., Simonds, E. F., Chen, T. J., Shenfeld, D. K., Nolan, G. P., and Pe'er, D. Single-cell trajectory detection uncovers progression and regulatory coordination in human B cell development. *Cell*, 157(3):714–725, 2014.

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., and Courville, A. Improved training of Wasserstein GANs, 2017.

Guo, M., Bao, E. L., Wagner, M., Whitsett, J. A., and Xu, Y. Slice: determining cell differentiation and lineage based on single cell entropy. *Nucleic Acids Research*, 45(7): e54–e54, 2017.

Ji, Z. and Ji, H. Tscan: Pseudo-time reconstruction and evaluation in single-cell RNA-seq analysis. *Nucleic Acids research*, 44(13):e117–e117, 2016.

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

Marco, E., Karp, R. L., Guo, G., Robson, P., Hart, A. H., Trippa, L., and Yuan, G.-C. Bifurcation analysis of single-cell gene expression data reveals epigenetic landscape. *Proceedings of the National Academy of Sciences*, 111 (52):E5643–E5650, 2014.

Mirza, M. and Osindero, S. Conditional generative adversarial nets, 2014.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. Pytorch: An imperative style, high-performance deep learning library. In Wallach, H., Larochelle, H., Beygelzimer, A., d Álché-Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 32*, pp. 8024–8035. Curran Associates, Inc., 2019.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E.

Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Qiu, X., Mao, Q., Tang, Y., Wang, L., Chawla, R., Pliner, H. A., and Trapnell, C. Reversed graph embedding resolves complex single-cell trajectories. *Nature Methods*, 14(10):979, 2017.

Schiebinger, G., Shu, J., Tabaka, M., Cleary, B., Subramanian, V., Solomon, A., Liu, S., Lin, S., Berube, P., Lee, L., et al. Reconstruction of developmental landscapes by optimal-transport analysis of single-cell gene expression sheds light on cellular reprogramming. *Cell*, 176(4): 928–943.e22, 2019.

Setty, M., Tadmor, M. D., Reich-Zeliger, S., Angel, O., Salame, T. M., Kathail, P., Choi, K., Bendall, S., Friedman, N., and Pe'er, D. Wishbone identifies bifurcating developmental trajectories from single-cell data. *Nature Biotechnology*, 34(6):637, 2016.

Shin, J., Berg, D. A., Zhu, Y., Shin, J. Y., Song, J., Bonaguidi, M. A., Enikolopov, G., Nauen, D. W., Christian, K. M., Ming, G.-l., et al. Single-cell RNA-seq with waterfall reveals molecular cascades underlying adult neurogenesis. *Cell Stem Cell*, 17(3):360–372, 2015.

Trapnell, C., Cacchiarelli, D., Grimsby, J., Pokharel, P., Li, S., Morse, M., Lennon, N. J., Livak, K. J., Mikkelsen, T. S., and Rinn, J. L. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nature Biotechnology*, 32(4):381, 2014.

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

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

Wolf, F. A., Hamey, F. K., Plass, M., Solana, J., Dahlin, J. S., Göttgens, B., Rajewsky, N., Simon, L., and Theis, F. J. Paga: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. *Genome Biology*, 20(1):59, 2019.

Yang, K. D. and Uhler, C. Scalable unbalanced optimal transport using generative adversarial networks. *International Conference on Learning Representations*, 2019.

Yang, K. D., Damodaran, K., Venkatchalapathy, S., Soylemezoglu, A. C., Shivashankar, G., and Uhler, C. Autoencoder and optimal transport to infer single-cell trajectories of biological processes. *PLoS Computational Biology*, 16:e1007828, 2020.