Quantum Theory and Application of Contextual Optimal Transport

Nicola Mariella[1,2,*] **Jannis Born**[2,*]

Albert Akhriev[1,2] **Francesco Tacchino**[1,2] **Christa Zoufal**[1,2] **Eugene Koskin**[3]

Ivano Tavernelli[1,2] **Stefan Woerner**[1,2] **Marianna Rapsomaniki**[2] **Sergiy Zhuk**[1,2]

[*]Shared first-authors: nicola.mariella@ibm.com, jab@zurich.ibm.com
[1]IBM QUANTUM
[2]IBM RESEARCH EUROPE
[3]UNIVERSITY COLLEGE DUBLIN

Abstract

Optimal Transport (OT) has fueled machine learning (ML) applications across various domains. In cases where paired data measurements $(\boldsymbol{\mu}, \boldsymbol{\nu})$ are coupled to a context variable $\boldsymbol{p}_i$, one may aspire to *learn* a global transportation map, parameterized through the context to facilitate prediction of target states even from unseen context. Existing approaches for this task leverage Brenier's theorem and utilize Neural OT. Here, we follow a radically different approach inspired by quantum computing principles to develop a Quantum formulation for learning transportation plans parameterized by a context variable. This is achieved through exploiting a natural link between doubly stochastic matrices and unitary operators. The latter can be directly related to recent results in quantum learning theory suggesting intrinsic advantages in modelling constrained problems with quantum methods. We verify our methodology on synthetic data, emulating the task of predicting single-cell perturbation responses parameterized through drug dosage as context. Our experimental comparisons to a baseline reveal that our method can capture dose-induced variations in cell distributions, even to some extent when extrapolating to dosages outside the interval seen during training. In summary, this work assesses the feasibility of learning to predict contextualized transportation plans through a novel quantum computing approach.

1 Introduction

Optimal transport (OT) [1] provides a mathematical framework for finding optimal transportation plans that minimize the cost of moving resources from a source to a target distribution. The cost is defined as a distance or a dissimilarity measure between the source and target points, and the optimal transport plan aims to minimize this cost while satisfying certain constraints. OT theory has found applications across several fields, including economics, statistics, biology, computer science, and image processing. In biology, OT has recently gained popularity in single-cell analysis, an area of research rich in problems of mapping cellular distributions across distinct states, timepoints, or spatial contexts [2]. Notable biological tasks solved using OT theory include reconstructing trajectories of cellular evolution [3], predicting cell responses to therapeutic interventions [4, 5], inferring spatial and signaling relationships between cells [6] and aligning datasets across different omic modalities [7, 8].

Accepted to the Optimal Transport and Machine Learning Workshop at NeurIPS 2023.

In these applications, the source and target distributions are measurements of biomolecules (e.g., mRNA, proteins) at a single-cell level, with or without spatial or temporal resolution [9].

In many recent OT applications, data measurements $\boldsymbol{\mu}_i$ and $\boldsymbol{\nu}_i$ are coupled to a context variable $\mathbf{p}_i$ that induced $\boldsymbol{\mu}_i$ to develop into $\boldsymbol{\nu}_i$. In such cases, one might aspire to *learn* a global transportation map T that is parameterized through $\mathbf{p}_i$ and thus facilitates the *prediction* of target states $\hat{\boldsymbol{\nu}}_j$ from source states $\boldsymbol{\mu}_j$, even for an unseen context variable $\mathbf{p}_j$. This line of work is largely based on Brenier's theorem [10], a seminal theorem in OT that postulated the existence of an optimal and unique OT map T given by the gradient of a convex function under certain conditions, i.e., $T = \nabla f_\theta$. As established by Makkuva et al. [11], OT maps between two distributions can be learned through neural OT solvers by using a minimax optimization where f_θ is an input convex neural network [12]. A notable example of such a neural OT approach is `CondOT` [5], an OT framework that estimates transportation maps conditioned on a context variable and learned from paired quasi-probability distributions $(\boldsymbol{\mu}_i, \boldsymbol{\nu}_i)$, each linked to a context variable $\mathbf{p}_j$[1]. However, one key limitation of neural OT approaches stems from the fact that Brenier's theorem only holds when the cost is the squared Euclidean distance [14], which is not always meaningful, especially for high-dimensional data.

On a separate realm, quantum computers offer a new computing paradigm with the potential to become practically useful in ML [15, 16, 17, 18, 19, 20] and fuel applications in, e.g., life sciences [21] or high-energy physics [22, 23]. Here, we propose a quantum contextual OT approach that is inspired by a previously unreported natural link between OT and unitary operators, a fundamental concept in quantum computing. This link relates to the structure of the OT maps and allows to turn the analytical problem of computing OT plans into a parameterizable approach to *estimate* transportation plans. In contrast to existing neural OT methods, our quantum formulation does not depend on Brenier's theorem and thus does not impose any restriction on the ground cost C. Furthermore, our approach directly estimates full transportation plans, with the advantage of increased interpretability, e.g., methods of homological algebra could be applied to study the topology of predicted maps.

The remainder of this paper commences with describing the contextual OT problem and then proceeds to lay the theoretical foundations of the quantum formulation, detailing the circuits and the ansatz[2] for encoding doubly stochastic matrices (DSMs) and transportation plans. As a proof-of-concept, we present an application of our methodology on synthetic data that describes the effect of drug dosage perturbation on the composition of different cellular populations. In our setup, we use drug dosage as a context variable $\boldsymbol{p}_i$ and instantiate $\boldsymbol{\mu}_i$ and $\boldsymbol{\nu}_i$ as distributions over cell types for a population of cells. As our work describes a fundamentally novel approach to learn OT maps, our objective is to assess the feasibility of learning to predict contextualized transportation plans through a quantum approach, and *not* (yet) to compete with established neural OT approaches for drug perturbation prediction.

2 Optimal Transport theory

2.1 Preliminaries

In the Kantorovich relaxation of the Monge problem [14], $C \in \mathbb{R}^{n\times m}$ is a non-negative matrix where $C_{i,j}$ corresponds to the cost of mass displacement from entity i to some other j. Let $\boldsymbol{\mu} \in \mathbb{R}^n_{++}$ and $\boldsymbol{\nu} \in \mathbb{R}^m_{++}$ be positive real vectors, representing the quasi-probability discrete distributions for the source and destination entities. The discrete (regularized) Kantorovich's OT problem is then defined as follows

$$\min_{Q\in\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\nu})} \operatorname{Tr}\left(QC^\top\right) + \gamma h(Q), \tag{1a}$$

$$\text{s.t.} \quad \mathcal{N}(\boldsymbol{\mu},\boldsymbol{\nu}) = \left\{Q \in \mathbb{R}^{n\times m}_+ \middle| Q\mathbf{1}_m = \boldsymbol{\mu}, Q^\top \mathbf{1}_n = \boldsymbol{\nu}\right\}, \tag{1b}$$

where h is a regularization function [24] with its trade-off $\gamma \geq 0$. The set $\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\nu})$ is known as the *transportation polytope* (convex) [25], and its elements are the *transportation maps*. Given a map $T \in \mathcal{N}(\boldsymbol{\mu},\boldsymbol{\nu})$, we interpret each entry $T_{i,j}$ as the mass moved from source i to destination j. In this work a special case of the transportation polytope will often emerge, that is the *Birkhoff polytope*, defined as $\Omega_n \coloneqq \mathcal{N}(\mathbf{1}_n, \mathbf{1}_n)$, with its elements called *doubly stochastic matrices* (DSMs).

[1] Throughout this work we will use the term "contextual" rather than "conditional" to differentiate from the OT problem of computing transportation plans between two conditional probabilities [13]

[2] In the context of quantum computing, an ansatz is a parametric quantum circuit used to approximate a quantum state related to the problem at hand.

2.2 The Contextual OT problem

Let $\mathcal{K}_d \coloneqq \Delta_d \cap \mathbb{R}^d_{++}$ denote the subset of the probability simplex with vectors presenting non-zero components[3]. We consider a dataset of contextualised measures each represented by a tuple $(\mathbf{p}_i, (\boldsymbol{\mu}_i, \boldsymbol{\nu}_i)) \in \mathcal{X} \times \mathcal{K}_d^2$, where $\mathbf{p}_i \in \mathcal{X} \subseteq \mathbb{R}^s$ defines the context. The initial and final states $\boldsymbol{\mu}_i$ and $\boldsymbol{\nu}_i$ are assumed from the same simplex $\mathcal{K}_d$. The ground metric matrix C is not required to be constant across all data samples, and can be interpreted as a materialization of the perturbation. Given an unobserved perturbation $\mathbf{p}_{\text{new}}$ and initial state $\boldsymbol{\mu}_{\text{new}}$, the objective is to predict a transportation map $T^\star \in \mathcal{N}(\boldsymbol{\mu}_{\text{new}}, \boldsymbol{\nu}^\star)$, which marginalization yields the final states $\boldsymbol{\nu}^\star$. At training time, we can use classical OT solvers to obtain an OT map for each sample, providing a list of tuples $(\mathbf{p}_i, T_i)$ where $T_i \in \mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{\nu}_i)$ is the solution to the i-th OT problem.

Herein, we design a quantum circuit for predicting a DSM $Q \in \Omega_d$ given a context $\mathbf{p}_{\text{new}}$ (Section 3.1). Given the unobserved distribution $\boldsymbol{\mu}_{\text{new}}$, we extract a structure from the DSM that can be rescaled with $\boldsymbol{\mu}_{\text{new}}$ to the predicted transportation map fulfilling the required initial marginal (Section 3.2).

3 Quantum formulation

Our quantum formulation leverages the following fundamental concept. Let $\overline{(\cdot)}$ denote the complex conjugate of the argument (i.e. in the case of a matrix argument, the transpose of the adjoint), and $(\cdot) \odot (\cdot)$ the Hadamard product between matrices. If U is a unitary matrix, then $U \odot \overline{U} \in \Omega_n$ is a DSM and such matrix is called *unistochastic*. The set of unistochastic matrices is a non-convex proper subset of the Birkhoff polytope, however the $n \times n$ permutations matrices all belong to such set, consequently its convex hull corresponds to the Birkhoff polytope, that is $\text{conv}\,(\mathfrak{S}_n) = \Omega_n$. The constraints required for an arbitrary DSM to be unistochastic are still an open problem[4].

Hence we can represent (with some approximation) the solution of the assignment problem using unitary operators. This principle produces DSM independently of the construction of the unitary U, which offers great freedom in the choice of the ansatz for U supporting both variational and kernel-based learning strategies [27]. Furthermore, such natural link between transportation maps and unitary operators may lead to quantum models enjoying better expressivity compared to classical counterparts. In fact, our problem can directly be related to recent results in quantum learning theory [28], where a separation between so-called quantum contextual and non-contextual models is proven under the assumption of specific constraints being present in the label space.

3.1 Quantum circuit for the Birkhoff polytope

We assume that the structures on n qubits have a dimension comparable to the input $d = 2^n$, where d is the number of entities for the discrete distributions considered in Section 2.2. Let U_p be a parametric unitary operator acting on the bipartite Hilbert space $\left(\mathbb{C}^2\right)^{\otimes m} \otimes \left(\mathbb{C}^2\right)^{\otimes n}$, with m being a non-negative integer such that the classical simulation of a quantum circuit on $m+n$ qubits is unfeasible in general. The operator U_p depends on the input vector $\mathbf{x}$ (perturbation) as well as on the learning parameters $\boldsymbol{\theta}$. To prove the construction, we consider the *Operator-Schmidt decomposition* of U_p determined by the quantum mechanical sub-systems A_1, B_1, consisting of respectively m and n qubits. So

$$U_p(\mathbf{x}, \boldsymbol{\theta}) = \sum_{i=1}^{d^2} \lambda_i V_i \otimes W_i, \tag{2}$$

with $\{V_i\}$ and $\{W_i\}$ being sets of unitary operators orthonormal w.r.t. the Frobenius inner product. Through singular value decomposition we have $\lambda_i \geq 0$, with unitarity implying $\sum_i \lambda_i^2 = 1$. Notably the matrix U_p depends on the input and the parameters vectors, then the components of Operator-Schmidt decomposition λ_i, V_i and W_i are functions of $(\mathbf{x}, \boldsymbol{\theta})$. Moreover, to assure the consistency of the formulation we impose the *Schmidt rank* of U_p (i.e. the number of strictly positive λ_i) to be

[3] This is a common requirement in OT, for avoiding degeneracy [14, Remark 2.1].

[4] For some specific instances, it has been numerically obtained that the unistochastic matrices cover $\approx 75\%$ of the Birkhoff polytope [26].

greater than one. Using the unitary U_p and the interleaved bell states $|b_n\rangle$ and $|b_m\rangle$ (defined in (11)), we obtain the following state

$$|\varphi\rangle = \left(\mathbb{I}_2^{\otimes m} \otimes U_p \otimes \mathbb{I}_2^{\otimes n}\right) \cdot \left(|b_m\rangle \otimes |b_n\rangle\right) \tag{3a}$$

$$\underset{(2)}{=} \sum_k \lambda_k \left(\mathbb{I}_2^{\otimes m} \otimes V_k\right) |b_m\rangle \otimes \left(W_k \otimes \mathbb{I}_2^{\otimes n}\right) |b_n\rangle \tag{3b}$$

$$\underset{(12a)}{=} \sum_k \lambda_k \frac{\mathrm{vec_r}\left(V_k^\top\right)}{\sqrt{2^m}} \otimes \frac{\mathrm{vec_r}\left(W_k\right)}{\sqrt{2^n}}, \tag{3c}$$

where $\mathrm{vec_r}(\cdot)$ is the vectorization operator defined in Section A.2. Now, we partition the Hilbert space on which $|\varphi\rangle$ lays into two subsystems; the first is denoted as A_2 and consists of the first $2m$ qubits. The second is denoted as B_2 and takes the last $2n$ qubits. We obtain the mixed state ρ by applying the partial trace over the system A_2, to the pure state $|\varphi\rangle\langle\varphi|$, that is

$$\rho = \mathrm{Tr}_{A_2}\left(|\varphi\rangle\langle\varphi|\right) \tag{4a}$$

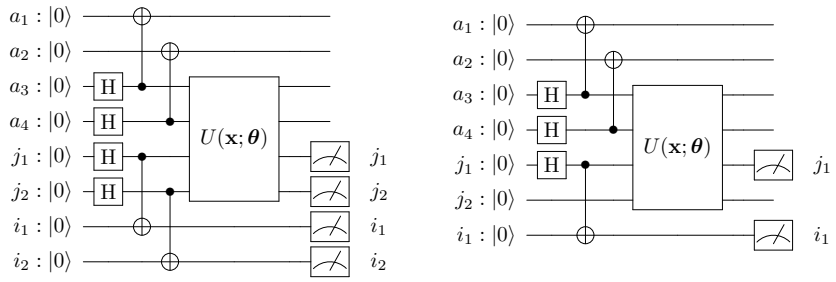
$$= \frac{1}{2^{m+n}} \sum_{i,j} \lambda_i \lambda_j \,\mathrm{Tr}\left(V_i V_j^\dagger\right) \mathrm{vec_r}\left(W_i\right) \mathrm{vec_r}\left(W_j\right)^\dagger \tag{4b}$$

$$\underset{(13)}{=} \frac{1}{2^n} \sum_i \lambda_i^2 \,\mathrm{vec_r}\left(W_i\right) \mathrm{vec_r}\left(W_i\right)^\dagger. \tag{4c}$$

Recall that by the Operator-Schmidt decomposition and unitarity of U_p we have that $\sum_i \lambda_i^2 = 1$. Given that the action of the unitary $U_p(\mathbf{x}, \boldsymbol{\theta})$ is generally not classically efficiently simulable, the state ρ has the potential to represent correlations that cannot be captured with classical models. The recovery of the DSM is completed with the lemma that follows.

Lemma 3.1. *Let* $p(i,j) := 2^n \,\mathrm{Tr}(\rho\,|ij\rangle\langle ij|)$ *for* $i,j \in [0\,..\,d-1]$. *Then the matrix* $Q = \sum_{i,j} p(i,j)\,|i\rangle\langle j|$ *is doubly stochastic.*

In the latter, the outer product $|i\rangle\langle j|$ defining the entries of the resulting DSM, is to be interpreted as the rank 1 matrix $\mathbf{e}_i\mathbf{e}_j^\top$ where the tuple (i,j) corresponds to the indexing of the matrix entry. Also, $\mathbf{e}_i$ denotes the i-th canonical basis vector for the vector space $\mathbb{R}^{2^n}$. The circuit structure resulting from the formulation is depicted in Figure 1a.



(a) Example of circuit structure for encoding DSMs (Section 3.1).

(b) Example of circuit structure for the case of the embedded transportation map (Section 3.2).

Figure 1: Circuit structures for the transportation map prediction. The registers $\{i_k\} \cup \{j_k\}$ represent the bits for the index (i,j) related to the entry $Q_{i,j}$ of the resulting DSM. The registry $\{a_k\}$ refers to the $2m$ qubits as per Section 3.1. We remark that the registry i has been added for construction reasons, however in practice it can be removed and substituted with a classical uniform sampling over the registry j. Consequently, the number of required qubits can be reduced (referring left picture) to $2m + n$.

3.2 Embedding of transportation maps

Since in our applications, the initial distribution $\boldsymbol{\mu}$ is user-provided at inference time, the problem is then twofold, namely the embedding of transportation maps into DSM to fit the representation

presented in Section 3.1, and the prediction of maps which can be rescaled to an arbitrary initial distribution. Starting from a training set $\{(\mathbf{p}_i, T_i)\}$ (i.e., tuples of contexts and transportation maps), we assume that $T_i \in \mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{\nu}_i)$ with $\boldsymbol{\mu}_i, \boldsymbol{\nu}_i \in \mathbb{R}^d_{++}$, that is the margins of the transportation maps are strictly positive[5]. Let $\mathbf{v} \in \mathbb{R}^d$, then we denote $D_{\mathbf{v}} = \operatorname{diag}(\mathbf{v})$ the $d \times d$ diagonal matrix having the elements of the vector $\mathbf{v}$ as diagonal elements. Now, given T_i as defined above, we define

$$\widehat{T}_i \coloneqq D_{\boldsymbol{\mu}_i}^{-1} T_i, \tag{5}$$

and observe that $\widehat{T}_i \mathbf{1}_d = D_{\boldsymbol{\mu}_i}^{-1} T_i \mathbf{1}_d = D_{\boldsymbol{\mu}_i}^{-1} \boldsymbol{\mu}_i = \mathbf{1}_d$, that is $\widehat{T}_i \in \mathcal{N}(\mathbf{1}_d, \boldsymbol{\nu}'_i)$ is a right stochastic matrix[6], with $\boldsymbol{\nu}'_i = \widehat{T}_i^\top \mathbf{1}_d$. At inference, when given a perturbation $\mathbf{p}_i$, the model predicts a right stochastic matrix $\widehat{T} \in \mathcal{N}(\mathbf{1}_d, \boldsymbol{\nu}')$ for some $\boldsymbol{\nu}' \in \mathbb{R}^d_{++}$. The latter, alongside the user-provided initial distribution $\boldsymbol{\mu}$, determines the final predicted map $T = D_{\boldsymbol{\mu}} \widehat{T}$. One can immediately verify that $T\mathbf{1}_d = D_{\boldsymbol{\mu}} \widehat{T} \mathbf{1}_d = D_{\boldsymbol{\mu}} \mathbf{1}_d = \boldsymbol{\mu}$. In other words, we learn the pattern of the transportations in a margin-independent fashion and rescale to the required margin at inference time. We note that, when the context is $\mathbf{0}$ (null perturbation) then $\widehat{T}_i = \mathbb{I}_d$. Given the user-provided $\boldsymbol{\mu}$ we have $D_{\boldsymbol{\mu}} \widehat{T}_i = D_{\boldsymbol{\mu}}$, hence $D_{\boldsymbol{\mu}} \mathbf{1}_d = D_{\boldsymbol{\mu}}^\top \mathbf{1}_d = \boldsymbol{\mu} = \boldsymbol{\nu}$, ergo the initial and final distributions agree (consistent with the notion of null perturbation), inducing a stationarity inductive bias.

To complete the structure, we now expand on the link between the $d \times d$ right stochastic matrix $\widehat{T}$ and a $2d \times 2d$ doubly stochastic matrix Q. This step is necessary since the formulation in Section 3.1 produces DSM only. First, consider the block decomposition of a DSM that follows

$$Q = \begin{pmatrix} Q_1 & Q_2 \\ Q_3 & Q_4 \end{pmatrix} \in \Omega_{2d}, \tag{6}$$

with $Q_i \in \mathbb{R}^{d \times d}_+$. Now, note that $\begin{pmatrix} Q_1 & Q_2 \end{pmatrix} \mathbf{1}_{2d} = \mathbf{1}_d$ implies $(Q_1 + Q_2)\,\mathbf{1}_d = \mathbf{1}_d$. We embed the right stochastic matrix $\widehat{T}_i$ into the sum $Q_1 + Q_2$ of the top quadrants of a DSM $Q \in \Omega_{2d}$. Since this structure does not consider Q_3 and Q_4, Section A.3 proposes a more parameter-efficient ansatz.

To obtain the number of required qubits, let m be the number of qubits (as per Section 3.1) that makes the function space achievable by the ansatz, hard to be computed classically. Also, let $d = 2^n$ and consider $d \times d$ transportation maps, then using the reduction introduced in Figure 1b (additionally we remove registry i), the overall circuit requires $(n+1) + 2m$ qubits.

3.3 Training objective

Let $f : \mathcal{X} \to \Omega_d^{(r)}$ be a function from the set of perturbations $\mathcal{X}$ to the set of $d \times d$ row-stochastic matrices, and let $\mathcal{F}$ be the function space of such functions related to our model. Then, given the training set $\{(\mathbf{p}_i, T_i)\}$ we define our learning problem as

$$f^\star = \operatorname*{argmin}_{f \in \mathcal{F}} \sum_i \| D_{\boldsymbol{\mu}_i} f(\mathbf{p}_i) - T_i \|_F, \tag{7}$$

where $\boldsymbol{\mu}_i = T_i \mathbf{1}_d$ is the initial distribution for the i-th training sample and $\| \cdot \|_F$ is the Frobenius matrix norm. At inference time, given the initial (quasi-)distribution $\boldsymbol{\mu} \in \mathbb{R}^d_{++}$ and the perturbation $\mathbf{p} \in \mathcal{X}$, the predicted transportation map is obtained as $T = D_{\boldsymbol{\mu}} f^\star(\mathbf{p})$. Note that for applications where one is interested primarily in the predicted target distribution $\boldsymbol{\nu}^\star$, the objective can easily be adjusted, e.g., through setting $f^\star = \operatorname{argmin} \sum_i \| D_{\boldsymbol{\mu}_i} f(\mathbf{p}_i) \mathbf{1}_d - \boldsymbol{\nu}_i \|$. In practice, we optimize the parameters of the ansatz representing the $f^\star$ through the quasi-Newton algorithm BFGS [29]. Since it can get expensive in terms of the required number of circuit evaluations to get gradients, one might use derivative-free alternatives, e.g. COBYLA [30].

Evaluation. To evaluate the accuracy of the transportation plans we use a relative Frobenius norm

$$F(\bar{T}_i, T_i) = \frac{\| \bar{T}_i - T_i \|_F}{\| \bar{T}_i \|_F} \tag{8}$$

where $\bar{T}_i = D_{\boldsymbol{\mu}_i} f^\star(\mathbf{p}_i)$. We report the relative norm because the absolute norm is unbound and we thus bypass constraints on the target norm. As secondary evaluation metric we report the first Wasserstein distance, acronymed W_1, on the flattened transportation plans.

[5] The constraint on the strict positivity of the margins is justified in Section 2.2.

[6] A non-negative $n \times n$ matrix Q is right stochastic when the rows sum to one, that is $Q\mathbf{1}_n = \mathbf{1}_n$.

4 Experimental setup

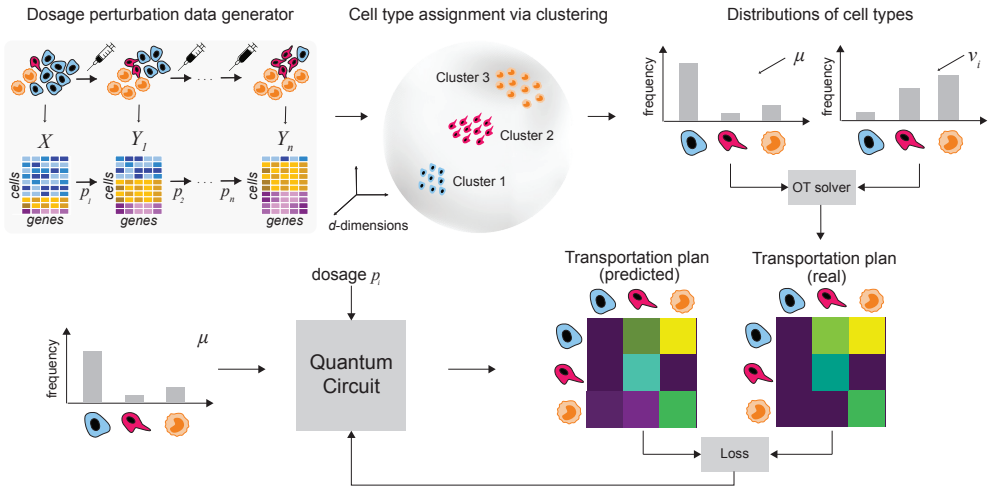


Figure 2: QontOT overview. We first employ a synthetic data generation scheme, where we perturb a population of cells using increasing dosage of a drug, resulting $(X_i, Y_i, \boldsymbol{p}_i)$ where X_i represents synthetic scRNA-seq measurements before and $Y_i \in \mathbb{R}^{n_2 \times l}$ synthetic scRNA-seq measurements after a drug perturbation administered with dosage $\boldsymbol{p}_i \in [0, 1]$. We cluster these measurements to identify all cell types present, and then compute, for each stage, the distribution of cell types before and after perturbation, i.e., $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$. Each pair $\boldsymbol{\mu}, \boldsymbol{\nu}_i$ is given as an input to a classical OT solver that computes a real transportation plan T_i. We feed to our quantum circuit the cluster distribution before perturbation $\boldsymbol{\mu}$ and the dosage $\boldsymbol{p}_i$ to infer a predicted transportation plan $\bar{T}_i$ and use the relative Frobenius norm as a loss function to train the QC.

We applied our proposed method on predicting changes in the composition of a population of cells following a drug perturbation and tested it on synthetic data as visualized in Figure 2. Starting from a mixed population of cells that belong to different cell types and live in a high-dimensional cell state space we assume that administering a drug has a direct effect on the composition of the cell population, by eliminating certain cell types or pushing some other cell types to proliferate. We denote as $\boldsymbol{\mu}, \boldsymbol{\nu}_i$ the cell type distribution of a cell population *before* and *after* the drug perturbation with a context variable $\boldsymbol{p}_i$. We note that in all experiments, $\boldsymbol{p}_i \in [0, 1]$ is the drug dosage.

4.1 Dosage perturbation data generator

We leverage the established single-cell RNA sequencing generator `Splatter` [31] to form a three-stage generator for drug dosage perturbation datasets:

1. First, `Splatter` samples raw expression counts ($X \in \mathbb{R}^{n_1 \times l}$, with n_1 cells and l genes) from zero-inflated negative binomial distributions (one per gene). Sufficient statistics of all underlying distributions (Poisson, Gamma, Chi-Square) can be controlled.
2. We aim to produce a tuple of $(X_i, Y_i, \boldsymbol{p}_i)$ where X_i holds unperturbed base states of n_1 cells and $Y_i \in \mathbb{R}^{n_2 \times l}$ holds perturbed states of n_2 cells, resulting from a drug perturbation administered with dosage $\boldsymbol{p}_i \in [0, 1]$. To derive the perturbed states Y_i, new base states $\bar{Y}_i$ are sampled with the same configuration used to generate X_i, mimicking that cells are being destroyed during measurement. Subsequently $Y_i = g(\bar{Y}_i, \boldsymbol{p}_i)$ where g is the total effect on the cells, governed by a combination of noise terms and the immediate effect $g_p(\cdot)$ of the perturbation. We assume that only 15% of the genes alter their expression upon perturbation. In this case, we apply g_p to the raw cell states, scaled by a response amplitude $\sim \mathcal{U}(0.3, 1)$. Moreover, 10% of the cells are generally unresponsive to the perturbation ($g_p = 0$). We investigate linear and non-linear perturbations, i.e., $g_{p_1}(x) = ax + b$ and a reciprocal root function $g_{p_2}(x) = ax^{-b}$ with $a, b > 0$, respectively. The hyperparameters of the experiments can be found in Appendix A.1.
3. We repeat stage 2 for each dosage by varying smoothly the immediate effect $g_p(\cdot)$ based on $\boldsymbol{p}_i$, resulting in a dataset $\{X_i, Y_i, \boldsymbol{p}_i\}_{i=1}^N$ of N tuples. Responsive genes are fixed across samples. The

base states X_i are *identical* across all samples of the dataset, mimicking the common experimental setting where only one control population was measured [32].

4.2 Cell type assignment via clustering

We represent each cell through a single cell type label, which can be easily achieved following any conventional clustering approach in the original $\mathbb{R}^l$ space. Following a cell clustering into d clusters (roughly corresponding to cell types), we compute $\boldsymbol{\mu}, \boldsymbol{\nu} \in \mathbb{R}^d$ as distributions over cluster labels for the population of cells before and after perturbation. In practice, we use k-Means clustering to cluster all cells of the training set and set $k = d = 16$ to adhere with the requirements of the quantum circuit, i.e., d must be a power of 2 (in the general case, if we pick a different number of clusters and $\log_2(k) \notin \mathbb{N}$, we can pad and set $d := 2^{\lceil \log_2(k) \rceil}$ and fix the transportation plan to be diagonal for the padded entries). We then solve Equation 1 and compute the OT map between $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$, using the entropically regularized Sinkhorn solver [24] from the `POT` library [33]. The cost matrix $C \in \mathbb{R}^{d \times d}$ is computed between the clustering centroids using a distance measure of choice, in our case either a Euclidean (L_2) or a cosine distance. Repeating this procedure for all dosages yields a dataset $\{T_i, \boldsymbol{p}_i\}_{i=1}^N$ of transportation plan-perturbation tuples, processed as described in Section 3.3. Note that, even though the perturbation functions g_p are simple (linear or inverse root) and only affect expression of a subset of cells and genes, the induced changes in the cell type distributions are significant and locally continuous (cf. Figure A1).

An alternative approach to generate cell labels is to perform dimensionality reduction on the training cells and rasterize the obtained space into equidistant hypercubes. Cells are then assigned to the hypercube they originate from and the cost matrix is computed using the L_1 distance of the centers of the hypercubes, scaled by dimension weights (e.g., eigenvalues of a PCA). We test this procedure with a Kernel PCA [34] using a radial basis function kernel and two principal components.

Soft assignment and cell reconstruction To better capture the underlying structure of the data when generating cell labels, we employed fuzzy C-Means clustering [35] to yield a distribution of cluster assignments for each cell. Combined with the cluster label distribution $\boldsymbol{\nu}^\star$ obtained from the predicted $\bar{T}_i$, this allows expression values in the original gene space to be reconstructed: For each source cell, the mass assigned to each cluster is distributed using the corresponding row of the predicted transportation plan. These new cluster assignment vectors are summed up across all original clusters. Last, the expression value of each gene is calculated as an average of all cluster centroids, weighted by the predicted soft cluster assignment. Note that this procedure can be performed for both aforementioned cell label generation techniques, clustering and dimensionality reduction.

5 Experimental results

The first set of experiments aims to verify that our method QontOT (for **Q**uantum C**ont**extual **O**ptimal **T**ransport) can learn to predict transportation maps *contextualized* through a perturbation variable. The results are produced using a PyTorch-based custom state vector simulator [36]. We compare QontOT to a baseline which always predicts the same transportation plan, obtained by solving the regularized OT problem (Equation 1) on all training samples at once, disregarding the context.

The results in Table 1 show that QontOT outperforms the baseline in all cases by a wide margin. Unlike related work that leverages Brenier's theorem [5], our method is not limited to squared Euclidean cost and works equally when computing the cost through cosine distances of the centroids. Interestingly, QontOT outperforms the baseline not only in the linear but also in the non-linear perturbation case, an arguably more realistic and challenging scenario, as g_{p_2} induces strong up-regulation of genes with low initial expression but very low up-regulation if genes are already highly expressed. Moreover, we observe a general trend that more layers in the ansatz (cf. Section 3.2) yield better performance, leaving room for further improvement. Some exemplary real and predicted transportation plans (cf. Figure 3a) serve as a testimony that QontOT can learn context-dependent shifts in cell type frequencies, by capturing the change in the distribution of cluster labels induced by the perturbation. Assessing the relation between dosage and error (cf. Figure 3b) reveals that predicting the effect of stronger perturbations (higher dosages) is more challenging. This is expected because in the control condition ($\boldsymbol{p}_i = 0$), the cell ID distribution remains identical, subject only to stochastic effects in data generation and selection of cells for the batch. The baseline model, however,

Perturb.	Distance	Method	Layers	W_1	Frobenius
Linear (g_{p_1})	L_2	Baseline	–	$0.0036_{\pm 0.001}$	$0.80_{\pm 0.06}$
Linear (g_{p_1})	L_2	QontOT	4	$0.0029_{\pm 0.001}$	$0.70_{\pm 0.15}$
Linear (g_{p_1})	L_2	QontOT	6	$0.0028_{\pm 0.001}$	$0.67_{\pm 0.13}$
Linear (g_{p_1})	Cosine	Baseline	–	$0.0032_{\pm 0.000}$	$0.84_{\pm 0.07}$
Linear (g_{p_1})	Cosine	QontOT	4	$0.0029_{\pm 0.000}$	$0.72_{\pm 0.14}$
Linear (g_{p_1})	Cosine	QontOT	6	$0.0029_{\pm 0.000}$	$0.69_{\pm 0.14}$
Nonlinear (g_{p_2})	L_2	Baseline	–	$0.0024_{\pm 0.001}$	$0.75_{\pm 0.10}$
Nonlinear (g_{p_2})	L_2	QontOT	6	$0.0016_{\pm 0.001}$	$0.48_{\pm 0.19}$
Nonlinear (g_{p_2})	L_2	QontOT	10	$0.0014_{\pm 0.001}$	$0.43_{\pm 0.20}$

Table 1: Transportation plan prediction. Performance in predicting transportation plans for unseen context, comparing our approach (QontOT) to a baseline model. Each block of three lines denotes one dataset, evaluated with two QontOT models and the baseline. Different distance metrics were used to derive the cost matrix from the K-means centroids and both linear and non-linear perturbation effects were recovered. W_1 denotes first Wasserstein distance and Frobenius is the relative Frobenius norm (cf. Equation 8). Layers refers to the number of layers in the ansatz. Each experiment was performed three times with different random seeds.

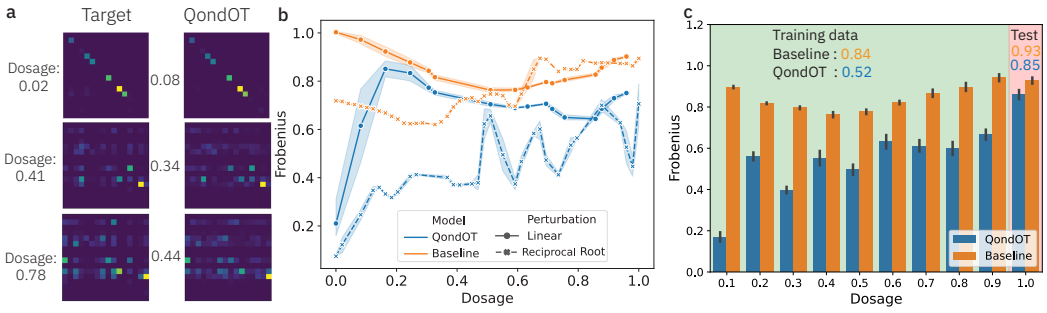


Figure 3: Capturing variation in cell cluster ID distributions via transportation plan prediction. a) Three predicted transportation plans from the nonlinear perturbation dataset are shown next to their unseen ground truth. **b)** Frobenius distance of real and predicted transportation across unseen dosages are shown for QontOT and the baseline. **c)** Out-of-distribution scenario. When QontOT is evaluated on dosages outside the training data scenario, the performance decreases but still remains well above the baseline.

performs relatively well for mediocre dosages but struggles with extreme values. Given the higher difficulty (i.e., transportation cost) in predicting transportation plans for large dosages, we performed an extrapolation experiment where dosages $\boldsymbol{p}_i \in [0, 0.9]$ were used for training and $\boldsymbol{p}_i \in [0.9, 1]$ were used for evaluation. The barplot in Figure 3c reveals that the test error is higher than the training error, however QontOT still clearly outperforms the baseline.

Note that, if cells are assigned unique cluster labels, QontOT predicts distributions over cluster assignments but this can hardly be mapped back to gene expression values on a single-cell level. As a remedy, we utilize a soft assignment approach, implemented through soft clustering (e.g., fuzzy C-Means [35]) or a dimensionality reduction, followed by a propagation procedure to obtain predicted single-cell expression values (cf. Section 4.2). In Figure 4a, we verify that when predicting transportation maps for soft assignments, our results from Figure 3 are confirmed. While the same general trends can be seen, an interesting observation is the high error of QontOT for low dosages in the Kernel PCA setting. This error can be attributed to the stationarity inductive bias of the method (diagonal transportation plans for $p_i = 0$). This is generally a desirable assumption of the method, but the kernelPCA seems to not always preserve local similarity of cells, inducing distribution shifts even for control states. Some examples of reconstructed gene expression are displayed in Figure 4b and c, revealing that QontOT captures the general response of the gene induced by the drug dosage.

6 Discussion

In this work we have first introduced a principled approach to represent transportation maps on quantum computers. Such representation is justified by the constraints defining the transportation maps, which can be related to recent work on quantum contextuality and inductive bias in quantum

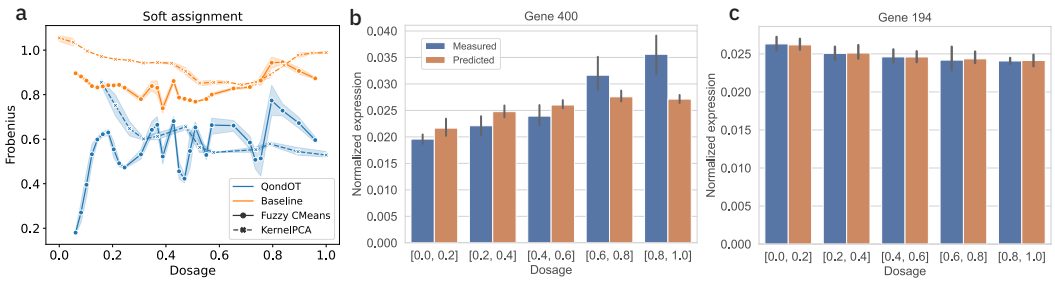


Figure 4: Soft assignment. a) Performance on different approaches of soft assignment of cells to their group. **b)**, **c)** Soft assignment facilitates the reconstruction of gene expression data from predicted cell label distributions. For two exemplary genes, measured and predicted gene expression across unseen dosages are shown. Expressions were averaged across all genes.

machine learning [28]. We then proposed an ansatz for learning to predict such OT maps conditioned on a context variable, without requiring access to the cost matrix. We provide empirical evidence supporting that our methodology can successfully learn to predict contextualized transportation maps. Exemplified on synthetic data of single-cell drug dosage perturbations, our method is able to predict transportation plans representing distributional shifts in cell type assignments. While further validation on more realistic and challenging datasets as well as better classical comparisons are clearly needed, our work constitutes, to the best of our knowledge, the first approach to bridge OT and ML on quantum computers. Notably our approach does not impose general constraints on the dimensionality of the context variable(s), thus more complex perturbations such as continuous drug representations, combinatorial genetic perturbations or other covariates could be also employed. Given that, in our described application, the dosage-induced shifts in cluster assignments are also largely driven by the initial cell states (rather than only the dosage), future work will be devoted to make our ansatz fully parametric for $\boldsymbol{\mu}_i$, potentially through co-optimal transport [37, 38].

References

[1] C. Villani. *Optimal Transport: Old and New*. Grundlehren der mathematischen Wissenschaften. Springer Berlin Heidelberg, 2008. ISBN 9783540710509. URL https://books.google.ie/books?id=hV8o5R7_5tkC.

[2] Dominik Klein, Giovanni Palla, Marius Lange, Michal Klein, Zoe Piran, Manuel Gander, Laetitia Meng-Papaxanthos, Michael Sterr, Aimee Bastidas-Ponce, Marta Tarquis-Medina, et al. Mapping cells through time and space with moscot. *bioRxiv*, pages 2023–05, 2023.

[3] Geoffrey Schiebinger, Jian Shu, Marcin Tabaka, Brian Cleary, Vidya Subramanian, Aryeh Solomon, Joshua Gould, Siyan Liu, Stacie Lin, Peter Berube, et al. Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming. *Cell*, 176(4):928–943, 2019.

[4] 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*, pages 1–10, 2023.

[5] Charlotte Bunne, Andreas Krause, and Marco Cuturi. Supervised training of conditional monge maps. *Advances in Neural Information Processing Systems*, 35:6859–6872, 2022.

[6] Zixuan Cang and Qing Nie. Inferring spatial and signaling relationships between cells from single cell transcriptomic data. *Nature communications*, 11(1):2084, 2020.

[7] Kai Cao, Qiyu Gong, Yiguang Hong, and Lin Wan. A unified computational framework for single-cell data integration with optimal transport. *Nature Communications*, 13(1):7419, 2022.

[8] Federico Gossi, Pushpak Pati, Panagiotis Chouvardas, Adriano Luca Martinelli, Marianna Kruithof-de Julio, and Maria Anna Rapsomaniki. Matching single cells across modalities with contrastive learning and optimal transport. *Briefings in bioinformatics*, 24(3):bbad130, 2023.

[9] Mirjana Efremova and Sarah A Teichmann. Computational methods for single-cell omics across modalities. *Nature methods*, 17(1):14–17, 2020.

[10] Yann Brenier. Décomposition polaire et réarrangement monotone des champs de vecteurs. *CR Acad. Sci. Paris Sér. I Math.*, 305:805–808, 1987.

[11] Ashok Makkuva, Amirhossein Taghvaei, Sewoong Oh, and Jason Lee. Optimal transport mapping via input convex neural networks. In *International Conference on Machine Learning*, pages 6672–6681. PMLR, 2020.

[12] Brandon Amos, Lei Xu, and J Zico Kolter. Input convex neural networks. In *International Conference on Machine Learning*, pages 146–155. PMLR, 2017.

[13] Esteban G Tabak, Giulio Trigila, and Wenjun Zhao. Data driven conditional optimal transport. *Machine Learning*, 110:3135–3155, 2021.

[14] Gabriel Peyré and Marco Cuturi. Computational optimal transport. 2018. doi: 10.48550/ARXIV.1803.00567. URL https://arxiv.org/abs/1803.00567.

[15] Amira Abbas, David Sutter, Christa Zoufal, Aurélien Lucchi, Alessio Figalli, and Stefan Woerner. The power of quantum neural networks. *Nature Computational Science*, 1(6):403–409, 2021.

[16] Thaddeus D Ladd, Fedor Jelezko, Raymond Laflamme, Yasunobu Nakamura, Christopher Monroe, and Jeremy Lloyd O'Brien. Quantum computers. *nature*, 464(7285):45–53, 2010.

[17] Vojtěch Havlíček, Antonio D. Córcoles, Kristan Temme, Aram W. Harrow, Abhinav Kandala, Jerry M. Chow, and Jay M. Gambetta. Supervised learning with quantum-enhanced feature spaces. *Nature*, 567(7747):209–212, 2019. doi: 10.1038/s41586-019-0980-2. URL https://doi.org/10.1038/s41586-019-0980-2.

[18] Yunchao Liu, Srinivasan Arunachalam, and Kristan Temme. A rigorous and robust quantum speed-up in supervised machine learning. *Nature Physics*, 2021. ISSN 1745-2481.

[19] Aram W. Harrow, Avinatan Hassidim, and Seth Lloyd. Quantum algorithm for linear systems of equations. *Physical Review Letters*, 103(15), 2009.

[20] Hsin-Yuan Huang, Michael Broughton, Masoud Mohseni, Ryan Babbush, Sergio Boixo, Hartmut Neven, and Jarrod R. McClean. Power of data in quantum machine learning. *Nature Communications*, 12(1):2631, 2021.

[21] Saugata Basu, Jannis Born, Aritra Bose, Sara Capponi, Dimitra Chalkia, Timothy A Chan, Hakan Doga, Mark Goldsmith, Tanvi Gujarati, Aldo Guzman-Saenz, et al. Towards quantum-enabled cell-centric therapeutics. *arXiv preprint arXiv:2307.05734*, 2023.

[22] Alberto Di Meglio, Karl Jansen, Ivano Tavernelli, Constantia Alexandrou, Srinivasan Arunachalam, Christian W Bauer, Kerstin Borras, Stefano Carrazza, Arianna Crippa, Vincent Croft, et al. Quantum computing for high-energy physics: State of the art and challenges. summary of the qc4hep working group. *arXiv preprint arXiv:2307.03236*, 2023.

[23] Elie Bermot, Christa Zoufal, Michele Grossi, Julian Schuhmacher, Francesco Tacchino, Sofia Vallecorsa, and Ivano Tavernelli. Quantum generative adversarial networks for anomaly detection in high energy physics. *arXiv preprint - arXiv:2304.14439*, 2023.

[24] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transportation distances. 2013. doi: 10.48550/ARXIV.1306.0895. URL https://arxiv.org/abs/1306.0895.

[25] Richard A. Brualdi. *Combinatorial Matrix Classes*. Encyclopedia of Mathematics and its Applications. Cambridge University Press, 2006.

[26] Charles Dunkl and Karol Życzkowski. Volume of the set of unistochastic matrices of order 3 and the mean jarlskog invariant. *Journal of Mathematical Physics*, 50(12), December 2009. ISSN 1089-7658. doi: 10.1063/1.3272543. URL http://dx.doi.org/10.1063/1.3272543.

[27] Nicola Mariella, Jannis Born, Marianna Rapsomaniki, Stefan Wörner, and Sergiy Zhuk. Quantum regression for contextual optimal transport. In *Quantum Techniques in Machine Learning (QTML)*, Geneva, Switzerland, November 2023. In press.

[28] Joseph Bowles, Victoria J Wright, Máté Farkas, Nathan Killoran, and Maria Schuld. Contextuality and inductive bias in quantum machine learning, 2023.

[29] Roger Fletcher. *Practical methods of optimization*. John Wiley & Sons, 2000.

[30] Michael JD Powell. *A direct search optimization method that models the objective and constraint functions by linear interpolation*. Springer, 1994.

[31] Luke Zappia, Belinda Phipson, and Alicia Oshlack. Splatter: simulation of single-cell rna sequencing data. *Genome biology*, 18(1):174, 2017.

[32] 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.

[33] Rémi Flamary, Nicolas Courty, Alexandre Gramfort, Mokhtar Z. Alaya, Aurélie Boisbunon, Stanislas Chambon, Laetitia Chapel, Adrien Corenflos, Kilian Fatras, Nemo Fournier, Léo Gautheron, Nathalie T.H. Gayraud, Hicham Janati, Alain Rakotomamonjy, Ievgen Redko, Antoine Rolet, Antony Schutz, Vivien Seguy, Danica J. Sutherland, Romain Tavenard, Alexander Tong, and Titouan Vayer. Pot: Python optimal transport. *Journal of Machine Learning Research*, 22(78):1–8, 2021. URL http://jmlr.org/papers/v22/20-451.html.

[34] Bernhard Schölkopf, Alexander Smola, and Klaus-Robert Müller. Kernel principal component analysis. In *International conference on artificial neural networks*, pages 583–588. Springer, 1997.

[35] James C Bezdek. *Pattern recognition with fuzzy objective function algorithms*. Springer Science & Business Media, 2013.

[36] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

[37] Vayer Titouan, Ievgen Redko, Rémi Flamary, and Nicolas Courty. Co-optimal transport. *Advances in neural information processing systems*, 33:17559–17570, 2020.

[38] Quang Huy Tran, Hicham Janati, Nicolas Courty, Rémi Flamary, Ievgen Redko, Pinar Demetci, and Ritambhara Singh. Unbalanced co-optimal transport. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 10006–10016, 2023.

[39] Qiskit contributors. Qiskit: An open-source framework for quantum computing, 2023.

[40] Michael A. Nielsen and Isaac L. Chuang. *Quantum Computation and Quantum Information: 10th Anniversary Edition*. Cambridge University Press, USA, 10th edition, 2011. ISBN 1107002176.

[41] Adriano Barenco, Charles H. Bennett, Richard Cleve, David P. DiVincenzo, Norman Margolus, Peter Shor, Tycho Sleator, John A. Smolin, and Harald Weinfurter. Elementary gates for quantum computation. *Physical Review A*, 52(5):3457–3467, nov 1995. doi: 10.1103/physreva.52.3457. URL https://doi.org/10.1103%2Fphysreva.52.3457.

A Experimental details

A.1 Datasets and hyperparameters

For the results shown in Table 1 and Figure 3, we simulate 300 genes and 1000 cells across 50 unique dosages, equidistantly spaced in $[0, 1]$. 15% of the genes respond to the perturbation $g_p(\cdot)$ but 10% of the cells are set as unresponsive. The sinkhorn regularization $\gamma = 0.001$. For the linear case, $f_{p_1}(x) = 3x + 1$ and for the non-linear case $f_{p_2}(x) = 100x^{-0.2}$. For each dosage, four batches of 500 cells each were created, summing to 200 samples which were split randomly with 20% test data. For the extrapolation experiment in Figure 3**c**, the 10% of samples with the highest dosage was used as test data. For the (soft) clustering and KernelPCA in Figure 4, a dataset with 500 genes, 2000 cells, 50 dosages and 500 cells per batch were used (the remaining parameters were identical). The ansatz has been implemented in `Qiskit` 0.43.0 [39] and the remaining source code is written in `PyTorch` 2.0.1 [36]. For practical reasons, simulations were performed on classical hardware (CPU).

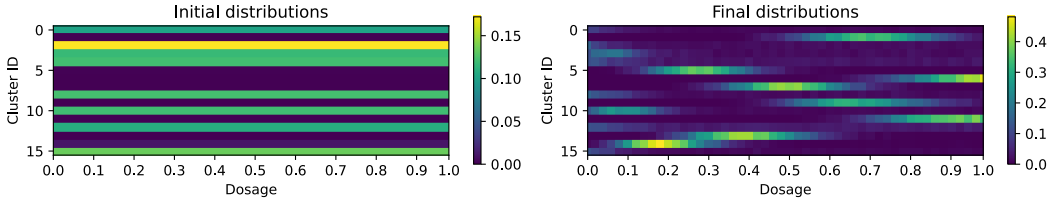


Figure A1: Exemplary cell type distributions for source and target cell populations. While the distribution of cell types in the unperturbed tissue is static (see left), the perturbed cells produce highly dissimilar distributions that, however, exhibit some local continuity for similar dosages. Plot generated with data from linear perturbations and euclidean cost.

A.2 Linear algebra

Let Δ_n denote the simplex in $n-1$ dimensions, that is the set

$$\Delta_n = \left\{ \mathbf{v} \in \mathbb{R}_+^n \middle| \mathbf{1}_n^\top \mathbf{v} = 1 \right\}. \tag{9}$$

The elements $Q \in \Omega_n$ are *doubly stochastic matrices* (DSM), also the extreme points of the polytope are permutation matrices. A DSM $Q \in \Omega_n$ can be decomposed as

$$Q = \sum_{i=1}^{N} \lambda_i P_i, \tag{10}$$

for some $\boldsymbol{\lambda} \in \Delta_N$, $n \times n$ permutation matrices $\{P_i\}$, and the number of extreme points $N \leq n^2$. We note that the decomposition is not unique.

Fundamental for the quantum formulation is the set $\mathfrak{S}_n$ of $n \times n$ *unistochastic matrices*. Given any $n \times n$ unitary matrix U, the matrix obtained by substituting each element of U with its absolute value squared, is unistochastic. In other words, let $U \in \mathfrak{U}(n)$, then $U \odot \overline{U}$ is doubly stochastic, where $\overline{U} = \left(U^\dagger\right)^\top$. The latter result is an implication of unitarity. The set of unistochastic matrices is a non-convex proper subset of the Birkhoff polytope, however the $n \times n$ permutations matrices all belong to such set, consequently its convex hull corresponds to the Birkhoff polytope, that is $\mathrm{conv}\left(\mathfrak{S}_n\right) = \Omega_n$. The constraints required for an arbitrary DSM to be unistochastic are still an open problem.

We denote with J_n the $n \times n$ matrix of ones, that is $J_n = \mathbf{1}_n \mathbf{1}_n^\top$.

Let $|b_n\rangle$ denote the quantum state consisting of interleaved Bell's states on $\left(\mathbb{C}^2\right)^{\otimes n} \otimes \left(\mathbb{C}^2\right)^{\otimes n}$, so

$$|b_n\rangle = \frac{1}{\sqrt{2^n}} \sum_{i=0}^{2^n-1} |i\rangle \otimes |i\rangle. \tag{11}$$

Note that $|b_n\rangle$ corresponds to the vectorization of the identity operator up to a scalar multiple, that is $\frac{1}{\sqrt{2^n}} \mathrm{vec_r}\left(\mathbb{I}_2^{\otimes n}\right) = |b_n\rangle$, where $\mathrm{vec_r}$ denotes the *row-major vectorization operator*. Moreover, given a $n \times n$

matrix M in $\mathbb{C}$, we will be using the identities [7]

$$\operatorname{vec}_{\mathrm{r}}(M) = (M \otimes \mathbb{I}_n) \operatorname{vec}_{\mathrm{r}}(\mathbb{I}_n) \tag{12a}$$

$$\operatorname{vec}_{\mathrm{r}}\left(M^\top\right) = (\mathbb{I}_n \otimes M) \operatorname{vec}_{\mathrm{r}}(\mathbb{I}_n). \tag{12b}$$

The following lemma establishes a relation between unitary operators and their vectorization.

Lemma A.1. *Let $\{U_i\}$ be a set of unitary operators $U_i \in \mathfrak{U}(n)$ such that $\operatorname{Tr}\left(U_i U_j^\dagger\right) = n\delta_{ij}$, that is the unitaries are orthogonal w.r.t. the Frobenius inner product. Then the set $\{\operatorname{vec}_{\mathrm{r}}(U_i)\} \subset \mathbb{C}^{n^2}$ consists of orthogonal vectors, that is*

$$\operatorname{vec}_{\mathrm{r}}(U_j)^\dagger \operatorname{vec}_{\mathrm{r}}(U_i) = \operatorname{Tr}\left(U_i U_j^\dagger\right) = n\delta_{ij}. \tag{13}$$

Let σ_i with $i = [1\,..\,3]$ be the Pauli operators [40] commonly denoted with σ_x, σ_y and σ_z, respectively. Also we define $\sigma_0 = \mathbb{I}_2$. The subscript of the σ to determine the Pauli will be indicated interchangeably as symbol or integer index.

A.3 The checkerboard ansatz

We propose an ansatz construction which is compatible with the structure of the embedding of transportation maps expanded in Section 3.2. For some positive integer k, we define the subset G_k of unitary operators as

$$G_k \coloneqq \{U \in \mathsf{U}(2k) \,|\, U = \mathbb{I}_2 \otimes A + \sigma_x \otimes B\}, \tag{14}$$

where A, B are $k \times k$ matrices, not necessarily unitary. In other words, the operators in $U \in G_k$ have the following block matrix form

$$U = \begin{pmatrix} A & B \\ B & A \end{pmatrix}, \tag{15}$$

which is clearly inherited by the corresponding unistochastic $U \odot \overline{U}$.

The next lemma shows that the set G_k is a subgroup of even degree of the unitary group.

Lemma A.2. *The set G_k is non-empty and endowed with a group structure under operator composition, for all positive integers k.*

Proof. It is immediately verifiable that $\mathbb{I}_{2k} \in G_k$, that is the set G_k is non-empty and it contains the identity element w.r.t. matrix multiplication. Also the composition of operators carries the associativity as required. Finally we verify the closure. Let $U_1, U_2 \in G_k$ such that $U_i = \mathbb{I}_2 \otimes A_i + \sigma_x \otimes B_i$ for $i = 1, 2$, then

$$U_1 U_2 = (\mathbb{I}_2 \otimes A_1 + \sigma_x \otimes B_1)(\mathbb{I}_2 \otimes A_2 + \sigma_x \otimes B_2) \tag{16a}$$

$$= \mathbb{I}_2 \otimes (A_1 A_2 + B_1 B_2) + \sigma_x \otimes (A_1 B_2 + B_1 A_2), \tag{16b}$$

which corresponds to the pattern in (15). Hence $U_1 U_2 \in G_k$. □

The result that follows shows that the structure is preserved under the tensor product.

Lemma A.3. *Let $U_1 \in G_{k_1}$ and $U_2 \in G_{k_2}$, for some positive integers k_1 and k_2. Then $U_1 \otimes U_2 \in G_{2k_1k_2}$.*

Proof. Let $U_1 \in G_{k_1}$ and $U_2 \in G_{k_2}$ such that $U_i = \mathbb{I}_2 \otimes A_i + \sigma_x \otimes B_i$ for $i = 1, 2$, then

$$U_1 \otimes U_2 = (\mathbb{I}_2 \otimes A_1 + \sigma_x \otimes B_1) \otimes (\mathbb{I}_2 \otimes A_2 + \sigma_x \otimes B_2) \tag{17a}$$

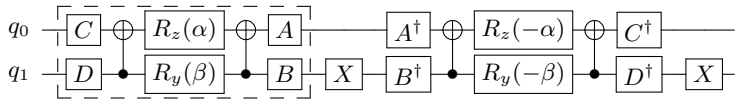
$$= \mathbb{I}_2 \otimes A + \sigma_x \otimes B, \tag{17b}$$

with $A = A_1 \otimes (\mathbb{I}_2 \otimes A_2 + \sigma_x \otimes B_2)$ and $B = B_1 \otimes (\mathbb{I}_2 \otimes A_2 + \sigma_x \otimes B_2)$. Hence it follows that $U_1 \otimes U_2$ fulfils the pattern in (15) and since A, B are linear maps in $\mathbb{C}^{2k_1k_2}$, then $U_1 \otimes U_2 \in G_{2k_1k_2}$. □

[7] In the context of quantum information theory the identity in (12a) is known as the Choi-Jamiołkowski correspondence.

A.3.1 Ansatz's two-qubit generator

We obtain a two-qubit circuit $U_g \in G_2$, that by Lemma A.2 and A.3 can be used as a generator for the more general G_{2^k} with $k \geq 1$. From the definition in (14) we obtain the symmetry $U \in G_2 \implies (\sigma_x \otimes \mathbb{I}_2)U(\sigma_x \otimes \mathbb{I}_2) = U$. Using the latter and the general unitary circuit with 2 CNOTs (highlighted) [41], we solve the following circuit equation

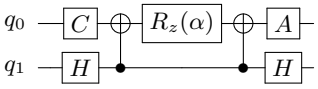


$$= \mathbb{I}_2^{\otimes 2}, \tag{18}$$

where A, B, C, D are arbitrary single qubit (special) unitaries and $\alpha, \beta \in \mathbb{R}$. Now, since the operator $\sigma_z \otimes \mathbb{I}_2$ commutes with the CNOT gate when the Pauli σ_z acts on the controlling qubit, we impose on (18) the conditions

$$B^\dagger \sigma_x B = \sigma_z, \tag{19a}$$
$$R_y(-\beta)\sigma_z R_y(\beta) = \sigma_z, \tag{19b}$$
$$D^\dagger \sigma_z D = \sigma_x. \tag{19c}$$

Then a solution is $B = D = H$ and $\beta = 0$, where H is the Hadamard operator on a single qubit. Hence the generator circuit takes the following form



(20)

where $A, C \in \mathsf{SU}(2)$ and $\alpha \in \mathbb{R}$.

Finally, by using Lemma A.2 and A.3, and the generator block in (20), we construct the ansatz as exemplified in Figure A2.

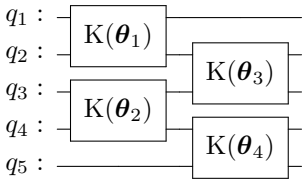


Figure A2: An example of depth and connectivity efficient layer for the Checkerboard ansatz. Here the blocks K correspond to the 2-qubits circuit in (20) and the vectors $\boldsymbol{\theta}_i$ are seven dimensional vectors parameterizing gates A, C and R_z of (20).

Proofs

Proof of Lemma A.1. Considering the constraint $\mathrm{Tr}\left(U_i U_j^\dagger\right) = n\delta_{ij}$ we obtain

$$\mathrm{vec_r}(U_j)^\dagger \,\mathrm{vec_r}(U_i) = \sum_k \left(\langle k| U_j^\dagger \otimes \langle k|\right) \sum_t \left(U_i |t\rangle \otimes |t\rangle\right) \tag{21a}$$
$$= \sum_{k,t} \left(\langle k| U_j^\dagger U_i |t\rangle \otimes \langle k|t\rangle\right) \tag{21b}$$
$$= \sum_k \left(\langle k| U_j^\dagger U_i |k\rangle\right) = \mathrm{Tr}\left(U_i U_j^\dagger\right) = n\delta_{ij}. \tag{21c}$$

□

Proof of Lemma 3.1. We expand the function $p : [0\,..\,d-1] \times [0\,..\,d-1] \to [0,1]$ defined in . . . , so

$$p(i,j) := 2^n \,\mathrm{Tr}(\rho\, |ij\rangle\langle ij|) \tag{22a}$$
$$= \sum_k \lambda_k^2 \,\langle ij|\,\mathrm{vec_r}\,(W_k)\,\mathrm{vec_r}\,(W_k)^\dagger\,|ij\rangle \tag{22b}$$
$$= \sum_k \lambda_k^2 \,\langle i|\left(W_k \odot \overline{W_k}\right)|j\rangle\,. \tag{22c}$$

The positivity of the entries of the DSM is clear from the definition of $p(i,j)$. We prove the rows sum constraint for $Q = \sum_{i,j} p(i,j) \left|i\right\rangle \left\langle j\right|$, that is

$$Q\mathbf{1}_d = \sum_{i,j} p(i,j) \left|i\right\rangle \tag{23a}$$

$$= \sum_i \left|i\right\rangle \cdot \left(\sum_k \lambda_k^2 \left\langle i\right| \sum_j \left(W_k \odot \overline{W_k} \right) \left|j\right\rangle \right), \tag{23b}$$

where the rightmost sum equals the vector $\mathbf{1}_m$ since $W_k \odot \overline{W_k}$ is unistochastic, also $\sum_k \lambda_k^2 = 1$ (following from (2)), hence $Q\mathbf{1}_d = \mathbf{1}_d$. Similarly the same holds for the columns sum constraint, hence the claim follows. □