NEURAL ADAPTIVE OPTIMAL TRANSPORT: DUALITY, BRENIER POTENTIALS AND APPLICATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Classical optimal transport is a cornerstone of modern machine learning, yet its rigid requirement for full-mass or fixed-mass conservation severely limits its robustness against outliers and distributional shifts. While recent discrete elastic optimal transport enables adaptive mass transport via inequality constraints and mixed-sign costs, it lacks a continuous theoretical foundation, leaving critical questions on duality, Brenier potentials, and high-dimensional solvability unanswered. To bridge this gap, we propose Neural Adaptive Optimal Transport (NAOT), a novel framework that unifies discrete elastic optimal transport with continuous optimal transport theory. Building upon Brenier's seminal theory, we derive the duality theory and Brenier-type theorems, which shed light on designing an efficient and scalable algorithm based on input convex neural networks to find transport map. NAOT outperforms the state-of-the-art methods at predicting single-cell perturbation responses on the benchmark dataset, demonstrating its superiority on adaptive distribution alignment. Our work provides the first continuous framework of adaptive optimal transport, offering both theoretical insights and a scalable algorithm with broad applicability in artificial intelligence.

1 INTRODUCTION

Optimal transport (OT) has emerged as a core tool of artificial intelligence, owing to its unique ability to compare probability distributions while respecting the underlying geometry of the data space (Villani, 2003). Despite this versatility, classical OT suffers from the rigid full-mass constraint which makes it sensitive to outliers, noise, and distributional shifts, preventing any form of partial or adaptive matching. To mitigate this issue, two major extensions have been proposed. Partial optimal transport (Caffarelli & McCann, 2010; Figalli, 2010) relaxes the full-mass constraint by transporting a fixed, user-specified amount of mass, but choosing this budget a priori is difficult without domain knowledge. Unbalanced optimal transport (Liero et al., 2017) replaces the hard marginal constraints with soft divergence-based penalties; however, the appropriate penalty strength remains elusive and problem-dependent. More recently, elastic optimal transport (ELOT) (Yang et al., 2026) introduced a discrete formulation that combines two key relaxations: inequality mass constraints and a mixed-sign ground cost. This allows the amount of transported mass to be determined adaptively, based on the intrinsic structure of the problem, thereby overcoming the fixed-mass limitations.

Although the discrete ELOT (Yang et al., 2026) rooted in linear programming is logically intuitive and high-performing, it leaves a substantial theoretical vacuum concerning the continuous counterpart. More broadly, both the theoretical foundation and the scalable algorithm for adaptive-mass optimal transport remain open areas. To this end, we propose Neural Adatptive Optimal Transport (NAOT) to establish the continuous theoretical foundation of adaptive optimal transport. The transition from discrete to continuous formulations unlocks deep mathematical insights: duality theory over function spaces, Brenier's theorem that characterizes the optimal transport map as the gradient of a convex potential. These theoretical pillars not only illuminate the intrinsic structure of the problem but also underpin scalable algorithms for large-scale problems. We formulate the continuous elastic optimal transport problem with inequality marginal constraints and a mixed-sign cost function. Building upon Brenier's seminal work (Brenier, 1991), we derive a duality theory and prove a Brenier-type theorem for the quadratic cost. Specifically, we show that the optimal transport map for NAOT is the gradient of a convex potential. Leveraging the theoretical results, we develop an efficient and scalable algorithm based on input-convex neural networks (ICNNs). By param-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

eterizing the dual potentials with ICNNs, we obtain the transport map as the gradient of the network output with respect to its input. This approach naturally enforces convexity by restricting the search to compact convex function classes, avoids the prohibitive cost of evaluating conjugate functions per sample, and scales gracefully to large-scale problems, which is a critical advantage over classical linear-programming-based solvers. We demonstrate the practical utility of NAOT on the benchmark sciPlex3 dataset (Srivatsan et al., 2020) for predicting cellular responses to anti-cancer drugs, which remains one of the grand challenges of biology (Schiebinger et al., 2019). By enabling adaptive-mass transport, NAOT accurately aligns the distribution of control cells with that of perturbed cells, while simultaneously accounting for cell growth and death. Our extensive experiments show that NAOT outperforms state-of-the-art methods in predicting perturbation responses across various drug dosages and cell lines, and it generalizes well to unseen dosages.

NAOT provides a general-purpose framework for adaptive distribution alignment which is a fundamental challenge of machine learning, with impact across multiple areas of artificial intelligence. By unifying the discrete and continuous perspectives of adaptive OT, our work not only fills a crucial theoretical gap but also delivers a practical tool that is both principled and computationally feasible for large-scale problems. The contributions of the work are highlighted as follows:

- We propose Neural Adaptive Optimal Transport, establishing a continuous foundation for adaptive optimal transport. We derive duality theory and a Brenier-type theorem, proving the optimal map is the gradient of a convex potential.
- We develop an ICNN-based solver that parameterizes dual potentials and obtains the transport map via automatic differentiation, avoiding per-sample conjugate evaluations and scaling efficiently to massive datasets.
- We apply NAOT to single-cell perturbation prediction, adaptively aligning control and perturbed cells while accounting for proliferation and death. Experiments on sciPlex3 benchmark show superior performance across both seen and unseen drug dosages.

2 Neural Adaptive Optimal Transport

In this section, we propose the novel neural adaptive optimal transport (NAOT) method. We first introduce the general formulation of NAOT and then derive the duality theory and Brenier potential. Based on the theoretical analysis, we design a neural solver to learn the transport map efficiently. The review of related work and preliminaries could be found in Appendices A and B, respectively.

2.1 Formulation

Suppose $X, Z \subset \mathbb{R}^d$ are domains in Euclidean space. $\mathcal{P}(X)$ denotes the set of nonnegative Borel measures on a space X. Let $\mu \in \mathcal{P}(X)$ and $\nu \in \mathcal{P}(Z)$ be two Borel measures with finite second-order moments, and we have $d\mu = f(x)dx$ and $d\nu = g(z)dz$, where $f(x)$ and $g(z)$ are density functions. Whenever A is the Borel subset of X, $\mu[A]$ denotes the mass located inside A. Denote the space of bounded continuous functions in X by $\mathcal{C}_b(X)$, and the set of all convex functions in $\mathcal{C}_b(X)$ by $\mathcal{F}(X)$. We model the transport plans by nonnegative Borel measure $\gamma \in \mathcal{P}(X \times Z)$, where $d\gamma(x, z)$ measures the amount of mass transferred from location x to location z. Denote $spt(\gamma)$ the support of γ, which refers to the smallest closed subset of $X \times Z$ carrying the full mass of γ, and define $spt(\mu)$ and $spt(\nu)$ accordingly. The cost function $c(x, z)$ tells how much it costs to move one unit of mass from location $x \in \mathcal{X}$ to location $z \in \mathcal{Z}$.

Given two Borel measures $\mu \in \mathcal{P}(X)$ and $\nu \in \mathcal{P}(Z)$ on the source X and target Z respectively, as well as the bounded continuous cost function $c(x, z)$, the primary problem of neural adaptive optimal transport is to find the optimal transport plan $\gamma \in \mathcal{P}(X \times Z)$ which moves the mass adaptively from the source to the target at minimal cost:

$$\min_{\gamma \in \Gamma_{\leq}(\mu,\nu)} \int_{X \times Z} c(x, z) d\gamma(x, z) \tag{1}$$

where the set of admissible transport plans is denoted by $\Gamma_{\leq}(\mu, \nu)$ whose left and right marginals are dominated by μ and ν respectively, i.e.

$$\Gamma_{\leq}(\mu, \nu) = \left\{ \gamma \in \mathcal{P}(X \times Z) \;\middle|\; \begin{array}{l} \gamma[A \times Z] \leq \mu[A] \\ \gamma[X \times B] \leq \nu[B] \end{array} \right\} \tag{2}$$

for all Borel subsets $A \subset X$ and $B \subset Z$. Notice that the inequality constraints are used in $\Gamma_{\leq}(\mu, \nu)$. That is to say, if $\mu \in \mathcal{P}(X)$ and $\nu \in \mathcal{P}(Z)$ are probability measures ($\mu[X] = \nu[Z] = 1$), the transport plan $\gamma \in \mathcal{P}(X \times Z)$ is not necessarily a probability measure, and we have $\gamma[X \times Z] \leq 1$ in this case. Similar to ELOT (Yang et al., 2026), NAOT requires the costs to be mixed-sign to avoid the potential degeneracy. This could be a limitation in certain tasks, but it is widely applicable to various scenarios where the cost could be derived from multiple feature spaces (e.g., original feature space and latent feature space, or marginal and conditional distributions).

2.2 Duality Theory and Brenier Potential

We conduct theoretical analysis for NAOT, which not only provides insights into the intrinsic structure of NAOT but also sheds light on developing an efficient and scalable algorithm to solve the problem. We start from the general cost function and focus on NAOT with quadratic cost.

Theorem 1 (**Duality of NAOT with General Cost**). *Minimizing the primary NAOT problem in Equation 1 is equivalent to maximizing its dual problem, i.e.,*

$$\min_{\gamma \in \Gamma_{\leq}(\mu,\nu)} \int_{X \times Z} c(x,z) d\gamma(x,z) = \max_{\substack{\phi(x)+\psi(z) \leq c(x,z) \\ \phi(x), \psi(z) \leq 0}} \int_X \phi(x) d\mu(x) + \int_Z \psi(z) d\nu(z) \tag{3}$$

where $\phi(x) \in C_b(X)$ and $\psi(z) \in C_b(Z)$ are bounded continuous functions.

The proof can be found in Appendix C.1. Duality theory plays a central role in optimal transport theory. It transforms the original complex 'transport plan' problem into a more tractable 'potential function' optimization problem. The primal problem seeks a transport plan defined on the joint space of the two distributions, which is extremely high-dimensional. In contrast, the dual problem only requires solving for two potential functions defined on the lower-dimensional spaces. Therefore, dual theory not only provides rigorous theoretical foundations to analyze the solution structure of transport map, but also enables the scalable algorithms that make OT practically useful, thereby laying the theoretical groundwork for the application of OT in various areas of artificial intelligence.

Next, we derive the duality theory for NAOT with quadratic cost:

$$c(x,z) = \frac{1}{2}\|x - z\|^2 - \alpha \langle Mx, Mz \rangle \tag{4}$$

where α is a non-negative coefficient to trade off the impact of the original feature space or the latent space, and M is the projection matrix mapping the original feature space to the latent space. The projection matrix M is learnable and task-dependent, which could be used to encode the prior knowledge of the specific task.

Theorem 2 (**Duality of NAOT with Quadratic Cost**). *The primary NAOT problem (Equation 1) with quadratic cost (Equation 4) is equivalent (up to an additive constant) to its dual form defined as follows*

$$\min_{\substack{\phi(x)+\psi(z) \geq x^T Q z \\ \phi(x) \geq \frac{1}{2}\|x\|^2, \psi(z) \geq \frac{1}{2}\|z\|^2}} \int_X \phi(x) d\mu(x) + \int_Z \psi(z) d\nu(z) \tag{5}$$

where $\phi(x) \in \mathcal{C}_b(X)$ and $\psi(z) \in \mathcal{C}_b(Z)$ are bounded continuous functions and $Q = I + \alpha M^T M$.

The proof can be found in Appendix C.2. In Theorem 2, we derive the duality theory for NAOT with quadratic cost. However, the constraints in the duality of NAOT are notoriously challenging because they require the evaluation of the functions at every point of the domain, which is not scalable. Next we generalize the Brenier theorem (Brenier, 1991) from the classical OT to NAOT and show that the optimal transport map of NAOT is the gradient of a convex function, the so-called Brenier potential. This makes it feasible to develop an efficient algorithm to find the optimal transport map for NAOT.

The conjugates of $\phi(x)$ and $\psi(z)$ are defined as follows:

$$\phi^*(z) = \sup_{x \in X} x^T \big(I + \alpha M^T M\big) z - \phi(x), \tag{6}$$

$$\psi^*(x) = \sup_{z \in Z} x^T \big(I + \alpha M^T M\big) z - \psi(z). \tag{7}$$

Theorem 3 (**Optimal Transport Map**). *Given the measures μ and ν with compact supports $X, Z \subset \mathbb{R}^d$, and the cost function $c(x,z) = \frac{1}{2}\|x-z\|^2 - \alpha\langle Mx, Mz\rangle$, NAOT's optimal transport map T from μ to ν exists. It is the gradient of a convex function, the so-called Brenier potential, i.e.,*

$$z = T(x) = \big(I + \alpha M^T M\big)^{-1} \cdot \nabla\phi(x) = \big(I + \alpha M^T M\big)^{-1} \cdot \nabla\psi^*(x). \tag{8}$$

Likewise, for the inverse optimal transport map T^{-1}, we have

$$x = T^{-1}(z) = \big(I + \alpha M^T M\big)^{-1} \cdot \nabla\psi(x) = \big(I + \alpha M^T M\big)^{-1} \cdot \nabla\phi^*(z). \tag{9}$$

The proof can be found in Appendix C.3. Building upon the fundamental connection from Brenier theorem established in Theorem 3, we are ready to eliminate the constraints in the Kantorovich duality by restricting the search over the set of all convex functions. It allows us to use input convex neural networks (Makkuva et al., 2020) to solve the neural adaptive optimal transport problem.

According to the definition of the conjugate function in Equation 6, the constrained optimization problem over two potentials can be formulated as a convex program over a single potential

$$\min_{\substack{\phi(x)+\psi(z)\geq x^T Q z \\ \phi(x)\in\mathcal{C}_b(X),\psi(z)\in\mathcal{C}_b(Z)}} \int_X \phi(x)d\mu(x) + \int_Z \psi(z)d\nu(z) = \min_{\substack{\phi(x)\in\mathcal{C}_b(X) \\ \phi^*(z)\in\mathcal{C}_b(Z)}} \int_X \phi(x)d\mu(x) + \int_Z \phi^*(z)d\nu(z). \tag{10}$$

Next, we show that the search can be restricted in the set of convex functions.

Theorem 4 (**NAOT over Convex Functions**). *The optimization of the duality of NAOT defined in Equation 5 is equivalent to*

$$\min_{\substack{\phi(x)\in\mathcal{F}(X) \\ \phi(x)\geq\frac{1}{2}\|x\|^2}} \max_{\substack{\psi(z)\in\mathcal{F}(Z) \\ \psi(z)\geq\frac{1}{2}\|z\|^2}} \int_X \phi(x)d\mu(x) + \int_Z \Big[\big\langle z, \nabla\psi(z)\big\rangle - \phi\Big(Q^{-1}\nabla\psi(z)\Big)\Big]d\nu(z) \tag{11}$$

where $Q = I + \alpha M^T M$.

The proof can be found in Appendix C.4. Until now, by summing up Theorems 2 and 4, it is straightforward to prove Theorem 5, which is omitted.

Theorem 5 (**Neural Adaptive Optimal Transport**). *The optimization of the primary NAOT problem with quadratic cost*

$$\min_{\gamma\in\Gamma_\leq(\mu,\nu)} \int_{X\times Z} \left[\frac{1}{2}\|x-z\|^2 - \alpha\big\langle Mx, Mz\big\rangle\right]d\gamma(x,z) \tag{12}$$

is equivalent to

$$\min_{\substack{\phi(x)\in\mathcal{F}(X) \\ \phi(x)\geq\frac{1}{2}\|x\|^2}} \max_{\substack{\psi(z)\in\mathcal{F}(Z) \\ \psi(z)\geq\frac{1}{2}\|z\|^2}} \int_X \phi(x)d\mu(x) + \int_Z \Big[\big\langle z, \nabla\psi(z)\big\rangle - \phi\Big(Q^{-1}\nabla\psi(z)\Big)\Big]d\nu(z) \tag{13}$$

where $Q = I + \alpha M^T M$.

Theorem 5 shows that the original NAOT problem can be solved by optimizing the Brenier potential functions, $\phi(x)$ and $\psi(z)$, over the set of convex functions $\mathcal{F}(X)$ and $\mathcal{F}(Z)$. By parameterizing these potentials with input convex neural networks (Amos et al., 2017), it restricts the search to compact convex function classes. This enables efficient gradient-based optimization via automatic differentiation, avoiding costly per-sample conjugate evaluations required by classical dual solvers.

2.3 Neural Solver

In this subsection, we propose the novel neural solver based on input convex neural networks to learn the transport map of NAOT.

For an observed dataset, let $X = \{x_i\}_{i=1}^n$ be a set of samples in the source domain, and $Z = \{z_j\}_{j=1}^m$ a set of samples in the target domain. The measures for the source and target domains are denoted by $\mu = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$ and $\nu = \frac{1}{m}\sum_{j=1}^m \delta_{z_j}$ respectively, where δ is the Dirac function. We propose to optimize the following objective which is the empirical counterpart of Equation 13:

$$\min_{\substack{\phi(x)\in\mathcal{F}(X) \\ \phi(x)\geq\frac{1}{2}\|x\|^2}} \max_{\substack{\psi(z)\in\mathcal{F}(Z) \\ \psi(z)\geq\frac{1}{2}\|z\|^2}} J_{emp}(\phi,\psi) \tag{14}$$

where

$$J_{emp}(\phi,\psi)=\frac{1}{n}\sum_{i=1}^{n}\phi(x_i)+\frac{1}{m}\sum_{j=1}^{m}\big\langle z_j,\nabla\psi(z_j)\big\rangle-\phi\big(Q^{-1}\nabla\psi(z_j)\big). \tag{15}$$

As a common practice, we relax the constraints $\phi(x)\geq\frac{1}{2}\|x\|^2,\psi(z)\geq\frac{1}{2}\|z\|^2$, and instead minimize the penalty term for those violating the constraints:

$$R_c(\phi,\psi)=\frac{1}{n}\sum_{i=1}^{n}max^2\Big\{\frac{1}{2}\big\|x_i\big\|^2-\phi(x_i),0\Big\}+\frac{1}{m}\sum_{j=1}^{m}max^2\Big\{\frac{1}{2}\big\|z_j\big\|^2-\psi(z_j),0\Big\}. \tag{16}$$

The overall objective is

$$\min_{\phi(x)\in\mathcal{F}(X)}\max_{\psi(z)\in\mathcal{F}(Z)}J(\phi,\psi)\equiv\min_{\phi(x)\in\mathcal{F}(X)}\max_{\psi(z)\in\mathcal{F}(Z)}J_{emp}(\phi,\psi)+\sigma R_c(\phi,\psi) \tag{17}$$

In general, we use two input convex neural networks (ICNNs) (Amos et al., 2017) to approximate the convex functions ϕ and ψ, denoted by ICNN_ϕ and ICNN_ψ respectively. In the main loop, it plays the minmax game to update the two ICNNs iteratively. Thereafter the transport map T or the inverse transport map T^{-1} can be derived from the gradient of the potentials, ϕ and ψ, respectively.

ICNNs are a class of scalar-valued neural network $f(x;\theta)$ such that the function $x\to f(x;\theta)\in\mathbb{R}$ is convex, where θ denotes the learnable parameters of neural network. The neural network architecture for an ICNN is as follows. Given an input $x\in\mathbb{R}^d$, the mapping $x\to f(x;\theta)$ is given by a L-layer feed-forward neural network using the following equations for $l=0,1,\cdots,L-1$: $h_{l+1}=\xi_l(W_l^h h_l+W_l^x x+b_l)$ and $f(x;\theta)=h_L$, where h_l denotes the layer activations which is initialized with $h_0=\mathbf{0}$, and ξ_l is the entry-wise activation function. According to (Amos et al., 2017), $f(x;\theta)$ is convex in x provided: i) all entries of the weights W_l^h are non-negative; ii) activation function ξ_0 is convex; iii) $\xi_l(l=1,\cdots,L-1)$ is convex and non-decreasing. ICNNs are rich enough and any convex function over a compact domain can be approximated in sup norm by an ICNN to the desired accuracy (Chen et al., 2019).

Note that different ICNN variants could be used in our algorithm. In the experiments, we use partially input convex neural networks (PICNN) (Makkuva et al., 2020; Bunne et al., 2022) to incorporate context variables c such as the drug dosages or cell lines to modulate a family of convex functions $f(x,c;\theta)$. The L-layer PICNN architecture is defined by recurrences for $l=0,1,\cdots,L-1$:

$$\begin{aligned}
u_{l+1}&=\xi_l^u(\hat{U}_l u_l+\hat{b}_l^u)\\
h_{l+1}&=\xi_l^h\Big(W_l^h\big(h_l\circ[U_l^h u_l+b_l^h]_+\big)+W_l^x\big(x\circ[U_l^x u_l+b_l^x]_+\big)+W_l^u u_l+b_l^u\Big)\\
f(x,c;\theta)&=h_L
\end{aligned}$$

where u_l and h_l denote the hidden units which are initialized with $u_0=c,h_0=\mathbf{0}$, and $\circ$ denotes the element-wise product. ξ_l^u and ξ_l^h denote the entry-wise activation functions. $[\cdot]_+$ is the softplus function. PICNNs are neural networks that can be evaluated over a pair of inputs (x,c), but which are only required to be convex with respect to x. As in ICNN, we need the weights in W_l^h to be non-negative. The hidden states u_l act as context-dependent modulators. The PICNN architecture can be seen as an ICNN whose weights and biases are modulated by the context variables.

3 Application: Single-Cell Perturbation Prediction

Single-cell RNA sequencing (scRNA-seq) (Tanay & Regev, 2017) has significantly advanced our knowledge of biological systems. The scRNA-seq technique profiles gene expression in millions of cells across tissues and species. The advent of scRNA-seq can provide measurements for individual cells in control and perturbed conditions. However, scRNA-seq destroys cells to obtain the measurements, resulting in unpaired control and perturbed single-cell observations. Hence, modeling single-cell perturbation responses with access to unpaired populations of control and perturbed cells remains one of the grand challenges of biology (Schiebinger et al., 2019). Characterizing perturbation responses at the single-cell level is crucial to many biological applications (Schiebinger et al., 2019; Cui et al., 2024; Huang & Liu, 2024) such as predicting a patient's response to drug treatments, and understanding how and why individual tumor cells evade cancer therapies.

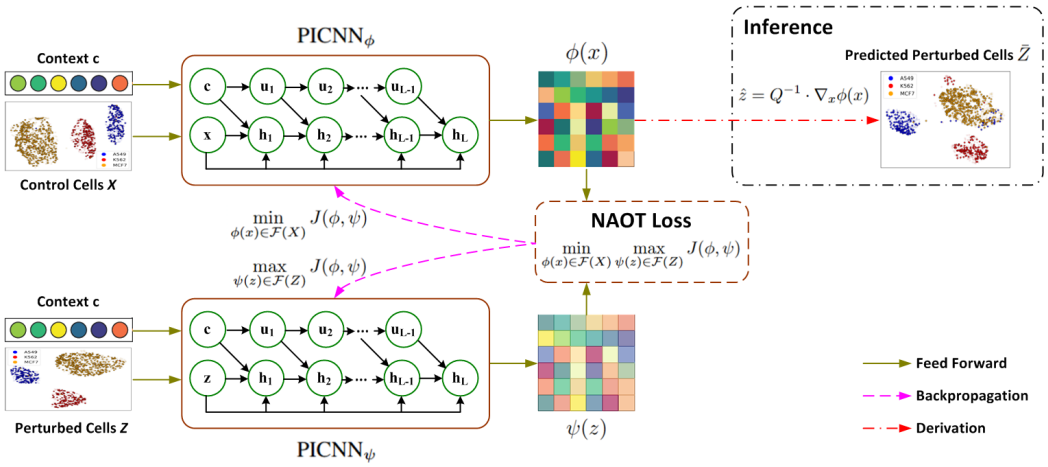


Figure 1: Overview of NAOT framework for single-cell perturbation response prediction. Given unpaired profiles of control cells X and perturbed cells Z together with context variable c, NAOT aims to generate the predicted perturbed cells $\bar{Z}$ which is similar to the observed perturbed cells Z at the distribution level. It solves an adaptive-mass optimal-transport problem by playing a minmax game to iteratively update two partially input-convex neural networks (PICNN$_\phi$, PICNN$_\psi$). The optimal transport map $T(\cdot, c)$ is obtained via automatic differentiation as the gradient of learned convex potentials. In the inference stage, we use $T(\cdot, c)$ to map each control-cell gene-expression profile to its predicted perturbed state as $\bar{z} = T(x, c) = Q^{-1} \cdot \nabla_x \phi(x)$.

However, the prominent methods (Bunne et al., 2023; 2022) based on optimal transport fall short of solving the highly complex pairing problem of cell populations while simultaneously accounting for the dynamic nature of cells such as proliferation and death. Optimal transport aims to find an optimal way to move a pile of sand (source distribution) into a hole (target distribution). But the application of optimal transport to cell populations differs in the key respect: unlike sand, cells can proliferate and die. Thus, it requires the system to either transport or create and destroy the mass to accommodate cell growth and death.

To this end, we apply neural adaptive optimal transport to predict single-cell perturbation responses. Figure 1 provides an overview of the NAOT framework. For the observed single-cell data, let $\mu = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i}$ and $\nu = \frac{1}{m}\sum_{j=1}^{m} \delta_{z_j}$ denote the empirical distributions of control cells and perturbed cells respectively, where $x_i \in \mathbb{R}^d$ and $z_j \in \mathbb{R}^d$ represent the gene expression profiles of cells. The task of predicting single-cell perturbation responses is to learn a transport map that maps each control cell to its corresponding perturbed state. Unlike classical OT, NAOT only requires that the transported mass does not exceed the available mass in either distribution, i.e., $\gamma \mathbf{1}_m \leq \mu$ and $\gamma^T \mathbf{1}_n \leq \nu$. This relaxation is biologically meaningful: it allows for cell proliferation (where the target distribution has more mass than the source) and cell death (where the target distribution has less mass), both of which are natural phenomena in response to drug perturbations. The ground cost function in NAOT for perturbation prediction is designed to capture both the feature-level similarity and the label-level similarity between cells. Specifically, the first term of Equation 4 encourages the transport of cells that are close in the gene expression space, while the second term promotes alignment in the projected space, which can capture higher-order biological relationships.

To incorporate contextual information such as drug dosage or cell line, we parameterize the transport map as a function of a context variable c. Specifically, we learn a family of transport maps $T(\cdot, c)$ conditioned on the context c, which enables NAOT to generalize to unseen contexts. By conditioning on the context, NAOT learns a continuous mapping from control cells to perturbed cells that varies smoothly with the context, allowing for interpolation and extrapolation to new conditions. The end-to-end training algorithm is summarized in Appendix D, where the transport plan and the neural network parameters are optimized jointly.

In summary, NAOT offers a biologically-grounded and scalable solution to the problem of single-cell perturbation prediction. By enabling adaptive-mass transport, it overcomes the limitations of

existing OT-based methods and provides accurate predictions in the presence of cell proliferation and death. Furthermore, by leveraging the duality theory and the Brenier potential, NAOT solves the adaptive transport problem efficiently using partially input convex neural networks.

4 EXPERIMENTS

We conduct extensive experiments to verify the effectiveness of NAOT on predicting cellular perturbation responses conditioned on various contexts, as well as its generalization on unseen contexts.

The experiments are evaluated on the sciPlex3 dataset (Srivatsan et al., 2020) which is the most popular benchmark for modeling cellular perturbation responses. The sciPlex3 dataset includes measurements from 762,039 cells, spanning three human cancer cell lines. For the evaluation metrics, it is common to adopt the discrepancy metrics such as maximum mean discrepancy (MMD), Wasserstein distance (WD), perturbation signatures (ℓ_2(PS)), and Fréchet inception distance (FID), where lower values indicate better performance. We compare NAOT with three strong baselines: CELLOT (Bunne et al., 2023), CondOT (Bunne et al., 2022), and GENOT (Klein et al., 2024). Please refer to Appendix E.1 for details about experiment setup such as datasets, metrics, baselines, networks and hyperparameters.

4.1 PERTURBATION RESPONSES ON DRUG DOSAGE

In the first task, we demonstrate that NAOT outperforms state-of-the-art methods on predicting cellular perturbation responses conditioned on drug dosage.

To maintain comparability with existing baselines, we evaluated the performance of NAOT using observed data from cells perturbed by two typical anti-cancer compounds: givinostat and trametinib which were also used in CondOT (Bunne et al., 2022), conditioned on drug dosages (10 nM, 100 nM, 1,000 nM, 10,000 nM). The pre-processed dataset contains 17,565 unperturbed cells, 3,541 cells perturbed by givinostat, and 3,277 cells perturbed by trametinib. Given the measurements for the control cells and the perturbed cells with different dosages, we seek to learn a dosage-dependent transport map from the control cell population to the perturbed cell population. Note that in this task all drug dosages in test set are seen in the training phase.

As shown in Table 1, NAOT consistently surpasses the comparison methods, leading to more precise predictions of perturbed cell distributions across different dosage levels. This superiority is reflected in lower average metrics including MMD, Wasserstein distance, $\ell_2(PS)$, and FID. We attribute the advantage of NAOT to its ability to learn a transport map that considers explicitly the data geometries of cell populations, while respecting the dynamic nature of cell growth and proliferation. Both CELLOT (Bunne et al., 2023) and CondOT (Bunne et al., 2022) adopt classical optimal transport to align the control cell distribution with the perturbed cell distribution, demonstrating prominent performance on predicting cellular perturbation responses. However, the full-mass constraint in the classical optimal transport is not the most suitable to model cellular responses because it neglects the proliferation and death of cells. GENOT (Klein et al., 2024) applied unbalanced OT to model cellular growth and death or account for outliers, achieving competitive performance in some metrics such as Wasserstein distance and $\ell_2(PS)$ for drug Trametinib. It suggests that relaxing the full-mass preservation is prevalent in highly noisy measurements in single-cell genomics. NAOT introduces the adaptive-mass constraint and further unleashes the potential of optimal transport, leading to significant performance gain in predicting perturbation responses.

Table 1: Evaluation of predicting dosage-dependent perturbation responses.

Method	Drug	MMD ↓	WD ↓	ℓ_2(PS) ↓	FID ↓
CELLOT (Bunne et al., 2023)	Givinostat	0.1218 ± 0.0563	120.68 ± 32.02	8.67 ± 2.99	192.15 ± 72.75
CondOT (Bunne et al., 2022)		0.0556 ± 0.0131	47.28 ± 7.96	3.60 ± 0.41	48.34 ± 6.77
GENOT (Klein et al., 2024)		0.1572 ± 0.1027	50.63 ± 30.54	5.10 ± 4.48	85.78 ± 62.29
NAOT		**0.0421** ± 0.0074	**38.86** ± 7.91	**2.13** ± 0.65	**33.91** ± 6.39
CELLOT (Bunne et al., 2023)	Trametinib	0.0982 ± 0.0064	71.41 ± 1.27	6.70 ± 0.06	96.97 ± 6.20
CondOT (Bunne et al., 2022)		0.0510 ± 0.0059	47.02 ± 0.80	3.20 ± 0.08	45.96 ± 4.41
GENOT (Klein et al., 2024)		0.1052 ± 0.0237	**30.93 ± 6.03**	**1.76 ± 0.70**	46.71 ± 11.13
NAOT		**0.0416** ± 0.0069	42.43 ± 1.04	2.03 ± 0.15	**35.58** ± 4.75

4.2 Generalization to Unseen Drug Dosage

We further investigate NAOT's generalization performance in the out-of-distribution scenarios such as unseen drug dosages.

Handling out-of-distribution data is critical for ensuring the robustness of NAOT when deployed in real-world environments. In the cell biology domain, it is expensive to obtain all scRNA-seq measurements for cells perturbed with a variety of cancer drugs at various dosages. Therefore, it is crucial for the model to infer the perturbed cell distribution for the unseen drug dosages. Given the measurements for the control cells and the perturbed cells with different dosages, we seek to learn a global, parameterized transport map sensitive to the dosage. Note that in this task the drug dosage in test set is unseen in the training phase. Specifically, in this training phase, we train NAOT using observed data from cells perturbed by two drugs, conditioned on drug dosages (10 nM, 1,000 nM, 10,000 nM). In test phase, the observed cells perturbed with drug dosage 100 nM are used for evaluation. This setup allows us to assess the model's capacity to predict cellular responses to drug dosages not encountered in training phase.

Table 2 shows that NAOT is capable of generalizing effectively to previously unobserved dosages, highlighting its ability to predict perturbation responses via interpolation or extrapolation of the dosage values. Due to the distribution shift between perturbed populations in training and inference stages, learning the transport map in the out-of-distribution scenarios becomes more challenging. The consistently strong performance of NAOT, as evidenced in Table 2, reinforces its superiority over the comparison methods.

Table 2: Generalization to unseen drug dosage (100 nM).

Method	Drug	MMD ↓	WD ↓	ℓ_2(PS) ↓	FID ↓
CELLOT (Bunne et al., 2023)	Givinostat	0.0530 ± 0.0007	92.02 ± 0.65	6.44 ± 0.09	115.00 ± 1.03
CondOT (Bunne et al., 2022)		0.0155 ± 0.0006	42.13 ± 0.36	2.51 ± 0.06	19.88 ± 0.51
GENOT (Klein et al., 2024)		0.0742 ± 0.0081	$\mathbf{29.42 \pm 3.41}$	1.71 ± 0.12	40.38 ± 4.54
NAOT		$\mathbf{0.0084} \pm 0.0001$	36.10 ± 0.20	$\mathbf{1.14} \pm 0.02$	$\mathbf{10.53} \pm 0.41$
CELLOT (Bunne et al., 2023)	Trametinib	0.0661 ± 0.0012	65.12 ± 0.67	6.61 ± 0.07	73.14 ± 1.32
CondOT (Bunne et al., 2022)		0.0138 ± 0.0004	38.57 ± 0.25	2.01 ± 0.09	17.29 ± 0.43
GENOT (Klein et al., 2024)		0.0806 ± 0.0196	$\mathbf{27.58 \pm 5.06}$	$\mathbf{1.03 \pm 0.37}$	37.65 ± 8.01
NAOT		$\mathbf{0.0102} \pm 0.0001$	35.33 ± 0.10	1.25 ± 0.01	$\mathbf{10.73} \pm 0.14$

4.3 Perturbation Responses on Cell Line

Cancer drugs might cause divergent cellular responses in each cell line. Therefore, we evaluate NAOT in the task of predicting perturbation responses conditioned on cell line.

The cell population contains three well-characterized cancer cell lines, including A549, a human lung adenocarcinoma, K562, a chronic myelogenous leukemia, and MCF7, a mammary adenocarcinoma cell line. Given the measures for the control and the perturbed cells in different cell lines, we seek to learn a global, parameterized transport map sensitive to the cell lines.

As demonstrated in Table 3, NAOT outperforms the baseline models across a variety of evaluation metrics in most cases. We attribute this to NAOT's capability of aligning the cell distributions not only in the whole-population level but also in the subpopulation level. NAOT encodes the prior knowledge of cell lines into the ground cost function to learn the optimal transport map. Intuitively, the more similar the two cells are in both feature space and label space, the more probability mass will be transported between them. Therefore, NAOT is able to capture the subpopulation structures among the control and perturbed cells, leading to fine-grained distribution alignment. Furthermore, NAOT's distinctive mechanism of adaptive-mass transporting provides a more robust solution for distribution alignment between the control and perturbed cell populations.

4.4 Generalization versus Specialization on Drug Dosage

To explore the generalization and specialization abilities, we conduct a comparative evaluation of NAOT in both multi-dosage and single-dosage settings to predict dosage-dependent perturbation responses for two drugs, Trametinib and Givinostat. For the multi-dosage setting, the model is trained

Table 3: Evaluation of predicting perturbation responses conditioned on cell line.

Method	Drug	MMD ↓	WD ↓	ℓ_2(PS) ↓	FID ↓
CELLOT (Bunne et al., 2023)	Givinostat	0.0517 ± 0.1255	83.46 ± 29.56	5.66 ± 1.61	100.37 ± 44.79
CondOT (Bunne et al., 2022)		0.0135 ± 0.0026	42.32 ± 3.88	1.58 ± 0.41	17.38 ± 2.55
GENOT (Klein et al., 2024)		0.0874±0.0226	**36.24±6.28**	2.55±0.74	54.76±13.11
NAOT		**0.0113** ± 0.0030	40.73 ± 4.33	**1.34** ± 0.25	**13.71** ± 2.96
CELLOT (Bunne et al., 2023)	Trametinib	0.0604 ± 0.0079	59.01 ± 21.89	5.94 ± 1.12	61.08 ± 27.85
CondOT (Bunne et al., 2022)		0.0114 ± 0.0057	34.00 ± 6.25	**0.90** ± 0.24	10.92 ± 5.29
GENOT (Klein et al., 2024)		0.0806±0.0189	**27.59±5.55**	0.99±0.41	38.17±8.80
NAOT		**0.0106** ± 0.0048	36.50 ± 7.59	1.00 ± 0.24	**10.51** ± 5.13

on all dosage conditions. In contrast, the models are trained separately for each individual dosage in single-dosage setting. Although the single-dosage approach faces challenges in generalizing to novel contexts and requires that all contexts be explicitly defined, it establishes a pseudo lower bound that the multi-dosage model is expected to approach.

The results are shown in Table 4. NAOT exhibits a relatively small performance gap between the multi-dosage and single-dosage settings, indicating that the model can effectively encode conditional information during joint training. Meanwhile, CELLOT shows a slight performance drop under the multi-dosage setting, and although CondOT achieves some improvement, its overall performance remains inferior to that of NAOT. These results suggest that the adaptive mass mechanism enhances both distribution fitting accuracy and the stability of joint learning across multiple conditions. In contrast, GENOT undergoes notable performance degradation under multi-dosage joint training, indicating its limitation in terms of insufficient stability when jointly modeling multiple conditions.

Table 4: Predicting cellular perturbation responses in single-dosage and multi-dosage settings.

Method	**Setting**	MMD ↓			
		10 nM	**100 nM**	**1,000 nM**	**10,000 nM**
CELLOT (Bunne et al., 2023)	single-dosage	0.0832±0.0210	0.0742±0.0200	0.0716±0.0126	0.1376±0.0379
CondOT (Bunne et al., 2022)		0.0521±0.0075	0.0527±0.0044	0.0455±0.0023	0.0587±0.0094
GENOT (Klein et al., 2024)		0.0691±0.0156	0.0685±0.0133	0.0721±0.0187	0.1331±0.0516
NAOT		**0.0405±0.0078**	**0.0396±0.0044**	**0.0379±0.0041**	**0.0444±0.0012**
CELLOT (Bunne et al., 2023)	multi-dosage	0.0888±0.0081	0.0911±0.0023	0.0895±0.0055	0.1367±0.0478
CondOT (Bunne et al., 2022)		0.0496±0.0060	0.0505±0.0004	0.0483±0.0038	0.0572±0.0074
GENOT (Klein et al., 2024)		0.1001±0.0256	0.0973±0.0215	0.0977±0.0191	0.1946±0.1177
NAOT		**0.0407±0.0068**	**0.0399±0.0029**	**0.0390±0.0055**	**0.0457±0.0023**

More Results: Furthermore, through intuitive visualizations and case study, we qualitatively demonstrate that NAOT is able to align the control and perturbed cell distributions in the population, cell line, and gene levels, yielding accurate and nuanced predictions. We also evaluate the robustness of NAOT to the hyperparameters. Please refer to Appendix E for more detailed experiments such as visualization of distributional alignment, and hyperparameter sensitivity analysis.

5 Conclusion

We propose an innovative method to solve the highly complex alignment problem of cell populations, which is crucial to many biological applications. The proposed NAOT method outperforms the state-of-the-art methods in predicting single-cell perturbation responses. To put it in a much broader perspective, inferring cell perturbation responses is linked to a general problem of interest to many areas of artificial intelligence, i.e., how to build adaptive mappings between probability distributions. The proposed NAOT method offers a principled solution to tackle this problem by transporting adaptive-mass rather than full-mass or fixed-mass. Then we generalize Brenier's seminal theory from the classical optimal transport to the neural adaptive optimal transport with quadratic costs. It consolidates the theoretical foundation of NAOT and sheds light on developing a neural optimizer based on input convex networks. The optimization algorithm enjoys the properties of convexity and scalability, which are critical to the real applications of artificial intelligence. As a novel family member of optimal transport, we envision that NAOT will find broad applications in a variety of domains such as artificial intelligence, biology, economics, healthcare, physics, etc.

AI USE STATEMENT

(This section is **required** and does not count toward the page limit.)

In this work, we used generative AI tools for language polishing and editing of the manuscript, mainly for grammar checking and improving readability. We have not used generative AI tools for generating research ideas, designing methods, conducting experiments, analyzing data, or producing scientific results, and the remaining required disclosure tasks are not applicable to this work. We have reviewed all AI-assisted work. Specifically, all AI-polished text was manually verified by the authors for scientific accuracy, and potential plagiarism, and any AI-suggested edits were checked against the original content. We take responsibility for the final content of this work.

ETHICS STATEMENT

(This section is **recommended** and does not count toward the page limit.)

The authors declare no conflicts of interest or ethical concerns. This work uses only public data, and raises no privacy, bias, or legal issues.

REPRODUCIBILITY STATEMENT

(This section is **recommended** and does not count toward the page limit.)

The source codes of this work are included in supplementary materials.

REFERENCES

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 2623–2631, 2019.

Brandon Amos, Lei Xu, and J. Zico Kolter. Input convex neural networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, volume 70, pp. 146–155, 2017.

Martín Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pp. 214–223, 2017.

Yann Brenier. Polar factorization and monotone rearrangement of vector-valued functions. *Communications on Pure and Applied Mathematics*, 44:375–417, 1991.

Charlotte Bunne, Andreas Krause, and Marco Cuturi. Supervised training of conditional monge maps. In *Neural Information Processing Systems (NeurIPS)*, volume 35, pp. 6859–6872, 2022.

Charlotte Bunne, Stefan G Stark, Gabriele Gut, Jacobo Sarabia Del Castillo, Mitch Levesque, Kjong-Van Lehmann, Lucas Pelkmans, Andreas Krause, and Gunnar Rätsch. Learning single-cell perturbation responses using neural optimal transport. *Nature Methods*, 20:1759–1768, 2023.

Luis Caffarelli and Robert J. McCann. Free boundaries in optimal transport and Monge-Ampere obstacle problems. *Annals of Mathematics*, 171:673–730, 2010.

Xavier Aramayo Carrasco, Grigoriy Ksenofontov, Aleksei Leonov, Iaroslav Sergeevich Koshelev, and Alexander Korotin. Entering the era of discrete diffusion models: A benchmark for schrödinger bridges and entropic optimal transport. In *The Fourteenth International Conference on Learning Representations (ICLR)*, 2026.

Laetitia Chapel and Mokhtar Z. Alaya. Partial optimal tranport with applications on positive-unlabeled learning. In *Neural Information Processing Systems (NeurIPS)*, 2020.

Yize Chen, Yuanyuan Shi, and Baosen Zhang. Optimal control via neural networks: A convex approach. In *7th International Conference on Learning Representations (ICLR)*, 2019.

Nicolas Courty, Rémi Flamary, Devis Tuia, and Alain Rakotomamonjy. Optimal transport for domain adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(9):1853–1865, 2017.

Haotian Cui, Chloe Wang, Hassaan Maan, Kuan Pang, Fengning Luo, Nan Duan, and Bo Wang. scGPT: toward building a foundation model for single-cell multi-omics using generative AI. *Nature methods*, 21(8):1470–1480, 2024.

Xiao Cui, Mo Zhu, Yulei Qin, Liang Xie, Wengang Zhou, and Houqiang Li. Multi-level optimal transport for universal cross-tokenizer knowledge distillation on language models. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pp. 23724–23732, 2025.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Neural Information Processing Systems (NeurIPS)*, pp. 2292–2300, 2013a.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Neural Information Processing Systems (NeurIPS)*, pp. 2292–2300, 2013b.

Alessio Figalli. The optimal partial transport problem. *Archive for Rational Mechanics and Analysis*, 195:533–560, 2010.

Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Neural Information Processing Systems (NeurIPS)*, pp. 6626–6637, 2017.

Wei Huang and Hui Liu. Predicting single-cell cellular responses to perturbations using cycle consistency learning. *Bioinformatics*, 40:i462–i470, 2024.

L. Kantorovitch. On the translocation of masses. *C. R. (Dokl.) Acad. Sci. URSS*, 37:199–201, 1942.

Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Dominik Klein, Théo Uscidda, Fabian Theis, and Marco Cuturi. Genot: Entropic (gromov) wasserstein flow matching with applications to single-cell genomics. *Advances in Neural Information Processing Systems*, 37:103897–103944, 2024.

Qiaonan Li, Guihua Wen, and Pei Yang. From patch, sample to domain: Capture geometric structures for few-shot learning. *Pattern Recognition*, 148:110147, 2024.

Matthias Liero, Alexander Mielke, and Giuseppe Savaré. Optimal entropy-transport problems and a new Hellinger-Kantorovich distance between positive measures. *Inventiones mathematicae*, 211: 969 – 1117, 2017.

Mohammad Lotfollahi, F. Alexander Wolf, and Fabian J. Theis. scgen predicts single-cell perturbation responses. *Nature Methods*, 16:715–721, 2019.

Mohammad Lotfollahi, Anna Klimovskaia Susmelj, Carlo De Donno, Leon Hetzel, Yuge Ji, Ignacio L Ibarra, Sanjay R Srivatsan, Mohsen Naghipourfar, Riza M Daza, Beth Martin, Jay Shendure, Jose L McFaline-Figueroa, Pierre Boyeau, F Alexander Wolf, Nafissa Yakubova, Stephan Günnemann, Cole Trapnell, David Lopez-Paz, and Fabian J Theis. Predicting cellular responses to complex perturbations in high-throughput screens. *Molecular Systems Biology*, 19:e11517, 2023.

Ashok Makkuva, Amirhossein Taghvaei, Sewoong Oh, and Jason Lee. Optimal transport mapping via input convex neural networks. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, pp. 6672–6681, 2020.

Leland McInnes and John Healy. UMAP: uniform manifold approximation and projection for dimension reduction. *CoRR*, abs/1802.03426, 2018.

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

Debarghya Mukherjee, Aritra Guha, Justin M. Solomon, Yuekai Sun, and Mikhail Yurochkin. Outlier-robust optimal transport. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, pp. 7850–7860, 2021.

Sloan Nietert, Ziv Goldfeld, and Rachel Cummings. Outlier-robust optimal transport: Duality, structure, and statistical analysis. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 11691–11719, 2022.

Geoffrey Schiebinger, Jian Shu, Marcin Tabaka, Brian Cleary, Vidya Subramanian, Aryeh Solomon, Joshua Gould, Siyan Liu, Stacie Lin, Peter Berube, Lia Lee, Jenny Chen, Justin Brumbaugh, Philippe Rigollet, Konrad Hochedlinger, Rudolf Jaenisch, Aviv Regev, and Eric S. Lander. Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming. *Cell*, 176(4):928–943, 2019. ISSN 0092-8674.

Liangliang Shi, Zhengyan Shi, and Junchi Yan. Selkd: Selective knowledge distillation via optimal transport perspective. In *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025.

Sanjay R Srivatsan, José L McFaline-Figueroa, Vijay Ramani, Lauren Saunders, Junyue Cao, Jonathan Packer, Hannah A Pliner, Dana L Jackson, Riza M Daza, Lena Christiansen, et al. Massively multiplex chemical transcriptomics at single-cell resolution. *Science*, 367(6473):45–51, 2020.

Vasileios Stathias, Anna M Jermakowicz, Marie E Maloof, Michele Forlin, Winston Walters, Robert K Suter, Michael A Durante, Sion L Williams, J William Harbour, Claude-Henry Volmar, et al. Drug and disease signature integration identifies synergistic combinations in glioblastoma. *Nature communications*, 9(1):5315, 2018.

Amos Tanay and Aviv Regev. Scaling single-cell genomics from phenomenology to mechanism. *Nature*, 541:331–338, 2017.

Cédric Villani. Topics in optimal transportation. 2003.

Jian Wang, Tianhong Dai, Bingfeng Zhang, Siyue Yu, Eng Gee Lim, and Jimin Xiao. POT: prototypical optimal transport for weakly supervised semantic segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15055–15064, 2025.

Jingjing Xu, Hao Zhou, Chun Gan, Zaixiang Zheng, and Lei Li. Vocabulary learning via optimal transport for neural machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 7361–7373, 2021.

Pei Yang, Yuhang Zhuang, and Qi Tan. Elastic optimal transport: Theory, application, and empirical evaluation. In *The Fourteenth International Conference on Learning Representations (ICLR)*, 2026.

APPENDIX

A RELATED WORK

We review the related work on optimal transport and single-cell perturbation prediction.

A.1 OPTIMAL TRANSPORT

Optimal transport (Villani, 2003) is a mathematical framework for comparing and aligning probability distributions by moving mass from a source to a target at minimal cost. Originating from Monge's sandpile problem (Monge, 1781) and grounded in Kantorovich's duality theory (Kantorovitch, 1942), it induces a geometrically meaningful distance between distributions. With scalable approximate solvers (Cuturi, 2013a) and neural parameterizations, OT has been playing a core role in artificial intelligence.

Classical OT requires full mass conservation, making it sensitive to outliers and distributional shifts. Extensions such as partial, unbalanced, and elastic OT relax this constraint to enable fixed, soft, or adaptive mass transport. Partial optimal transport (Caffarelli & McCann, 2010; Figalli, 2010) relaxes the mass equality constraint by fixing a predetermined amount of mass to be transported. Unbalanced optimal transport (Liero et al., 2017) removes the mass conservation constraint entirely via divergence penalties. Robust optimal transport (Nietert et al., 2022; Mukherjee et al., 2021) is similar to unbalanced optimal transport in using the soft marginal constraints measured by f-divergence, while emphasizing handling the probability distributions possibly corrupted by outliers. ELOT (Yang et al., 2026) is a discrete elastic optimal transport framework that uses inequality mass constraints and mixed-sign costs to adaptively transport mass, overcoming full- or fixed-mass limitations for robust distribution alignment.

Optimal transport has been employed in multiple fields of artificial intelligence. To name a few, its applications span generative adversarial networks (Arjovsky et al., 2017), computer vision (Wang et al., 2025), natural language processing (Xu et al., 2021), semi-supervised learning (Chapel & Alaya, 2020), domain adaptation (Courty et al., 2017), few-shot learning (Li et al., 2024), knowledge distillation (Shi et al., 2025), diffusion model (Carrasco et al., 2026), and large language models (Cui et al., 2025). The OT-based machine learning methods depend on classical optimal transport such as Kantorovich optimal transport (Kantorovitch, 1942) or partial optimal transport (Figalli, 2010) for distribution alignment. As mentioned before, they will also suffer from the limitations of full-mass or fixed-mass constraints.

NAOT is the continuous counterpart and theoretical generalization of discrete ELOT (Yang et al., 2026), providing the missing duality, Brenier-type theorems, and a scalable neural solver for adaptive-mass optimal transport.

A.2 SINGLE-CELL PERTURBATION PREDICTION

Several computational approaches have been developed for predicting single-cell perturbation responses. scGen (Lotfollahi et al., 2019) combined variational autoencoders and latent space vector arithmetic to predict single-cell perturbation responses. The conditional perturbation autoencoder (CPA) (Lotfollahi et al., 2023) incorporated interpretable linear models with a deep learning model to learn factorized latent representations of both perturbations and covariates via adversarial training. Most recently, OT has proven indispensable in biomedical research, where it enables the alignment of single-cell gene-expression profiles to uncover developmental trajectories (Schiebinger et al., 2019) and cellular perturbation responses (Srivatsan et al., 2020). CELLOT (Bunne et al., 2023) leveraged the neural optimal transport (Makkuva et al., 2020) to learn the response of individual cells to a given perturbation by mapping these unpaired distributions. Building upon CELLOT (Bunne et al., 2023), CondOT (Bunne et al., 2022) estimated a family of transport maps conditioned on a context variable, allowing for generalization to unseen contexts. GENOT (Klein et al., 2024) is a flexible neural optimal transport framework using entropic (Gromov) Wasserstein flow matching and quadratic solvers. It learns stochastic transport plans, supports arbitrary costs, and enables robust single-cell genomics applications including development, drug response, and cross-modality translation, demonstrating significant potential for precision medicine.

B PRELIMINARY

As for preliminary, we review several typical formulations of optimal transport.

Let μ and ν be the probability measures of the source and the target domains respectively. The probability measure μ with the probability masses $\{p_i\}_{i=1}^n$ on the locations $\{x_i\}_{i=1}^n$ is formulated as $\mu = \sum_{i=1}^n p_i \delta_{x_i}$ where δ_{x_i} is the Dirac function located at x_i, and p_i belongs to the probability simplex, i.e. $\sum_{i=1}^n p_i = 1$. Likewise, the probability measure ν with the probability masses $\{q_j\}_{j=1}^m$ on the locations $\{z_j\}_{j=1}^m$ is formulated as $\nu = \sum_{j=1}^m q_j \delta_{z_j}$. The transport plans are formulated by the probability measures γ, where γ_{ij} measures the amount of mass transported from location x_i to location z_j. The bounded ground cost $\mathcal{C}_{ij}$ tells how much it costs to transport one unit of mass from x_i to z_j. We denote $\mathbb{R}_+^{n\times m}$ and $\mathbb{R}_{\pm}^{n\times m}$ as the non-negative and mixed-sign matrix, respectively. The d-dimensional vector of ones is denoted by $\mathbf{1}_d$.

Kantorovich's Optimal Transport. In the discrete setting, Kantorovich's optimal transport problem (Kantorovitch, 1942) is formulated as:

$$\min_{\gamma\in\Pi(\mu,\nu)} < \gamma, \mathcal{C} > \tag{18}$$

where $\mathcal{C} \in \mathbb{R}_+^{n\times m}$ and

$$\Pi(\mu,\nu) = \{\gamma \in \mathbb{R}_+^{n\times m} \mid \gamma \mathbf{1}_m = \mu, \gamma^T \mathbf{1}_n = \nu\}. \tag{19}$$

However, classical optimal transport suffers from the rigid requirement of equal total mass, making it sensitive to outliers and unsuitable for comparing distributions with different total masses or noisy measurements. There have been two main directions including partial optimal transport and unbalanced optimal transport to attempt to remove the constraints of full-mass preservation.

Partial Optimal Transport. Partial optimal transport (Caffarelli & McCann, 2010; Figalli, 2010) relaxes the full-mass constraints in Kantorovich's problem and preserves the fixed amount of mass:

$$\min_{\gamma\in\Pi_p(\mu,\nu)} < \gamma, \mathcal{C} > \tag{20}$$

where $\mathcal{C} \in \mathbb{R}_+^{n\times m}$ and

$$\Pi_p(\mu,\nu) = \{\gamma \in \mathbb{R}_+^{n\times m} \mid \gamma \mathbf{1}_m \le \mu, \gamma^T \mathbf{1}_n \le \nu, \mathbf{1}_n^T \gamma \mathbf{1}_m = s\} \tag{21}$$

where $s(0 \le s \le min(\|\mu\|_1, \|\nu\|_1))$ is the fixed budget of mass, which is user-defined. $\|\cdot\|_1$ is the l_1 vector norm. However, the prespecified mass fraction is often arbitrary and problem-dependent, lacking a principled way to choose it adaptively.

Unbalanced Optimal Transport. Unbalanced optimal transport (Liero et al., 2017) relaxes the 'hard' marginal constraints in Kantorovich's problem with the 'soft' penalties by using some divergence $\mathcal{D}_\phi$:

$$\min_{\gamma\in\mathbb{R}_+^{n\times m}} < \gamma, \mathcal{C} > + \tau_1 \mathcal{D}_\phi(\gamma \mathbf{1}_m \| \mu) + \tau_2 \mathcal{D}_\phi(\gamma^T \mathbf{1}_n \| \nu) \tag{22}$$

where $\mathcal{C} \in \mathbb{R}_+^{n\times m}$. The trade-off coefficients τ_1 and τ_2 are used to adjust the weights of the marginal divergences. It is usually unknown to what extent the 'soft' penalties should be imposed on the divergence.

Elastic Optimal Transport. As a new member of optimal transport family, elastic optimal transport (Yang et al., 2026) is distinctive in adaptive-mass preservation, facilitating partial distribution alignment. Elastic optimal transport is formulated as:

$$\mathcal{W}(\mu,\nu) = \min_{\gamma\in\Pi_e(\mu,\nu)} < \gamma, \mathcal{C} > \tag{23}$$

where $\mathcal{C} \in \mathbb{R}_{\pm}^{n\times m}$ and

$$\Pi_e(\mu,\nu) = \{\gamma \in \mathbb{R}_+^{n\times m} \mid \gamma \mathbf{1}_m \le \mu, \gamma^T \mathbf{1}_n \le \nu\}. \tag{24}$$

C PROOFS OF THEOREMS

C.1 PROOF OF DUALITY THEOREM WITH GENERAL COST (THEOREM 1)

We derive the duality theory for neural adaptive optimal transport. Before going into the formal derivation, we first borrow the idea from the seminal work (Caffarelli & McCann, 2010) to reformulate the partial mass transport problem into the complete mass transport problem. In contrast, NAOT overcomes the drawback of (Caffarelli & McCann, 2010) which needs to specify the fixed budget of mass, and is able to automatically find the optimal mass to be transferred.

Augmented Problem. Let's attach an isolated point $\hat{\infty}$ to X and Z, denoted by $\hat{X} = X \cup \{\hat{\infty}\}$ and $\hat{Z} = Z \cup \{\hat{\infty}\}$, and extend the cost function

$$\hat{c}(x,z) = \begin{cases} c(x,z) & if\ x \neq \hat{\infty}\ and\ z \neq \hat{\infty} \\ 0 & otherwise \end{cases}. \tag{25}$$

Extend the measures $d\mu(x) = f(x)dx$ and $d\nu(z) = g(z)dz$ to $\hat{X}$ and $\hat{Z}$ by adding a Dirac mass isolated at infinity

$$\hat{\mu} = \mu + \|g\|_{L^1}\delta_{\hat{\infty}} \tag{26}$$

$$\hat{\nu} = \nu + \|f\|_{L^1}\delta_{\hat{\infty}} \tag{27}$$

where δ is the Dirac function and $\|\cdot\|_{L^1}$ is the L^1 norm. A bijection between $\gamma \in \Gamma_{\leq}(\mu,\nu)$ and $\hat{\gamma} \in \Gamma(\hat{\mu},\hat{\nu})$ is given by

$$\hat{\gamma} = \gamma + (f - f_\gamma) \otimes \delta_{\hat{\infty}} + \delta_{\hat{\infty}} \otimes (g - g_\gamma) + m_\gamma \delta_{(\hat{\infty},\hat{\infty})} \tag{28}$$

where f_γ and g_γ represent the left and right marginals of γ respectively, and $\otimes$ is the Kronecker product. Due to mass conservation, we have $\|f_\gamma\|_{L^1} = \|g_\gamma\|_{L^1}$ and $m_\gamma = \gamma[X \times Z]$.

For the primary problem defined in Equation 1, its augmented problem is formulated as

$$\min_{\hat{\gamma}\in\Gamma(\hat{\mu},\hat{\nu})} \int_{\hat{X}\times\hat{Z}} \hat{c}(x,z)d\hat{\gamma}(x,z) \tag{29}$$

where the set of admissible transport plans is denoted by

$$\Gamma(\hat{\mu},\hat{\nu}) = \left\{ \hat{\gamma} \in P(\hat{X}\times\hat{Z}) \;\middle|\; \begin{array}{l} \hat{\gamma}[A\times\hat{Z}] = \hat{\mu}[A] \\ \hat{\gamma}[\hat{X}\times B] = \hat{\nu}[B] \end{array} \right\} \tag{30}$$

for all Borel subsets $A \subset \hat{X}$ and $B \subset \hat{Z}$.

The proof of Theorem 1 is as follows:

Proof. First, from the bijection between γ and $\hat{\gamma}$, it is easy to see that the primary elastic optimal transport problem is equivalent to the augmented problem, i.e.,

$$\min_{\gamma\in\Gamma_{\leq}(\mu,\nu)} \int_{X\times Z} c(x,z)d\gamma(x,z) = \min_{\hat{\gamma}\in\Gamma(\hat{\mu},\hat{\nu})} \int_{\hat{X}\times\hat{Z}} \hat{c}(x,z)d\hat{\gamma}(x,z) \tag{31}$$

Second, for the full-mass optimal transport, by Kantorovich duality (Kantorovitch, 1942), we can prove the equivalence between the augmented problem and its Kantorovich dual problem, i.e.,

$$\min_{\hat{\gamma}\in\Gamma(\hat{\mu},\hat{\nu})} \int_{\hat{X}\times\hat{Z}} \hat{c}(x,z)d\hat{\gamma}(x,z) = \max_{\hat{\phi}(x)+\hat{\psi}(z)\leq\hat{c}(x,z)} \int_{\hat{X}} \hat{\phi}(x)d\hat{\mu}(x) + \int_{\hat{Z}} \hat{\psi}(z)d\hat{\nu}(z) \tag{32}$$

where $\hat{\phi}(x) \in C_b(\hat{X})$ and $\hat{\psi}(z) \in C_b(\hat{Z})$ are bounded continuous functions. It is similar to the proof of standard Kantorovich duality (please refer to Section 1.1.5 of (Villani, 2003) for details). The basic idea is to remove the constraints by using the generalized Lagrange multiplier method. Note that as shown in the right-hand of Equation 32, the duality of the augmented problem requires that $\hat{\phi}(x) + \hat{\psi}(z) \leq \hat{c}(x,z)$, i.e., $\hat{\phi}(x) + \hat{\psi}(z) \leq c(x,z)$ if $x \neq \hat{\infty}$ and $z \neq \hat{\infty}$ as usually, and $\hat{\phi}(x) + \hat{\psi}(z) \leq 0$ otherwise to accommodate the isolated point ∞.

Third, it remains to prove the equivalence between the duality of the primary optimal transport problem and the duality of its augmented problem, i.e.,

$$\max_{\hat{\phi}(x)+\hat{\psi}(z)\leq\hat{c}(x,z)} \int_{\hat{X}} \hat{\phi}(x)d\hat{\mu}(x) + \int_{\hat{Z}} \hat{\psi}(z)d\hat{\nu}(z) = \max_{\substack{\phi(x)+\psi(z)\leq c(x,z)\\ \phi(x),\psi(z)\leq 0}} \int_X \phi(x)d\mu(x) + \int_Z \psi(z)d\nu(z) \tag{33}$$

where $\phi(x) \in C_b(X)$ and $\psi(z) \in C_b(Z)$ are bounded continuous functions.

For Equation 33, any competitor (ϕ, ψ) in the right-hand can be extended to $\hat{X}$ and $\hat{Z}$ by taking $\phi(\hat{\infty}) = \psi(\hat{\infty}) = 0$. This extension still satisfied the constraints of the left-hand side. The maximization of the left-hand side of Equation 33 over larger class of competitors can only dominate the maximization of its right-hand side.

For the left-hand side of Equation 33, since

$$\max \int_{\hat{X}} (\hat{\phi} + k)d\hat{\mu} + \int_{\hat{Z}} (\hat{\psi} - k)d\hat{\nu} = \max \int_{\hat{X}} \hat{\phi}d\hat{\mu} + \int_{\hat{Z}} \hat{\psi}d\hat{\nu}$$

for any $k \in \mathbb{R}$, we are free to assume $\hat{\phi}(\hat{\infty}) = 0$. By the constraint $\hat{\phi}(\hat{\infty}) + \hat{\psi}(z) \leq \hat{c}(\hat{\infty}, z) = 0$ for any $z \in \hat{Z}$, we obtain $\hat{\psi}(z) \leq 0$ throughout $z \in \hat{Z}$. At $z = \hat{\infty}$, the only constraint is that

$$\hat{\psi}(\hat{\infty}) \leq \inf_{x\in\hat{X}} \hat{c}(x, \hat{\infty}) - \hat{\phi}(x) = \inf_{x\in\hat{X}} -\hat{\phi}(x) = -\phi_{max}$$

and the equality can be assumed to hold for the optimizer $(\hat{\phi}, \hat{\psi})$ of the left-hand side of Equation 33. Thus

$$\begin{aligned}
&\int_{\hat{X}} \hat{\phi}d\hat{\mu} + \int_{\hat{Z}} \hat{\psi}d\hat{\nu} \\
=&\int_X \hat{\phi}d\mu + \hat{\phi}(\hat{\infty})\|g\|_{L^1} + \int_Z \hat{\psi}d\nu + \hat{\psi}(\hat{\infty})\|f\|_{L^1} \\
=&\int_Z \hat{\psi}d\nu + \int_X \hat{\phi}d\mu - \phi_{max}\|f\|_{L^1}
\end{aligned}$$

The sum of the last two terms is not positive since $\phi_{max} \geq \hat{\phi}(\hat{\infty}) = 0$. Replacing ϕ by $min\{\phi, 0\}$ pointwise always increases the above objective since $\phi - \phi_{max} \leq min\{\phi, 0\}$, and makes it easier to satisfy the constraints of the right-hand side of Equation 33. Therefore, we conclude $\phi_{max} = 0$. Thus

$$\int_{\hat{X}} \hat{\phi}d\hat{\mu} + \int_{\hat{Z}} \hat{\psi}d\hat{\nu} = \int_X \hat{\phi}d\mu + \int_Z \hat{\psi}d\nu$$

By now, we have shown that $\hat{\phi}(x) \leq 0(\forall x \in \hat{X})$ and $\hat{\psi}(z) \leq 0(\forall z \in \hat{Z})$. That is to say, the restriction (ϕ, ψ) of $(\hat{\phi}, \hat{\psi})$ to $X \times Z$ now satisfies the constraint of the right-hand side of Equation 33. Therefore, for Equation 33, the maximization of the right-hand side dominates the maximization of the left-hand size. Hence the two maximum values coincide, which completes the proof. □

The c-transform and $\bar{c}$-transform are defined as

$$\phi^c(z) = \inf_{x\in X} c(x, z) - \phi(x) \tag{34}$$

$$\psi^{\bar{c}}(x) = \inf_{z\in Z} c(x, z) - \psi(z) \tag{35}$$

Using c-transform and $\bar{c}$-transform, one can reformulate the duality of the elastic optimal transport problem over two potentials as an convex program over a single potential

$$\begin{aligned}
&\max_{\substack{\phi(x)+\psi(z)\leq c(x,z)\\ \phi(x),\psi(z)\leq 0}} \int_X \phi(x)d\mu(x) + \int_Z \psi(z)d\nu(z) \\
=&\max_{\phi(x),\phi^c(z)\leq 0} \int_X \phi(x)d\mu(x) + \int_Z \phi^c(z)d\nu(z) \\
=&\max_{\psi(z),\psi^{\bar{c}}(x)\leq 0} \int_X \psi^{\bar{c}}(x)d\mu(x) + \int_Z \psi(z)d\nu(z)
\end{aligned} \tag{36}$$

C.2 Proof of Duality Theory with Quadratic Cost (Theorem 2)

Proof. According to Theorem 1, we have

$$\begin{aligned}&\min_{\gamma\in\Gamma_{\leq}(\mu,\nu)}\int_{X\times Z}\left[\frac{1}{2}\|x-z\|^2-\alpha\langle Mx,Mz\rangle\right]d\gamma(x,z)\\=&\max_{\substack{\phi(x)+\psi(z)\leq\frac{1}{2}\|x-z\|^2-\alpha\langle Mx,Mz\rangle\\ \phi(x),\psi(z)\leq 0}}\int_X\phi(x)d\mu(x)+\int_Z\psi(z)d\nu(z)\end{aligned}$$

where $\phi(x)\in\mathcal{C}_b(X)$ and $\psi(z)\in\mathcal{C}_b(Z)$ are bounded continuous functions. The constraint in the above dual problem can be formulated as

$$\begin{aligned}&\phi(x)+\psi(z)\leq\frac{1}{2}\|x-z\|^2-\alpha\langle Mx,Mz\rangle\\\Leftrightarrow&\left[\frac{1}{2}\|x\|^2-\phi(x)\right]+\left[\frac{1}{2}\|z\|^2-\psi(z)\right]\geq x^T(I+\alpha M^TM)z.\end{aligned}$$

Reparameterizing $\frac{1}{2}\|x\|^2-\phi(x)$ and $\frac{1}{2}\|z\|^2-\psi(z)$ by $\phi(x)$ and $\psi(z)$ respectively in the dual problem and ignoring the constant term independent of (ϕ,ψ) yields

$$\min_{\substack{\phi(x)+\psi(z)\geq x^TQz\\ \phi(x)\geq\frac{1}{2}\|x\|^2,\psi(z)\geq\frac{1}{2}\|z\|^2}}\int_X\phi(x)d\mu(x)+\int_Z\psi(z)d\nu(z)$$

which completes the proof. □

C.3 Proof of Optimal Transport Map (Theorem 3)

Proof. According to Theorem 2, the dual form of NAOT with quadratic cost must satisfy the constraint $\phi(x)+\psi(z)\geq x^T\big(I+\alpha M^TM\big)z$ for $\forall x\in X$ and $\forall z\in Z$. Hence, we have

$$\psi(z)=\phi^*(z)=\sup_{x\in X}x^T\big(I+\alpha M^TM\big)z-\phi(x)\tag{37}$$

$$\phi(x)=\psi^*(x)=\sup_{z\in Z}x^T\big(I+\alpha M^TM\big)z-\psi(z).\tag{38}$$

Therefore, $\phi(x)$ and $\psi(z)$ are conjugates of each other.

Fixing x, $x^T\big(I+\alpha M^TM\big)z-\phi(x)$ is a linear function, hence $\phi^*(z)$ is the upper envelope of a group of linear functions, and thus is a convex function with Lipschitz condition since the gradient of each linear function is $\big(I+\alpha M^TM\big)x$ which is bounded. Likewise, $\psi^*(x)$ is also a convex function.

For any pair in the support of γ, i.e., $(x,z)\in spt(\gamma)$, the gradient of the right side of Equation 37 with regard to x vanished, i.e.,

$$\begin{aligned}&\nabla_x\left[x^T\big(I+\alpha M^TM\big)z-\phi(x)\right]=0\\\Leftrightarrow&\nabla\phi(x)=\nabla_x\left[x^T\big(I+\alpha M^TM\big)z\right]=\big(I+\alpha M^TM\big)z\end{aligned}$$

which implies that

$$z=T(x)=\big(I+\alpha M^TM\big)^{-1}\cdot\nabla\phi(x)=\big(I+\alpha M^TM\big)^{-1}\cdot\nabla\psi^*(x).$$

It shows that the optimal transport map T from μ to ν is the gradient of a convex function, the so-called Brenier potential. □

C.4 Proof of NAOT over Convex Functions (Theorem 4)

Proof. According to Theorem 3, for $(x,z)\in spt(\gamma)$ we have

$$x=\big(I+\alpha M^TM\big)^{-1}\cdot\nabla\phi^*(z).$$

Substituting it into the conjugate of $\phi(x)$ yields

$$\begin{aligned}\phi^*(z) &= \sup_{x\in X} x^T\big(I+\alpha M^TM\big)z - \phi(x)\\ &= \Big[\big(I+\alpha M^TM\big)^{-1}\nabla\phi^*(z)\Big]^T\big(I+\alpha M^TM\big)z\\ &\quad - \phi\Big[\big(I+\alpha M^TM\big)^{-1}\nabla\phi^*(z)\Big]\\ &= \big\langle z, \nabla\phi^*(z)\big\rangle - \phi\big(Q^{-1}\nabla\phi^*(z)\big)\end{aligned}$$

where $Q = I+\alpha M^TM$. Hence, for all convex function $\psi(z)\in\mathcal{F}(Z)$, by the definition of conjugate function we have

$$\big\langle z, \nabla\psi(z)\big\rangle - \phi\big(Q^{-1}\nabla\psi(z)\big) \le \big\langle z, \nabla\phi^*(z)\big\rangle - \phi\big(Q^{-1}\nabla\phi^*(z)\big)$$

for $\forall z\in Z$ such that $\psi(z)$ and $\phi^*(z)$ are differential at z. Since both convex function $\psi(z)$ and $\phi^*(z)$ are differential almost everywhere, for any convex function $\psi(z)\in\mathcal{F}(Z)$, we have

$$\begin{aligned}&\int_Z\Big[\big\langle z, \nabla\psi(z)\big\rangle - \phi\big(Q^{-1}\nabla\psi(z)\big)\Big]d\nu(z)\\ \le&\int_Z\Big[\big\langle z, \nabla\phi^*(z)\big\rangle - \phi\big(Q^{-1}\nabla\phi^*(z)\big)\Big]d\nu(z)\\ =&\int_Z\phi^*(z)d\nu(z)\end{aligned}$$

and the equality is achieved with $\psi=\phi^*$. Hence, we have

$$\int_Z\phi^*(z)d\nu(z) = \max_{\psi(z)\in\mathcal{F}(Z)}\int_Z\Big[\big\langle z, \nabla\psi(z)\big\rangle - \phi\big(Q^{-1}\nabla\psi(z)\big)\Big]d\nu(z).$$

Therefore, we have

$$\begin{aligned}&\min_{\substack{\phi(x)+\psi(z)\ge x^TQz\\ \phi(x)\in\mathcal{C}_b(X),\phi^*(z)\in\mathcal{C}_b(Z)}}\int_X\phi(x)d\mu(x)+\int_Z\psi(z)d\nu(z)\\ =&\min_{\substack{\phi(x)\in\mathcal{C}_b(X)\\ \phi^*(z)\in\mathcal{C}_b(Z)}}\int_X\phi(x)d\mu(x)+\int_Z\phi^*(z)d\nu(z)\\ =&\min_{\phi(x)\in\mathcal{F}(X)}\max_{\psi(z)\in\mathcal{F}(Z)}\int_X\phi(x)d\mu(x)\\ &+\int_Z\Big[\big\langle z, \nabla\psi(z)\big\rangle - \phi\big(Q^{-1}\nabla\psi(z)\big)\Big]d\nu(z).\end{aligned}$$

By adding the constraints $\phi(x)\ge\frac{1}{2}\|x\|^2, \psi(z)\ge\frac{1}{2}\|z\|^2$, it completes the proof. □

D NAOT ALGORITHM

The training procedure of NAOT is proposed in Algorithm 1. It aims to learn a transport map from a source control cell population to a target perturbed cell population. The algorithm begins by initializing two partially input convex neural networks, PICNN$_\phi$ and PICNN$_\psi$. It then runs a main loop for K_{max} iterations. In each iteration, a batch of control cells is sampled from the source measure. Next, an inner loop runs for K_ψ steps, sampling a batch of perturbed cells and updating PICNN$_\psi$ to maximize an objective function. Following the inner loop, PICNN$_\phi$ is updated to minimize the same objective. This min-max optimization process learns the optimal transport map.

NAOT formulates adaptive optimal transport as a convex-concave min-max problem over Brenier potentials, parameterized by input-convex neural networks. It restricts the search to compact convex function classes and avoids costly per-sample conjugate evaluations required by classical dual solvers, thus scaling to large datasets.

Algorithm 1 NAOT for Single-cell Perturbation Prediction

Input: Source measure of the control cell population $\mu = \frac{1}{n}\sum_{i=1}^{n}\delta_{x_i}$ and target measure of the perturbed cell population $\nu = \frac{1}{m}\sum_{j=1}^{m}\delta_{z_j}$, context variable c, trade-off parameters σ and α, K_ψ and K_{max}.
Output: Transport map T from source measure to target measure.

```
Initialize both PICNN_φ and PICNN_ψ.
for k = 1 to K_max do
    Sample a batch of control cells from the source measure.
    for k_ψ = 1 to K_ψ do
        Sample a batch of perturbed cells from the target measure.
        Update PICNN_ψ to maximize the objective defined in 17.
    end for
    Update PICNN_φ to minimize the objective defined in 17.
end for
```

E More Experimental Results

E.1 Experiment Setup

We detail the experimental setup including datasets, metrics, baselines, networks and hyperparameters.

E.1.1 Dataset

The sciPlex3 perturbation dataset (Srivatsan et al., 2020) is curated for modeling cellular perturbation responses. It profiles global transcriptional responses to thousands of independent perturbations at single-cell resolution via scRNA-seq techniques. The raw sciPlex3 dataset is accessible in the Gene Expression Omnibus (GEO) database under accession number GSM4150378 and the pre-processed version can be downloaded from github.com/bunnech/cellot. The pre-processed sciPlex3 dataset includes measurements from 762,039 cells across the top 1,000 highly variable genes, spanning three human cancer cell lines. Of these, 744,474 cells were treated with single compounds from a total of 188 drugs at four different dosages (10 nM, 100 nM, 1,000 nM, and 10,000 nM), while the remaining 17,565 cells served as untreated controls. For feature selection, we follow CondOT (Bunne et al., 2022) and select the top 50 marker genes which show strong differences between perturbed and unperturbed states for the evaluations.

E.1.2 Evaluation Metrics

Due to the lack of ground-truth pairs of control and perturbed cells on the single-cell level, it is common to assess the model performance on the distribution level via the discrepancy metrics such as Wasserstein distance, maximum mean discrepancy, perturbation signatures, and Fréchet inception distance.

Let $\mu = \frac{1}{n}\sum_{i=1}^{n}\delta_{x_i}$ represent the measure of the control cell population, $\nu = \frac{1}{m}\sum_{j=1}^{m}\delta_{z_j}$ the measure of the observed perturbed cell population, and $\bar{\nu} = \frac{1}{n}\sum_{i=1}^{n}\delta_{\bar{z}_i}$ the measure of the predicted perturbed cell population, where δ is the Dirac function. The following distributional metrics are used to measure the discrepancy between $\bar{\nu}$ and ν.

Maximum Mean Discrepancy (MMD) is a non-parametric measure that captures the difference between two distributions in a reproducing kernel Hilbert space (Gretton et al., 2012). The squared MMD between the predicted perturbation measure $\bar{\nu}$ and the observed perturbation measure ν is defined as:

$$\mathrm{MMD}(\bar{\nu}, \nu; \omega) = \mathbb{E}_{\bar{z}_i, \bar{z}_j}[\omega(\bar{z}_i, \bar{z}_j)] + \mathbb{E}_{z_i, z_j}[\omega(z_i, z_j)] - 2\mathbb{E}_{\bar{z}_i, z_j}[\omega(\bar{z}_i, z_j)] \tag{39}$$

where ω is the kernel function. Following CondOT (Bunne et al., 2022), we use the radial basis function kernel, and report the averaged MMD over kernel bandwidths including 0.5, 0.1, 0.01, and 0.005.

Wasserstein Distance is used to quantify the divergence between the predicted perturbation measure $\bar{\nu}$ and the observed perturbation measure ν. For efficient computation, we use the entropy-

regularized Wasserstein distance (Cuturi, 2013b) defined as:

$$\mathrm{WD}(\bar{\nu}, \nu) = \min_{\gamma \in \Gamma(\bar{\nu}, \nu)} \sum_{ij} \gamma_{ij} \|\bar{z}_i - z_j\|^2 + \epsilon \sum_{ij} \gamma_{ij} (\log \gamma_{ij} - 1) \tag{40}$$

where $\Gamma(\bar{\nu}, \nu) = \{\gamma \mid \gamma \mathbf{1}_m = \bar{\nu}, \gamma^T \mathbf{1}_n = \nu\}$ represents the set of admissible transport plans that map $\bar{\nu}$ to ν. The transport plan γ is a probability measure whose left and right marginals are equal to $\bar{\nu}$ and ν respectively. Following CondOT (Bunne et al., 2022), we set $\epsilon = 0.1$ for all experiments.

Perturbation Signatures (PS) is popular for measuring the effect of the perturbation on populations (Stathias et al., 2018). The perturbation signature between the control measure μ and the observed perturbed measure ν is defined as:

$$\mathrm{PS}(\nu, \mu) = \frac{1}{m} \sum_{z_j \in \nu} z_j - \frac{1}{n} \sum_{x_i \in \mu} x_i. \tag{41}$$

Then, the $\ell_2(PS)$ between the predicted perturbation measure $\bar{\nu}$ and the observed perturbation measure ν is defined as $\ell_2(\mathrm{PS})(\bar{\nu}, \nu) = \|\mathrm{PS}(\bar{\nu}, \mu) - \mathrm{PS}(\nu, \mu)\|_2$.

Fréchet Inception Distance (FID) is widely adopted in generative modeling to assess the quality of generated samples by comparing their distribution to that of real data (Heusel et al., 2017). The FID between $\bar{\nu}$ and ν is computed as:

$$\mathrm{FID} = \Big\| \frac{1}{m} \sum_{z_j \in \nu} z_j - \frac{1}{n} \sum_{\bar{z}_i \in \bar{\nu}} \bar{z}_i \Big\|^2 + \mathrm{Tr}\Big(\Sigma_\nu + \Sigma_{\bar{\nu}} - 2(\Sigma_\nu \Sigma_{\bar{\nu}})^{\frac{1}{2}} \Big) \tag{42}$$

where Σ_ν and $\Sigma_{\bar{\nu}}$ represent covariance matrices for the target perturbation distribution ν and the predicted perturbation distribution $\bar{\nu}$, respectively. Lower FID scores indicate that the predicted data is more similar to the observed data, both in terms of feature distribution and diversity.

Using various distributional metrics, we are able to comprehensively assess NAOT's predictive performance. However, computing these measurements directly on the whole populations has the drawback of neglecting the subpopulation structures. Instead, we use their conditioned versions which consider both whole-population distributions and the subpopulation distributions across different cell lines. Specifically, the cell populations are divided into several subpopulations according to the cell lines, then the metrics are applied to the subpopulations with the same cell line and the expectation over all cell lines is reported as the final distributional distance.

To exclude the influence of baseline differences among different cell lines, all metrics are first computed independently within each cell line, and then the results from the three cell lines are averaged to obtain the final reported values.

E.1.3 Comparison Baselines

We compare NAOT with three closely related methods: CELLOT (Bunne et al., 2023), CondOT (Bunne et al., 2022), and GENOT (Klein et al., 2024).

CELLOT (Bunne et al., 2023) is a pioneering model that employed neural optimal transport to predict cellular responses to perturbations at the single-cell level. CELLOT formulated the task of learning perturbation responses as an optimal transport problem, and used two standard input convex neural networks (Amos et al., 2017; Makkuva et al., 2020) to learn the transport map between control and perturbed distributions.

CondOT (Bunne et al., 2022) built upon CELLOT and achieved the state-of-the-art performance. It used partially input convex neural networks (Amos et al., 2017; Makkuva et al., 2020) to predict perturbation responses, conditioned on contexts such as drug dosage or cell line. This allows CondOT to incorporate the conditional inputs into learning a transport map between control and perturbed distributions and generalize to unseen conditions that were not present in the training data, expanding its applicability to more complex scenarios.

GENOT (Klein et al., 2024) learns flow matching using optimal transport couplings as supervisory signals, models perturbation responses through conditional probability distributions, and characterizes single-cell drug responses via stochastic generation. It allows unbalanced optimal transport to model cellular growth and death or account for outliers.

E.1.4 NETWORKS AND HYPERPARAMETERS

For the network architecture, we follow CondOT (Bunne et al., 2022) and implement the PICNN (Makkuva et al., 2020) architecture with four hidden layers and the number of hidden units in each layer is 64. The weights for both PICNN networks are initialized with uniform distribution between $[0, 0.1]$.

For hyperparameter selection, we primarily adhere to the validated settings from CondOT (Bunne et al., 2022). All networks are trained using the Adam optimizer (Kingma, 2014) with a learning rate of 0.0001, and momentum parameters $\beta_1 = 0.5$ and $\beta_2 = 0.9$, along with a non-negativity penalty coefficient $\lambda = 1$ for network parameters. The inner-loop iteration count in the NAOT algorithm is set to $K_\psi = 10$. Training is conducted over a total of $K_{max} = 20,000$ epochs, with performance metrics evaluated every 50 epochs. The trade-off coefficient α is optimized using the Optuna framework (Akiba et al., 2019), performing an automated search over the range [0.0, 1.0] with 20 trials. Since the cell line of each cell is known in our experiments, we use one parameterized fully connected layer to learn M by projecting the cell measurements from the original feature space to the label space.

For data splitting, we follow the methodology from CondOT, grouping data by drug type and partitioning each drug-specific subset into training and testing sets with an 80:20 split ratio.

E.2 VISUALIZATION OF CELL DISTRIBUTION ALIGNMENT

Through intuitive visualizations and case study, we qualitatively demonstrate that NAOT is able to align the control and perturbed cell distributions in the population, cell line, and gene levels, yielding accurate and nuanced predictions.

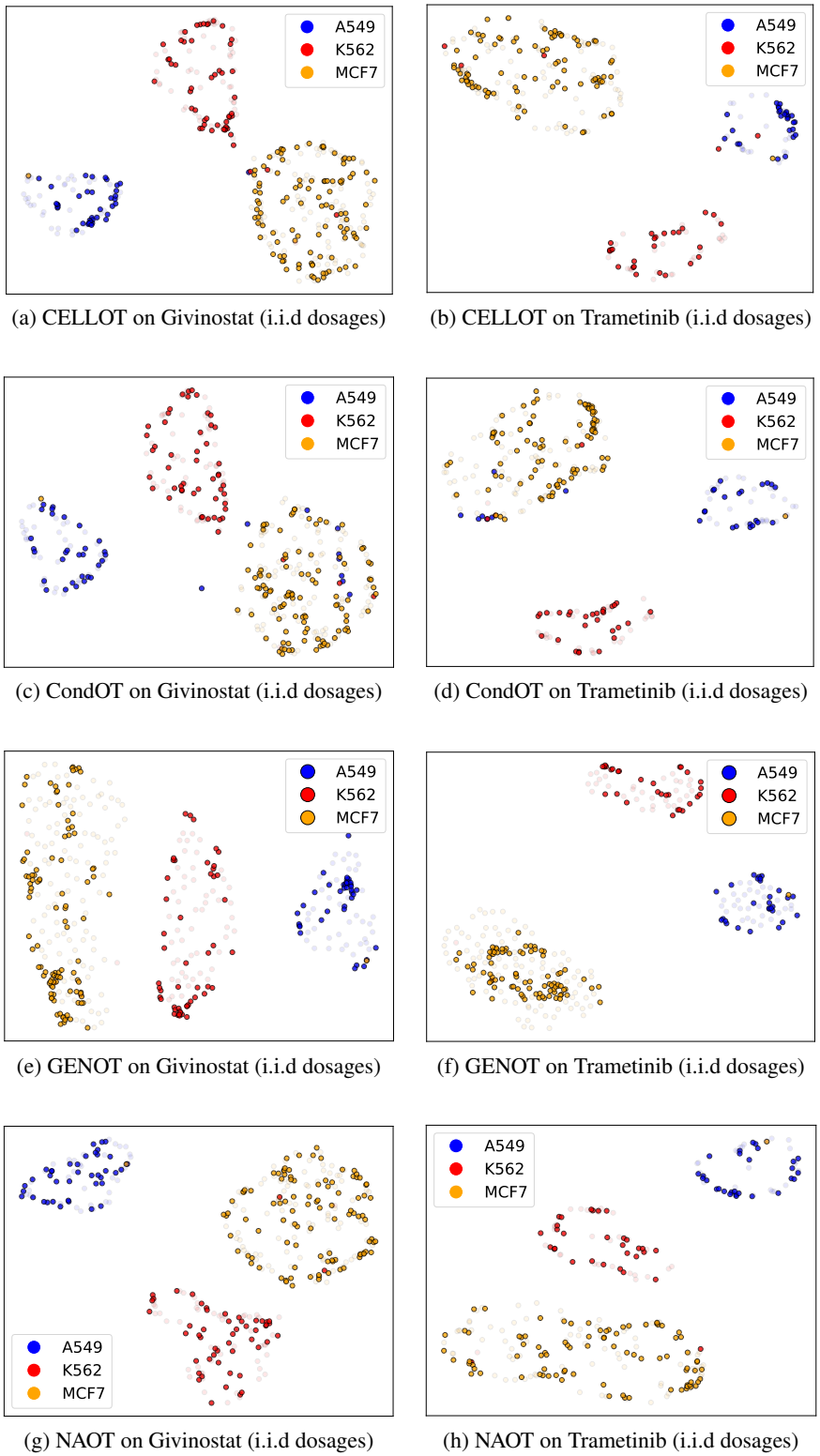
E.2.1 UMAP VISUALIZATION

We plot the uniform manifold approximation and projection (UMAP) (McInnes & Healy, 2018) embedding for both predicted and observed perturbed cells. Figure 2 and Figure 3 show the UMAP embeddings in the independent and identically distribution (seen dosages) and out-of-distribution (unseen dosages) settings respectively. We have two main observations from these figures. First, the perturbed cell distribution (in bright color) inferred by NAOT aligns well with the observed perturbed cell distribution (in grey color). Second, the boundary for cells in different cell lines is more clear than those of the comparison methods. In other word, NAOT simultaneously aligns the cells in both the population and subpopulation (cell line) level. In contrast, both CondOT and CELLOT are less successful in capturing the subpopulation distributions, and the cells in different cell lines may be mixed in the same cluster. In the out-of-distribution (Figure 3(e)-(f)), the data points predicted by GENOT are more obviously concentrated in the center of the distribution, indicating that the stochastic generation mechanism of GENOT mainly captures the average level of the distribution well, and its ability to precisely fit the distribution shape is insufficient.

E.2.2 MARGINAL DISTRIBUTION MATCHING

We go further into investigating how well NAOT aligns the marginal distributions of the predicted and observed target cells in the gene expression level. The marginal distributions for an exemplary gene across different cell lines are plotted in Figure 4, where x-axis denotes the expression level, and y-axis represents the density. As for the legends, 'source' and 'target' refer to the control and observed perturbed genes, while the others refer to predicted perturbed genes with the corresponding methods. By comparing the marginal distributions of the predicted and observed target cells on the individual gene, we observe that the gene of different cell lines shows highly various responses towards the cancer drug. This renders difficulty to the matching of marginal distributions.

The marginal distribution results show that NAOT more accurately captures the density shape of the true distribution in most cases, with both peak location and distribution width highly consistent with the true distribution, whereas CELLOT and CondOT exhibit obvious peak shifts in some cell lines. As for GENOT, its marginal distribution predictions deviate the farthest from the true gene expression marginal distributions among the four methods, with the most pronounced errors in both peak location and distribution width, further corroborating the limitations of GENOT in fine-grained distribution modeling.



(a) CELLOT on Givinostat (i.i.d dosages)

(b) CELLOT on Trametinib (i.i.d dosages)

(c) CondOT on Givinostat (i.i.d dosages)

(d) CondOT on Trametinib (i.i.d dosages)

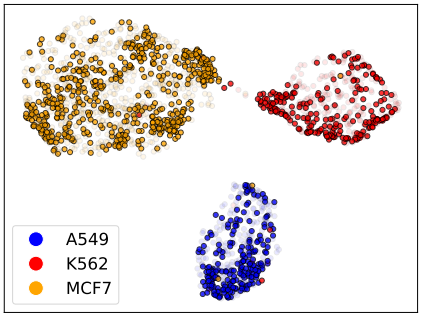
(e) GENOT on Givinostat (i.i.d dosages)

(f) GENOT on Trametinib (i.i.d dosages)

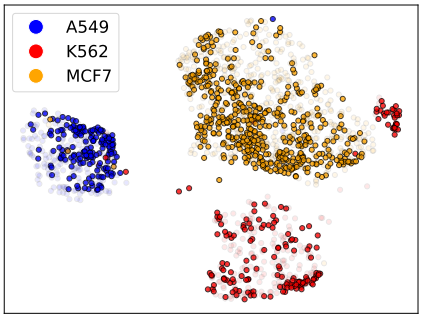
(g) NAOT on Givinostat (i.i.d dosages)

(h) NAOT on Trametinib (i.i.d dosages)

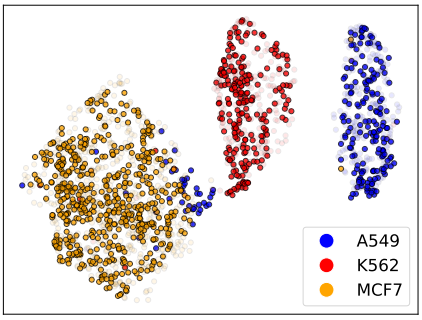
Figure 2: UMAP embeddings of observed target cells (in grey color) and perturbed cells (in bright color) predicted by CELLOT, CondOT, and NAOT in the independent and identically distribution (i.i.d) setting. The left and right panels correspond to the cancer drugs Givinostat and Trametinib respectively.



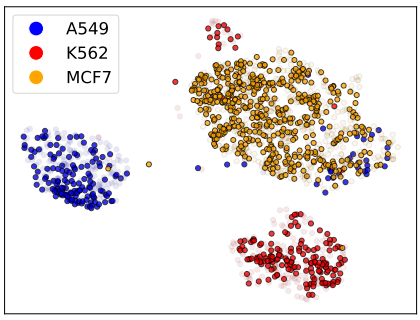
(a) CELLOT on Givinostat (OOD dosages)



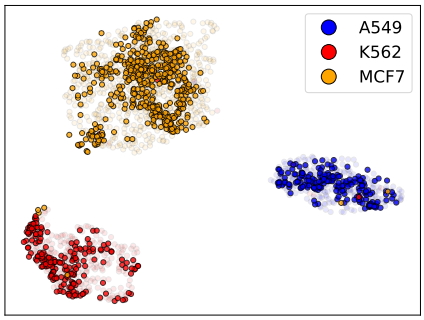
(b) CELLOT on Trametinib (OOD dosages)



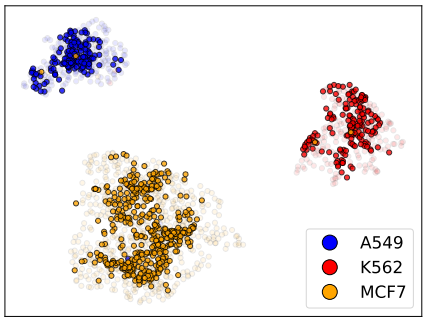
(c) CondOT on Givinostat (OOD dosages)



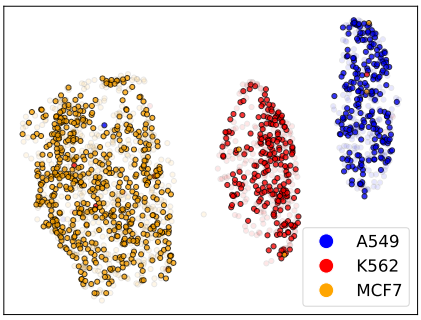
(d) CondOT on Trametinib (OOD dosages)



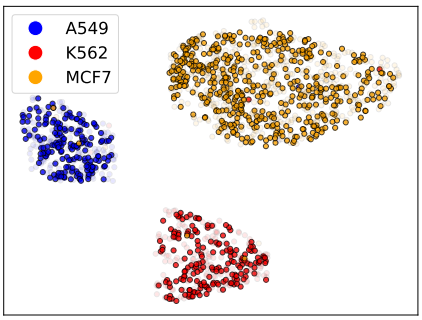
(e) GENOT on Givinostat (OOD dosages)



(f) GENOT on Trametinib (OOD dosages)

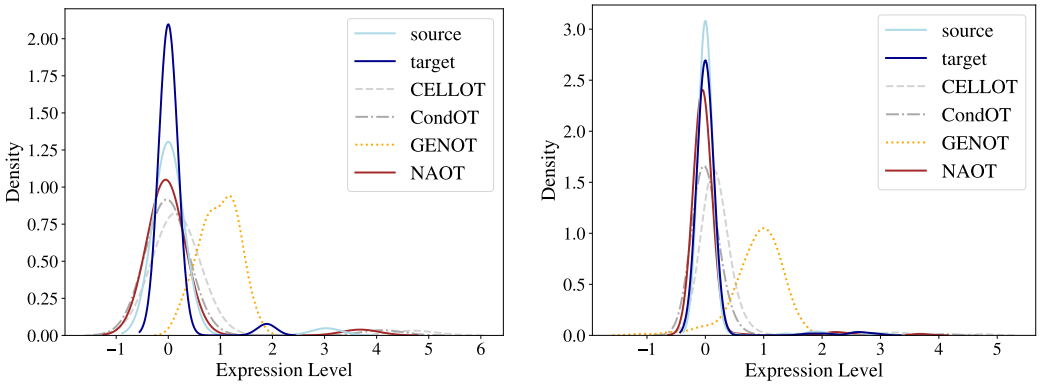


(g) NAOT on Givinostat (OOD dosages)



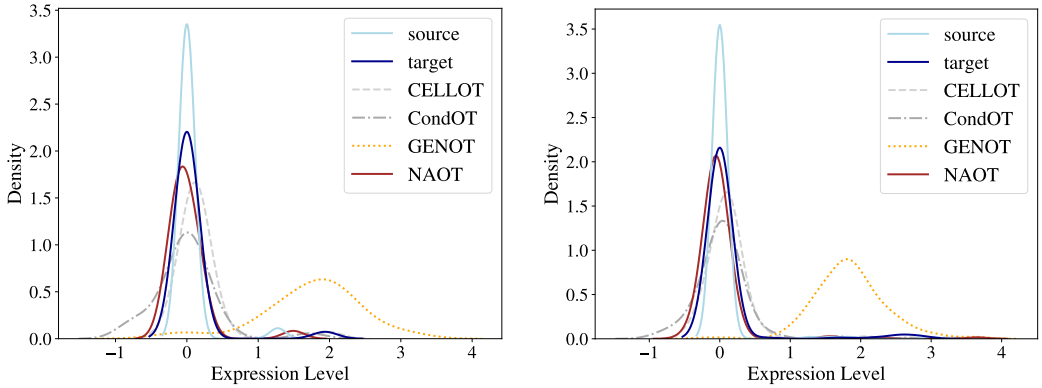
(h) NAOT on Trametinib (OOD dosages)

Figure 3: UMAP embeddings of observed target cells (in grey color) and perturbed cells (in bright color) predicted by CELLOT, CondOT, and NAOT in the out-of-distribution (OOD) setting. The left and right panels correspond to the cancer drugs Givinostat and Trametinib respectively.



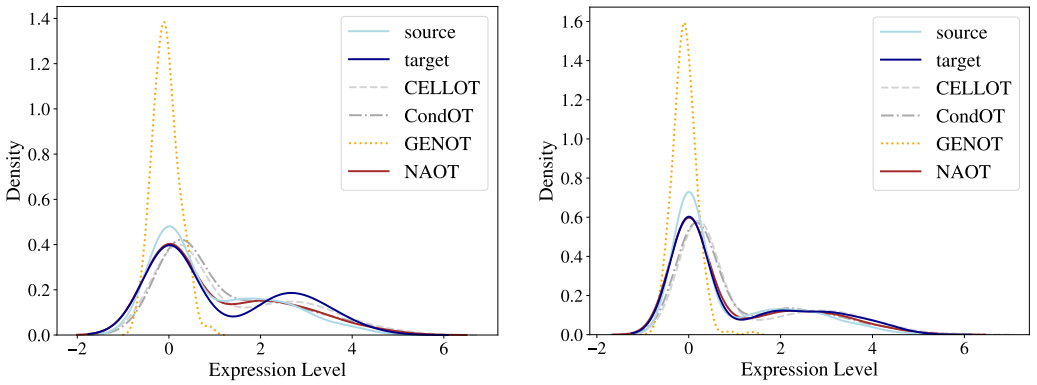
(a) A549 (i.i.d dosages)

(b) A549 (OOD dosages)



(c) K562 (i.i.d dosages)

(d) K562 (OOD dosages)



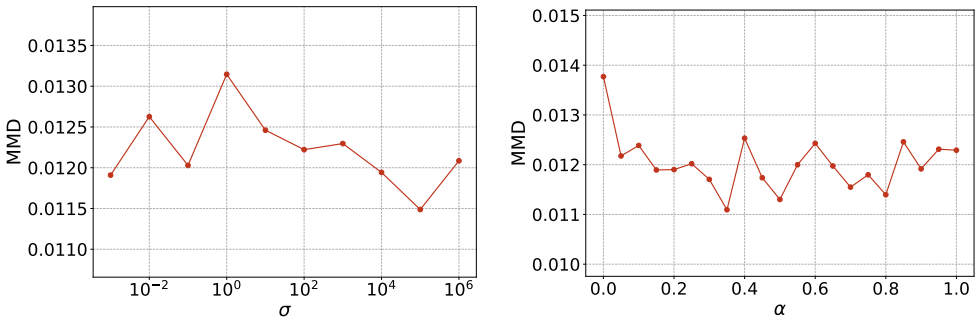
(e) MCF7 (i.i.d dosages)

(f) MCF7 (OOD dosages)

Figure 4: Marginal distributions of an exemplary gene ENSG00000115461.4 across different cell lines. The left and right panels correspond to settings where the drug dosages follow the independent identically distribution and out-of-distribution respectively.

E.3 Hyperparameter Sensitivity Analysis

NAOT involves two key hyperparameters: the convexity constraint penalty coefficient σ and the latent space similarity weight α. Figure 5 shows the trend of model performance as these two parameters vary.



(a) Sensitivity analysis of hyperparameter σ (b) Sensitivity analysis of hyperparameter α

Figure 5: Sensitivity analysis of the convexity constraint penalty coefficient σ and the latent space similarity weight α in NAOT

The model remains stable over a wide range of σ, indicating that the penalty-based convexity constraint works effectively. For α, when $\alpha = 0$, the model already outperforms the comparison methods, suggesting that the adaptive mass mechanism itself contributes positively. As α increases, model performance further improves and peaks near the optimal value, then declines slightly. This indicates that a moderate latent space similarity constraint can enhance biological consistency, but an overly strong constraint may suppress genuine differences in the expression space.