Distributional Matrix Completion for Gene Perturbation Prediction

Shengyi Li [1] Yiqun T. Chen [1,2]

[1]Department of Biostatistics [2]Department of Computer Science

Introduction

Goal: Predict **distributional** gene expression responses to perturbations

Challenges: *Unpaired* measurements from cell lysis, heterogeneous effects, and generalization to *unseen* perturbations

Prior work: Mostly target mean expression changes; few model full distributions at high computational cost

Solution: Nearest-neighbor search in perturbation embeddings + Zero-inflated quantile estimation + Copula modeling

Method

Motivation: Our work MAPLE builds on Dist-NN (Feitelberg et al., 2024), a recent method that imputes missing 1D distributions by finding similar rows based on observed entries and computing their Wasserstein barycenter using the quantile function.

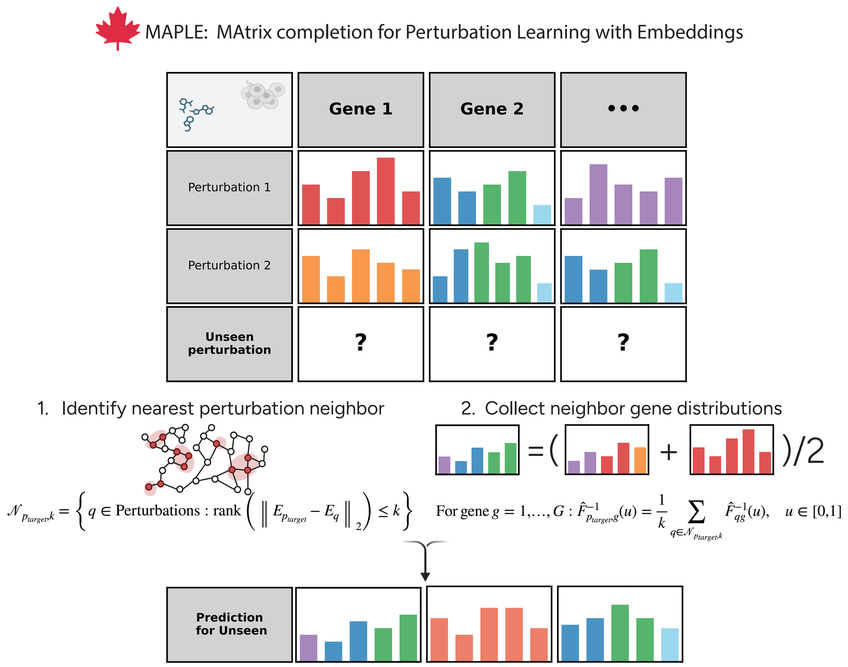

Figure 1. **Overview of MAPLE.**

- Identify perturbation neighbors via pretrained genetic (Chen and Zou, 2025) or chemical embeddings (Chithrananda et al., 2020)

$$\mathcal{N}_{p_{target},k} = \left\{ q \in \text{Perturbations} : \text{rank}\left(\left\| E_{p_{target}} - E_q \right\|_2 \right) \leq k \right\}$$

- Estimate marginal quantiles from neighbors

$$\text{For gene } g = 1,\ldots,G, \ \hat{F}^{-1}_{p_{target},g}(u) = \frac{1}{k} \sum_{q \in \mathcal{N}_{p_{target},k}} \hat{F}^{-1}_{qg}(u), \ u \in [0,1]$$

  - Non-parametric quantile estimation:

$$\hat{F}^{-1}_{qg}(u) = \begin{cases} 0, & u < \pi_0 \\ \hat{F}^{-1}_{qg,+}\left(\dfrac{u - \pi_0}{1 - \pi_0} \right), & u \geq \pi_0 \end{cases}, \ u \in [0,1]$$

  where $\pi_0 = \Pr(X_{qg} = 0)$, $\hat{F}_{qg,+}$ is the empirical CDF of non-zero values

- Sample univariate expression from the estimated marginal

$$\hat{X}_{p_{\text{target}},g} = \hat{F}^{-1}_{p_{\text{target}},g}(U_g), \ U_g \overset{i.i.d.}{\sim} \text{Unif}(0,1)$$

- Incorporate gene-gene correlations with copula models

$$\hat{X}'_{p_{\text{target}},g} = \hat{F}^{-1}_{p_{\text{target}},g}(U'_g), \ \text{where } U'_g \sim \text{Copula-structured Unif}(0,1)$$

  - Optimal transport (OT) copula:

$$U_{\text{OT}} = T^*_{\#}(U_{\text{Gauss}}), \ \text{where } T^*_{\#}\mu_{\text{Gauss}} = \mu_{\text{Emp}}$$

  where μ_{Gauss} is the distribution of $U_{\text{Gauss}} = \Phi(Z_g)$ with $Z_g \sim \mathcal{N}(0, \hat{\Sigma}_{\text{control}})$, and μ_{Emp} is the empirical joint rank distribution of neighbor perturbations.

Evaluation for each perturbation:

Let G be the number of genes, I the number of cells, and Y_g, $\hat{Y}_g$ the true and predicted distributions for gene g under a given perturbation.

- *Univariate Wasserstein-2 distance* (W_2^g): $\sqrt{\sum_{g=1}^{G} W_2^2(Y_g, \hat{Y}_g)}$
- *Multivariate Wasserstein-2 distance* (W_2^{mv}): $W_2(\mathbf{Y}, \hat{\mathbf{Y}})$, where $\mathbf{Y} = (Y_1, \ldots, Y_G)$
- *Coefficient of determination* (r^2): $1 - \frac{\sum_{g=1}^{G}(\mu_g - \hat{\mu}_g)^2}{\sum_{g=1}^{G}(\mu_g - \bar{\mu})^2}$, where $\mu_g = \frac{1}{I}\sum_{i=1}^{I} y_{g,i}$, $\hat{\mu}_g = \frac{1}{I}\sum_{i=1}^{I} \hat{y}_{g,i}$, and $\bar{\mu} = \frac{1}{G}\sum_{g=1}^{G} \mu_g$

Note: Non-parametric quantile estimation + OT copula is the optimal combination. The framework flexibly supports alternative quantile models (e.g., zero-inflated parametric) and copulas (e.g., Gaussian, empirical), with adaptable pretrained embeddings and neighbor sets.

Results & Discussion

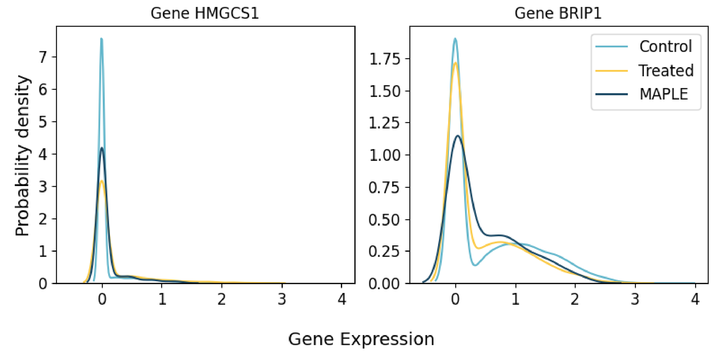

Figure 2. Marginal expression distributions for two DE genes treated by Givinostat.

Method	sci-Plex3			Adamson		
	$\overline{\mathbf{W}_2^{\text{mv}}}$	$\overline{\mathbf{W}_2^{g}}$	$\overline{\mathbf{r}^2}$	$\overline{\mathbf{W}_2^{\text{mv}}}$	$\overline{\mathbf{W}_2^{g}}$	$\overline{\mathbf{r}^{2*}}$
MAPLE (Non-parametric + OT copula)	9.4	**4.0**	0.72	**24.3**	**6.9**	0.58
Control baseline	12.0	7.4	0.30	31.2	8.5	N/A
chemCPA (Hetzel et al., 2022)	11.7	7.4	0.67	30.5	7.5	0.53
Biolord (Piran et al., 2024)	**8.1**	7.6	**0.86**	25.2	7.4	**0.61**

Table 1. **Performance comparison** on the sci-Plex3 (Srivatsan et al., 2020) (reported for 10 μM dosage) and Adamson (Adamson et al., 2016) datasets, evaluated over 10 nearby perturbations. Lower Wasserstein distances and higher $\overline{\mathbf{r}^2}$ values (control-centered for the Adamson dataset) indicate better reconstruction. **Bold**: best.

Our distributional model, combining flexible marginals with copulas, improves $\overline{W_2^g}$ and $\overline{r^2}$ over prior baselines like chemCPA (Hetzel et al., 2022). It also remains competitive with Biolord (Piran et al., 2024), a deep generative latent factor model, in $\overline{W_2^{\text{mv}}}$ for modeling joint dependencies.

Future work will improve both the estimation of gene-gene correlation structures and the theoretical foundations of our approach.

References

Britt Adamson, Thomas M Norman, Marco Jost, Min Y Cho, James K Nuñez, Yuwen Chen, Jacqueline E Villalta, Luke A Gilbert, Max A Horlbeck, Marco Y Hein, et al. A multiplexed single-cell crispr screening platform enables systematic dissection of the unfolded protein response. *Cell*, 167(7): 1867–1882, 2016.

Yiqun Chen and James Zou. Simple and effective embedding model for single-cell biology built from chatgpt. *Nature Biomedical Engineering*, 9(4): 483–493, 2025.

Seyone Chithrananda, Gabriel Grand, and Bharath Ramsundar. Chemberta: large-scale self-supervised pretraining for molecular property prediction. *arXiv preprint arXiv:2010.09885*, 2020.

Jacob Feitelberg, Kyuseong Choi, Anish Agarwal, and Raaz Dwivedi. Distributional matrix completion via nearest neighbors in the wasserstein space. *arXiv preprint arXiv:2410.13112v2*, 2024.

Leon Hetzel, Simon Böhm, Niki Kilbertus, Stephan Günnemann, Mohammad Lotfollahi, and Fabian J Theis. Predicting cellular responses to novel drug perturbations at a single-cell resolution. In *NeurIPS 2022*, 2022.

Zoe Piran, Niv Cohen, Yedid Hoshen, and Mor Nitzan. Disentanglement of single-cell data with biolord. *Nature Biotechnology*, 42(11):1678–1683, 2024.

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.