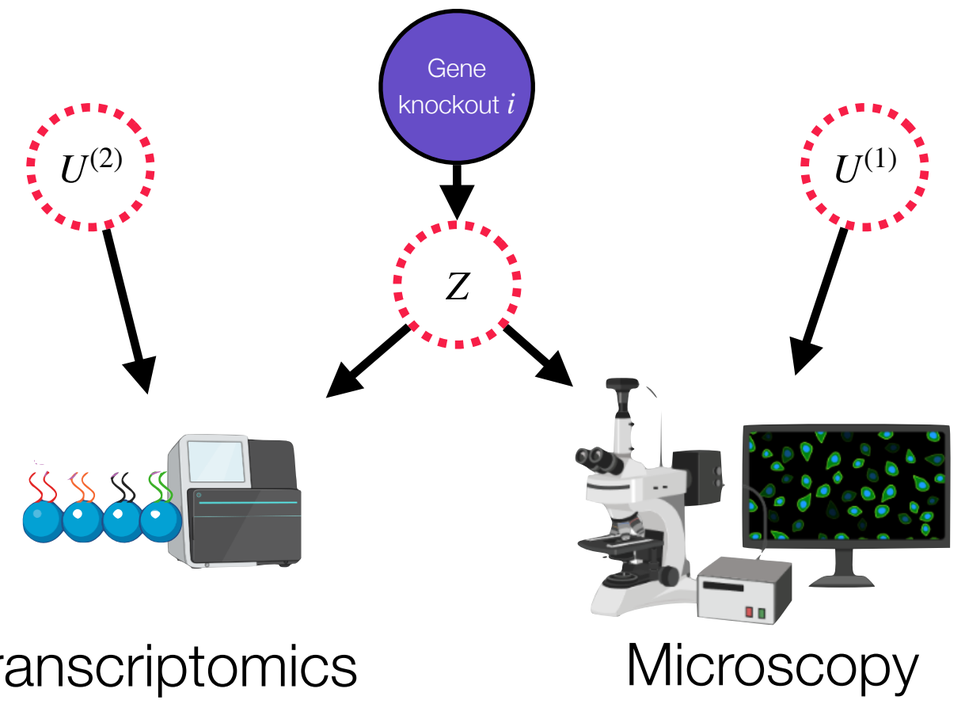
Propensity Score Alignment of Unpaired Multimodal Data

Valence Labs

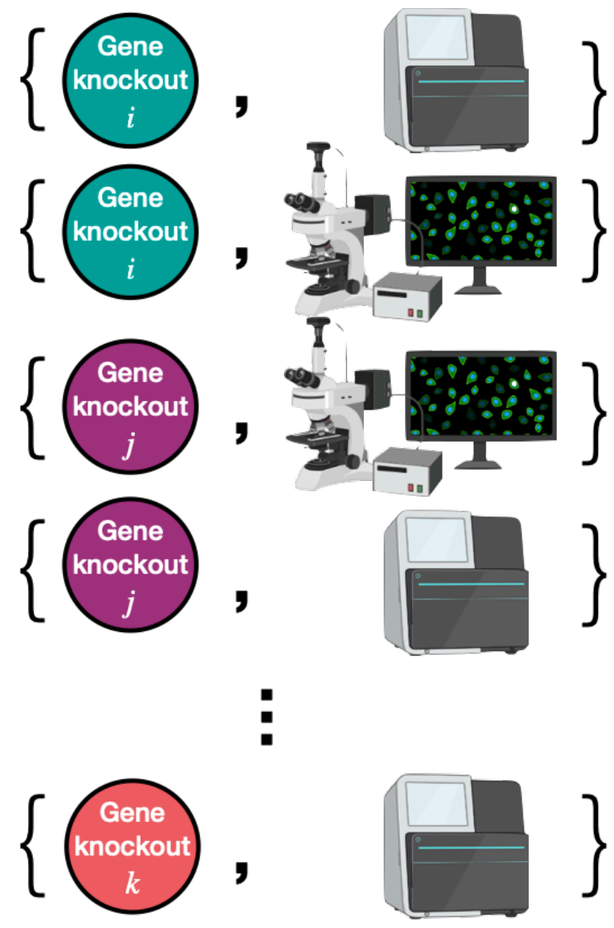
Johnny Xi, Jason Hartford

Setup: Latent Z represents biological information
U represent modality-specific measurement noise
Perturbations only perturb the common Z

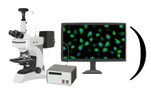


Problem: Given a dataset of unpaired (unimodal) samples, can we approximately reconstruct a matching?

Two big challenges:
1. How do we define a distance metric between different observation spaces?
2. How do we ensure that the metric focuses on biologically relevant info?

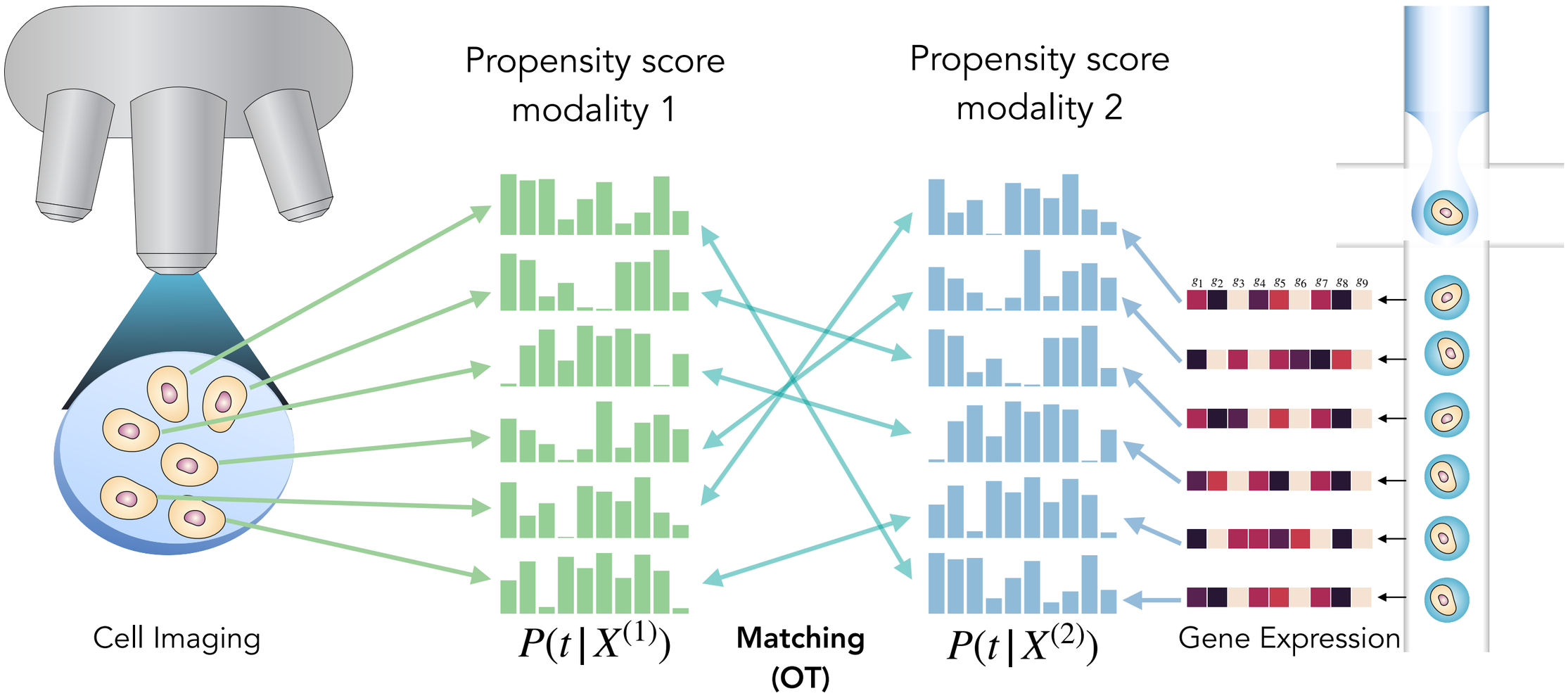


Methodology: Propensity Score Alignment

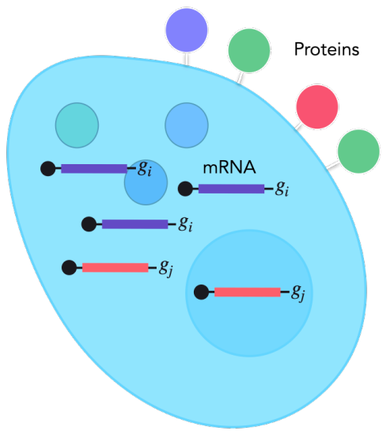
p(KO Gene i | ) = p(KO Gene i | ) = p(KO Gene i | Z) = "Propensity Score"

Match on estimated propensity score (trained separately for each modality)

1) same space; 2) contains **all "perturbable" information**: $I(t, Z^{(t)} \mid \pi(X^{(t)})) = 0$



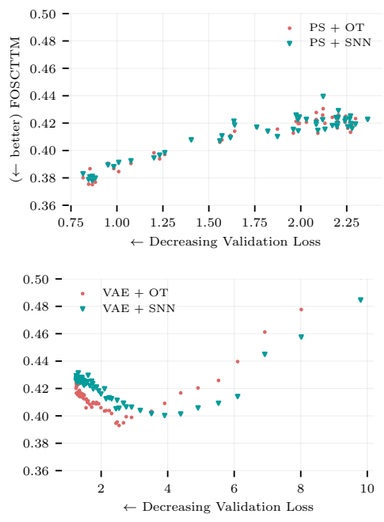
Results - CITE-Seq NeurIPS Challenge



Method	FOSCTTM (Median (Q1, Q3))	Trace (Median (Q1, Q3))
SCOT	0.4596	0.0200
GLUE+SNN	0.4412	0.0362
GLUE+OT	0.5309	0.0323
VAE+SNN	0.3816 (0.3760, 0.3822)	0.0612 (0.0588, 0.0634)
VAE+OT	0.3953 (0.3912, 0.4045)	0.0814 (0.0777, 0.8895)
PS+SNN	0.3126 (0.3121, 0.3160)	0.0941 (0.0880, 0.0989)
PS+OT	**0.3049 (0.3008, 0.3078)**	**0.1163 (0.1093, 0.1250)**

(Metrics require ground truth pairing)

Propensity score loss proxies matching metrics even without ground truth (i.e., in practice)



Train good classifiers, get good matchings!

Matching leads to strong cross modality translation.

Method	R^2 (MSE) (Med (Q1, Q3))	R^2 (Unbiased)[1] (Med (Q1, Q3))
Rand	0.1383 (0.1372, 0.1402)	0.1727 (0.1701, 0.1731)
VAE+OT	0.1493 (0.1179, 0.1724)	0.1142 (0.0786, 0.1594)
PS+OT	0.2174 (0.2062, 0.2228)	0.2331 (0.2069, 0.2504)
True Pairs	0.2243 (0.2234, 0.2257)	N/A[2]

Some gradient subtly: you need two samples to get unbiased gradient estimates

$$\mathcal{L}(\theta) := \sum_i (x_i^{(1)} - M_i f_\theta(x_j^{(2)}))^2.$$

$$\nabla\mathcal{L}(\theta) \approx -2\left(x_i^{(1)} - f_\theta(\dot{x}_j^{(2)})\right)\nabla_\theta f_\theta(\ddot{x}_j^{(2)}) \qquad \dot{x}_j^{(2)}, \ddot{x}_j^{(2)} \sim P(x_j^{(2)} | x_i^{(1)}).$$

Full paper here ➡