

STARDUST: Subcellular-level Tool for Analyzing RNA Distribution USing optimal Transport

Emma Chen, Doron Haviv, Dana Pe'er

Methods

To leverage the subcellular resolution of imaging-based spatial transcriptomics data and compare the spatial organization of transcripts between cells, we use a similarity metric based on optimal transport theory, specifically the Fused Gromov-Wasserstein (FGW) distance. This approach compares the spatial relationships of transcripts while accounting for their relative distance to references like the cell boundary, yielding a robust comparison metric that is invariant to rigid transformations (translation and rotation). The flexible featurization approach also enables considering transcripts from multiple genes while accounting for each transcript's gene identity.

Cell Featurization

Each cell k is featurized as a point cloud of size n_k , which includes all transcripts and reference points. Each point i in this cloud is characterized by two components: a feature value $h_{k,i} \in \mathbb{R}$ and a spatial coordinate vector $\mathbf{x}_{k,i} \in \mathbb{R}^2$. The feature value distinguishes points of different types, such as transcripts of different genes or reference points from the cell or nuclear boundaries. The spatial coordinate vector captures the subcellular position of the point within the cell. We can then represent the cell k as $\mu_k = (\mathbf{h}_k, \mathbf{D}_k)$, where \mathbf{h}_k is the n_k -dimensional vector containing all scalar feature values and \mathbf{D}_k is the $n_k \times n_k$ intra-cell inter-point distance matrix computed from $\{\mathbf{x}_{k,i}\}_{i=1}^{n_k}$ (for instance, $\mathbf{D}_{k,ij}$ represents the distance between points $\mathbf{x}_{k,i}$ and $\mathbf{x}_{k,j}$).

We will now detail how exactly we implement this featurization. The user must first specify which genes are of interest and which spatial references are available (cell boundary, nuclear boundary, etc). For each available reference, the user may also optionally specify the number of points to use during featurization. Accordingly, the model will sample points uniformly along the available reference boundaries, so that we have a total set of n_k points for each cell k that includes transcript points and reference points. Each point already has a spatial coordinate from the raw spatial transcriptomics data. Each point is then assigned a feature value according to its type, with the number of unique feature values equal to the number of distinct genes plus the number of available references. The actual feature values is arbitrary as long as the minimum distance between distinct values is sufficiently large (we will explain this in more detail when discussing the loss). Note that because we featurize references such as the cell boundary or the nuclear boundary, we capture not just spatial relationships between RNA transcripts, but also the relative position of transcripts to these references (e.g. transcripts might be clustered together in two cells, but the cluster could be centered in one cell and polarized to one side of the cell in the other).

Fused Gromov-Wasserstein Distance

The similarity in subcellular RNA distribution between two cells, represented as $\mu_1 = (\mathbf{h}_1, \mathbf{D}_1)$ and $\mu_2 = (\mathbf{h}_2, \mathbf{D}_2)$, is quantified by the Fused Gromov-Wasserstein (FGW) distance. The FGW distance seeks an optimal transport plan $\pi \in \mathbb{R}^{n_1 \times n_2}$ that matches the points of one cell to the points of the other. This plan minimizes a cost function that is a weighted combination of two

terms: a **Feature Fidelity Term** (Wasserstein) that prevents matching points with different feature values and a **Structural Preservation Term** (Gromov-Wasserstein) that minimizes the distortion in intra-cell inter-point distances.

The optimal transport plan π is found by solving the following optimization problem:

$$\mathcal{L}_{\text{FGW}}(\pi) = \underbrace{\alpha \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \pi_{ij} |h_{1,i} - h_{2,j}|^2}_{\text{Feature Fidelity Term}} + \underbrace{(1 - \alpha) \sum_{i,k=1}^{n_1} \sum_{j,l=1}^{n_2} \pi_{ij} \pi_{kl} |\mathbf{D}_{1,ik} - \mathbf{D}_{2,jl}|^2}_{\text{Structural Preservation Term}}$$

Here, the Feature Fidelity Term uses the squared difference between the scalar feature values to penalize matching points with different feature values. In our experiments, we set distinct feature values to be far apart so that feature fidelity is a hard requirement as the model tries to minimize distortions in spatial structure. The parameter $\alpha \in [0, 1]$ controls the balance between the two terms.

The optimization is performed under the constraints of a standard optimal transport problem:

$$\min_{\pi \in \Pi} \mathcal{L}_{\text{FGW}}(\pi)$$

where Π is the set of joint probability distributions with fixed marginals \mathbf{r}_1 and \mathbf{r}_2

$$\Pi = \left\{ \pi \in \mathbb{R}_+^{n_1 \times n_2} \mid \pi \mathbf{1}_{n_2} = \mathbf{r}_1, \quad \pi^\top \mathbf{1}_{n_1} = \mathbf{r}_2 \right\}$$

By default, we use uniform measures $\mathbf{r}_1 = 1/n_1 \mathbf{1}$ and $\mathbf{r}_2 = 1/n_2 \mathbf{1}$. However, users can manipulate the marginals to be non-uniform measures depending on how much they care about spatial relationships being preserved for transcripts of particular genes or how much they care about preserving distances to a set of specific reference points.

The final FGW distance is given by $d_{\text{FGW}}(\mu_1, \mu_2) = \min_{\pi \in \Pi} \mathcal{L}_{\text{FGW}}(\pi, \alpha)$. This distance essentially captures the minimum possible amount of distortion in subcellular spatial structure as we match points from one cell to points from another cell while respecting the point type.

This approach can be adapted to capture geometric relationships at other scales. For instance, in the case of cancer lesions, we would represent each lesion as a point cloud of cells. Then, the feature value would be gene expression vectors of cells instead of distinct scalar values, encouraging matching between cells with similar transcriptomic phenotypes.