Reflection on MLCB 2025 & Poster Overview

Shengyi Li, ScM in Biostatistics
Advised by Dr. Yiqun Chen

Genomic DS Working Group, Oct 14th

Conference Reflections

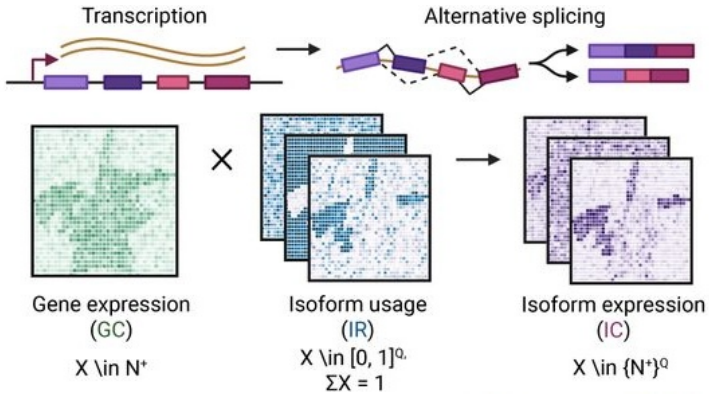
- Broad ML applications across structural biology, systems biology, drug design, algorithms, and benchmarking
- Extensive genomics work on RNA splicing, SNPs, regulatory networks, and promoter modeling
- Rapid growth of language models in protein and genomic modeling, embedding learning, and reasoning
- Highly engaging atmosphere with active discussions and close interactions among participants

Interesting Research Highlights

A computational framework for mapping isoform landscape and regulatory mechanisms from spatial transcriptomics data

Jiayu Su, Yiming Qu, Megan Schertzer, Haochen Yang, Jiahao Jiang, Tenzin Lhakhang, Theodore M. Nelson, Stella Park, Qiliang Lai, Xi Fu, Seung-won Choi, David A. Knowles, Raul Rabadan

doi: https://doi.org/10.1101/2025.05.02.651907



Screenshot of Figure 1

Scientific questions:

- Which isoforms show spatial heterogeneity in expression?
- Which RNA-binding proteins (RBPs) may regulate these patterns?

Interesting Research Highlights

A computational framework for mapping isoform landscape and regulatory mechanisms from spatial transcriptomics data

Jiayu Su, Yiming Qu, Megan Schertzer, Haochen Yang, Jiahao Jiang, Tenzin Lhakhang, Theodore M. Nelson, Stella Park, Qiliang Lai, Xi Fu, Seung-won Choi, David A. Knowles, Raul Rabadan

doi: https://doi.org/10.1101/2025.05.02.651907

Statistical formulation:

Use the Hilbert-Schmidt Independence Criterion (HSIC) to detect spatial dependence of isoform usage, and the conditional HSIC to assess whether it is explained by RBP expression.

Main datasets:

10x Visium mouse brain dataset (a coronal section) and Slide-seqV2 adult mouse hippocampus dataset (a single tissue section) for cross-platform validation,
and a Visium data of human DLPFC for cross-species validation.

Distributional Matrix Completion for Gene Perturbation Prediction

Scientific question:
Understanding gene expression responses to perturbations is fundamental to drug development and the study of cellular regulation, but large-scale experiments are costly and time-consuming.

Can we predict these perturbation responses computationally?

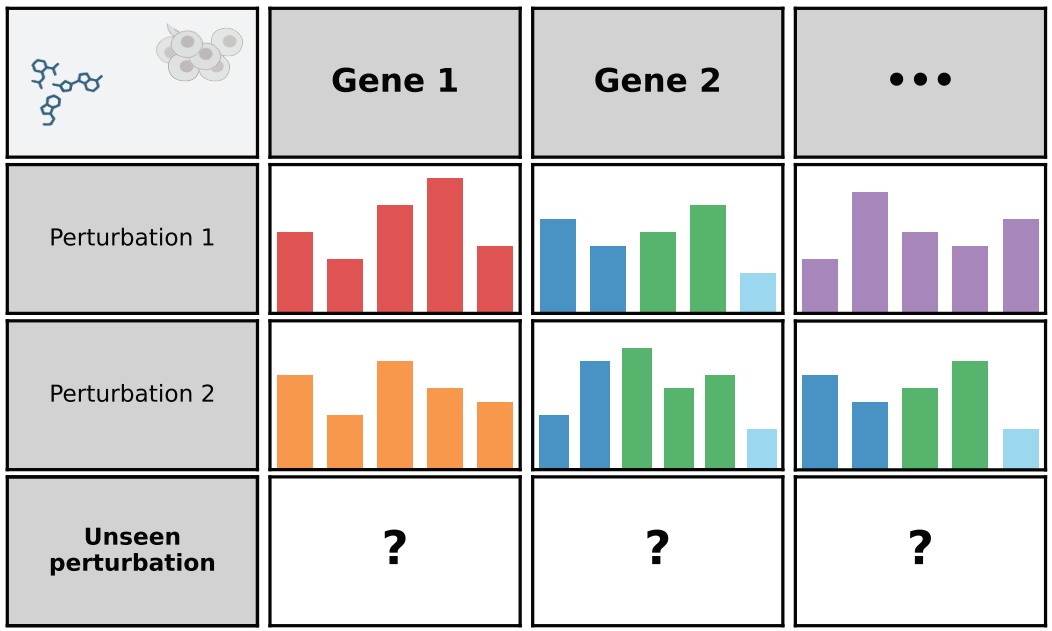
Motivation:
Predicting post-perturbation gene expression is challenging, as most ML frameworks capture only mean shifts rather than response heterogeneity, while those modeling full distributions are computationally expensive.

We aim to efficiently and accurately predict full expression distribution.

Data source and representation

Data source: Single-cell Perturb-seq data

Data representation:



Partially observed perturbation–gene response matrix

Background

p-Wasserstein distance:

The minimum cost of transporting one distribution to another, when the cost function is the p-th power of the Euclidean distance.

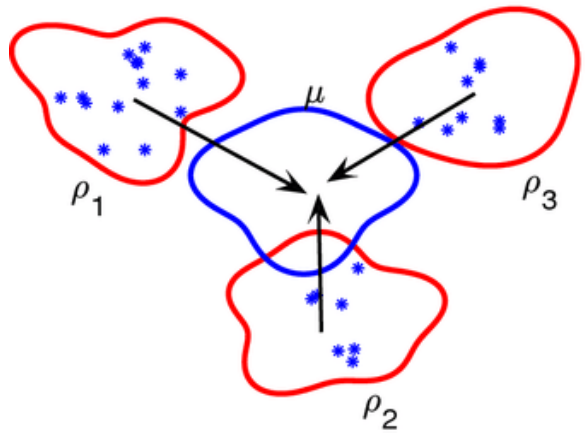
$$W_p^p(\mu,\nu) = \inf_{\pi\in\Pi(\mu,\nu)} \int_{\mathbb{R}^d\times\mathbb{R}^d} \|x-y\|^p \,\mathrm{d}\pi$$

In the 1D setting, where μ and ν are two probability measures on $\mathbb{R}$ with finite second moments, we have

$$W_p^p(\mu,\nu) = \int_0^1 |F_\mu^{-1}(t) - F_\nu^{-1}(t)|^p \,\mathrm{d}t$$

2-**Wasserstein barycenter:**

$$\mu^\star = \arg\min_{\mu} \sum_{i=1}^{N} W_2^2(\mu, \mu_i)$$



Source: [9].

In the 1D setting, we have [2]

$$F_{\mu^\star}^{-1}(t) = \frac{1}{N} \sum_{i=1}^{N} F_{\mu_i}^{-1}(t), \quad t \in [0, 1]$$

Related work:
Dist-NN[5], a recent method that imputes missing 1D distributions by finding similar rows from the observed entries and computing their Wasserstein barycenter in the target column using the quantile function.

Hypothesis:
Perturbations are considered similar when the knocked-out gene or the drug treatment is close in the embedding space.

Method Overview

For each unseen perturbation,

- **Stage 1: Marginal Prediction**
  - Identifies similar perturbations using pretrained embeddings
  - Estimates gene-wise expression quantiles based on these neighbors

- **Stage 2: Multivariate Prediction**
  - Integrates univariate distributions via a copula model to capture gene-gene dependencies

Method: Marginal Prediction

- Identify perturbation neighbors via pretrained genetic [3] or chemical [4] embeddings

$$\mathcal{N}_{p_{\text{target}},k} = \left\{\text{Top-}k \text{ nearest perturbations in the embedding space}\right\}$$

Method: Marginal Prediction

- Identify perturbation neighbors via pretrained genetic [3] or chemical [4] embeddings

$$\mathcal{N}_{p_{\text{target}},k} = \left\{\text{Top-}k \text{ nearest perturbations in the embedding space}\right\}$$

- Estimate marginal quantiles from neighbors

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

Method: Marginal Prediction

- Identify perturbation neighbors via pretrained genetic [3] or chemical [4] embeddings

  $$\mathcal{N}_{p_{\text{target}},k} = \left\{\text{Top-}k \text{ nearest perturbations in the embedding space}\right\}$$

- Estimate marginal quantiles from neighbors

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

  - Zero-inflated quantile estimation:

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

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

Method: Marginal Prediction

- Identify perturbation neighbors via pretrained genetic [3] or chemical [4] embeddings

$$\mathcal{N}_{p_{\text{target}},k} = \left\{\text{Top-}k \text{ nearest perturbations in the embedding space}\right\}$$

- Estimate marginal quantiles from neighbors

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

  - Zero-inflated quantile estimation:

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

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

- Sample univariate expression from the estimated quantiles (Probability integral transform: $F_X(X) \sim \text{Unif}(0,1)$)

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

Brief Introduction to Copulas

Sklar's Theorem. For any d-dimensional CDF F with marginals $F_1, \ldots, F_d$, there exists a copula C such that

$$F(x_1, \ldots, x_d) = C\big(F_1(x_1), \ldots, F_d(x_d)\big),$$

for all $x_i \in [-\infty, \infty]$ and $i = 1, \ldots, d$.

Writing $U_i = F_i(X_i)$, we have

$$(U_1, \ldots, U_d) \sim C$$

Method: Multivariate Prediction

- Incorporate gene-gene correlations with copula models

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

Method: Multivariate Prediction

- Incorporate gene-gene correlations with copula models

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

  - Optimal transport (OT) copula:

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

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

Evaluations

For each perturbation,

- *Univariate 2-Wasserstein distance* (W_2^g):
  $\sqrt{\sum_{g=1}^{G} W_2^2(Y_g, \hat{Y}_g)}$
- *Multivariate 2-Wasserstein 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
  $\mu_g = \frac{1}{I}\sum_{i=1}^{I} y_{g,i}$

Results

Method	**sci-Plex3** [8]			**Adamson** [1]		
	$\overline{\mathbf{W}_2^{\mathrm{mv}}}$	$\overline{\mathbf{W}_2^{g}}$	$\overline{\mathbf{r}^2}$	$\overline{\mathbf{W}_2^{\mathrm{mv}}}$	$\overline{\mathbf{W}_2^{g}}$	$\overline{\mathbf{r}^2}^*$
MAPLE (ZI+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 [6]	11.7	7.4	0.67	30.5	7.5	0.53
Biolord [7]	**8.1**	7.6	**0.86**	25.2	7.4	**0.61**

Table 1: Performance comparison

Next steps

- Refine the estimation of gene-gene correlation structures
- Extend the theoretical foundations of distributional matrix completion for this framework

References

[1] Britt Adamson et al. "A multiplexed single-cell CRISPR screening platform enables systematic dissection of the unfolded protein response". In: *Cell* 167.7 (2016), pp. 1867–1882.

[2] Martial Agueh and Guillaume Carlier. "Barycenters in the Wasserstein space". In: *SIAM Journal on Mathematical Analysis* 43.2 (2011), pp. 904–924.

[3] Yiqun Chen and James Zou. "Simple and effective embedding model for single-cell biology built from ChatGPT". In: *Nature Biomedical Engineering* 9.4 (2025), pp. 483–493.

[4] Seyone Chithrananda, Gabriel Grand, and Bharath Ramsundar. "ChemBERTa: large-scale self-supervised pretraining for molecular property prediction". In: *arXiv preprint arXiv:2010.09885* (2020).

[5] Jacob Feitelberg et al. "Distributional Matrix Completion via Nearest Neighbors in the Wasserstein Space". In: *arXiv preprint arXiv:2410.13112v2* (2024).

[6] Leon Hetzel et al. "Predicting Cellular Responses to Novel Drug Perturbations at a Single-Cell Resolution". In: *NeurIPS 2022*. 2022.

[7] Zoe Piran et al. "Disentanglement of single-cell data with biolord". In: *Nature Biotechnology* 42.11 (2024), pp. 1678–1683.

[8] Sanjay R Srivatsan et al. "Massively multiplex chemical transcriptomics at single-cell resolution". In: *Science* 367.6473 (2020), pp. 45–51.

[9] Esteban G. Tabak, Giulio Trigila, and Wenjia Zhao. "Conditional density estimation and simulation through optimal transport". In: *Machine Learning* 109.4 (2020), pp. 665–688. DOI: 10.1007/s10994-019-05866-3. URL: https://doi.org/10.1007/s10994-019-05866-3.

Thank You!