PROPENSITY SCORE ALIGNMENT OF UNPAIRED MULTIMODAL DATA

Johnny Xi [*]
University of British Columbia
johnny.xi@stat.ubc.ca

Jason Hartford
Valence Labs
jason@valencelabs.com

ABSTRACT

Multimodal representation learning techniques typically rely on paired samples to learn common representations, but paired samples are challenging to collect in fields such as biology where measurement devices often destroy the samples. This paper presents an approach to address the challenge of aligning unpaired samples across disparate modalities in multimodal representation learning. We draw an analogy between potential outcomes in causal inference and potential views in multimodal observations, which allows us to use Rubin's framework to estimate a common space in which to match samples. Our approach assumes we collect samples that are experimentally perturbed by treatments, and uses this to estimate a propensity score from each modality, which encapsulates all shared information between a latent state and treatment and can be used to define a distance between samples. We experiment with two alignment techniques that leverage this distance—shared nearest neighbours (SNN) and optimal transport (OT) matching—and find that OT matching results in significant improvements over state-of-the-art alignment approaches in both a synthetic multi-modal setting and in real-world data from NeurIPS Multimodal Single-Cell Integration Challenge.

1 INTRODUCTION

Large-scale multimodal representation learning techniques such as CLIP (Radford et al., 2021) have lead to remarkable improvements in zero-shot classification performance and have enabled the recent success in conditional generative models. However, the effectiveness of multimodal methods hinges on the availability of *paired* samples—such as images and their associated captions—across data modalities. This reliance on paired samples is most obvious in the InfoNCE loss (Gutmann & Hyvärinen, 2010; van den Oord et al., 2018) used in CLIP (Radford et al., 2021) which explicitly learns representations to maximize the true matching between images and their captions.

While paired image captioning data is abundant on the internet, paired multimodal data is often challenging to collect in scientific experiments. Take, for instance, experiments in biology, where unpaired data are the norm for technical reasons: RNA sequencing, protein expression assays, and the collection of microscopy images for cell painting experimental assays are all destructive processes. Because of this, we cannot collect multiple different measurements from the same cell, and can only explicitly group cells by their experimental condition.

If we could accurately match unpaired samples across modalities, we could use the aligned samples as proxies for paired samples and apply existing multimodal representation learning techniques. However, matching across modalities is challenging for two reasons. Firstly, observations from various modalities often exist in entirely different spaces. For instance, microscopy images of cells are in pixel space, while gene expression assays provide data in the form of mRNA abundance counts. Secondly, even if measurements were taken in the same space, determining an appropriate distance metric for matching remains a non-trivial task. For example, when matching images of distinct cells in different microscopy images, we would ideally want a metric that is more sensitive to features describing the underlying state of the cell than features related to biologically-irrelevant cell appearance, such as the orientation of the cell.

---

[*]Work done during an internship at Valence labs.

To understand the matching problem more abstractly, we can think of the multiple modalities as potential "views", $x^{(1)}(z) \in \mathcal{X}^{(1)}, x^{(2)}(z) \in \mathcal{X}^{(2)}$ of the same underlying latent state, $z \in \mathcal{Z}$, where we only get to observe one view for any individual unit i (e.g. an individual cell in our motivating example). Observing, or accurately inferring z would address the first issue by providing a common space $\mathcal{Z}$ in we could match samples. However, inferring the latent variable is hopelessly underspecified without making strong assumptions on the data generating processes of both modalities (Xi & Bloem-Reddy, 2023). Furthermore, for complex systems such as biological samples, z may still be extremely high-dimensional, and as a result, the second challenge remains: what is the "right" metric is for measuring similarity between samples in this space?

We address both challenges in this paper by appealing to classical ideas from Rubin (1974), in the case where we additionally observe a label for each unit $t \in \mathcal{T}$, e.g., indexing an experiment that it belongs to. By making the assumption that t perturbs the observations via their shared latent state, we identify an observable link between modalities with the same underlying z. Our key observation is that the propensity score, defined as $p(t|z)$, satisfies three properties: 1) It provides a common space for matching, 2) it maximally coarsens z, retaining only relevant information and 3) under certain assumptions, it is estimable from observations of individual modalities alone (Proposition 2.1).

In practice, the propensity score is straightforward to estimate: one simply trains two classifiers—one for each modality—to predict which treatment was applied to the respective modalities, and then match across modalities based on the similarity between predicted treatments within each treatment group. The proposed methodology is generic across modalities—it can be applied to match observations between any modality for which a classifier can be efficiently trained (e.g., text, video, tabular). However, we cannot naively use bipartite matching, as the same sample unit does not appear in both modalities. To address this, we use soft matching techniques to estimate the missing modality for each sample unit by combining multiple observations. We experiment with two different matching approaches from the recent literature: shared nearest neighbours (SNN) matching (Lance et al., 2022; Cao & Gao, 2022) and optimal transport (OT) matching Villani (2009); Schiebinger et al. (2019), and find that OT matching with distances defined on the proposenity score lead to significant improvement on real world multi-modal matching tasks from the NeurIPS Multimodal Single-Cell Integration Challenge Lance et al. (2022). We found similarly large improvements on synthetic image-based benchmarks.

Finally, we show how our matching procedure can be used for cross-modality prediction tasks. We estimate a function that translates between modalities by integrating over the coupling matrix that is output by our propensity score alignment procedure. Our approach uses a two-sample stochastic gradient estimator to get an unbiased estimate of the gradient of the resulting loss. In our experiments, this approach gave very strong performance, even outperforming methods that had access to the ground truth matching.

2 PROPENSITY SCORE MULTI-MODAL MATCHING

We consider multimodal settings where there exist two potential views, $X^{(e,t)} \in \mathcal{X}^{(e)}$ from two different modalities indexed by $e \in \{1, 2\}$ and experiment t that perturbs a shared latent state of these observations. The observations will typically be in very different spaces: for example, $\mathcal{X}^{(1,t)}$ may be the space of images of cells under a microscope, and $\mathcal{X}^{(2,t)}$ may be the space of gene expression data. This process defines a jointly distributed random variable $(X^{(1,t)}, X^{(2,t)}, e, t)$, from which we observe only a single modality, its index, and treatment assignment, $\{x_i^{(e_i,t_i)}, e_i, t_i\}_{i=1}^n$ (we will always denote random variables by upper-case letters, and samples by their corresponding lower-case letter). Our aim is to match or estimate the samples from the missing modality that would correspond to the realization of the missing random variable. Importantly, we match *within* treatment groups, and the treatment variable is only used to learn a common space in which to match.

We assume each modality arises from a common latent random variable Z as follows,

$$t \sim P_T, \quad Z^{(t)} \mid t \sim P_Z^{(t)}, \quad U^{(e)} \sim P_U^{(e)}, \quad U^{(e)} \perp\!\!\!\perp Z,$$
$$t \perp\!\!\!\perp U^{(e)}, \quad t \not\!\perp\!\!\!\perp Z, \quad X^{(e,t)} = f^{(e)}(Z^{(t)}, U^{(e)}), \tag{1}$$

where t indexes the experimental perturbations, $t \not\!\perp\!\!\!\perp Z$ ensures that t has a non-trival affect on the distribution of the latent variables, and we can take $t = 0$ to represent a base environment.
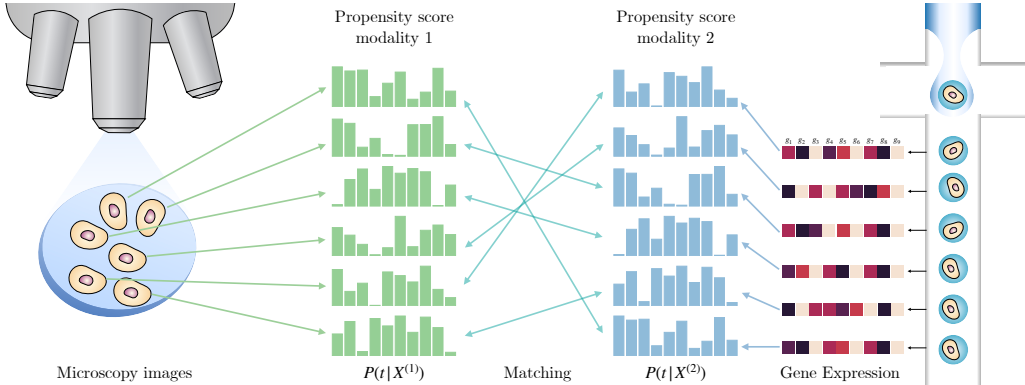
Figure 1: Our problem setting and proposed solution. We assume that two observed modalities (e.g., Microscopy images and RNA expression data) are generated in pairs according to some common latent variable, z_i, but we only observe the respective modalities separately. We further assume the latents are experimentally perturbed by a treatment, t, and we get observations from both modalities from a control state and at least one interventional state. Under our assumptions, training separate classifiers that predict the intervention label reveals the shared information $p(t|z_i)$, which allows us to re-pair the observed disconnected modalities.

Note that the structural equations f^e are deterministic after accounting for the randomness in Z and U: it represents purely the measurement process that captures the latent state. For example, in a microscopy image, this would be the microscope and camera that maps a cell to pixels. The modality specific noise variables, $U^{(1)} \perp\!\!\!\perp U^{(2)}$, play the role of measurement noise and modality-specific factors of variation: e.g. $U^{(1)}$, could describe the layout and orientation of cells on a slide.

Under this model, if Z were observable, an optimal matching can be constructed by simply matching the modalities with the most similar Z. However, Z is latent, and inference on the model described by Equation (1) is arguably more difficult than the matching problem itself due to theoretical difficulties such as identifiability and disentangling Z from the modality-specific noise terms $u^{(e)}$.

Instead, we see that the interventions t provide an observable link between the modalities, thereby revealing information about $Z^{(t)}$. Specifically, we use the propensity score with respect to t,

$$\pi(z) := P(t|Z = z) \in [0, 1]^{T+1}, \tag{2}$$

as a proxy for the latent Z. Now, although we cannot compute this directly as Z is latent, we make the observation that if $f^{(e)}$ is injective for $e = 1, 2$, then we can compute the compute the propensity score from each of the observed modalities, since it will be that $\pi(Z^{(t)}) = \pi(X^{(e,t)})$, regardless of $e = 1$ or $e = 2$. Not only does the propensity score reveal shared information, classical causal inference (Rubin, 1974) states that it captures *all* information shared between the latent and treatment, and does so minimally, in terms of having minimum dimension and entropy. We collect these observations into the following proposition.

Proposition 2.1. *In the model described by Equation* (1)*, further assume that $f^{(e)}$ are injective for $e = 1, 2$. Then, the propensity scores in either modality is equal to the propensity score given by $Z^{(t)}$, i.e., $\pi(X^{(1,t)}) = \pi(X^{(2,t)}) = \pi(Z^{(t)})$ as random variables. This implies*

$$I(t, Z^{(t)} \mid \pi(Z^{(t)})) = I(t, Z^{(t)} \mid \pi(X^{(t)})) = 0, \tag{3}$$

for each t, where I is the mutual information. Furthermore, any other function $b(Z^{(t)})$ satisfying $I(t, Z^{(t)} \mid b(Z^{(t)})) = 0$ is such that $\pi(Z^{(t)}) = f(b(Z^{(t)}))$.

The proof can be found in the Appendix. The above shows that computing the propensity score on either modality is equivalent to computing it on the unobserved shared latent, which captures all the shared information observable in t. The final statement implies that it is of minimal dimension and entropy, and thus it discards the modality-specific information that may be counterproductive to matching.

Number of Perturbations Note that point-wise equality of the propensity score $\pi(z_1) = \pi(z_2)$ does not necessarily imply equality of the latents $z_1 = z_2$, due to potential non-injectivity. In

fact, suppose that $Z^{(t)}$ is supported in $\mathbb{R}^d$. The propensity score $\pi(z)$ maps into the T-dimensional simplex, where T is the number of perturbations, which can be parametrized by coordinates in $\mathbb{R}^T$ (essentially, the logits of the classifier). Under mild conditions, the propensity score necessarily compresses information about $Z^{(t)}$ if the latent dimension exceeds the number of perturbations, echoing impossibility results from the causal representation learning literature (Squires et al., 2023).

Proposition 2.2. *Suppose that $P_Z^{(t)}$ has a smooth density $p(z|t)$ for each $t = 0, \dots, T$. Then, if $T < d$, the propensity score π, restricted to its strictly positive part, is non-injective.*

The proof can be found in the Appendix. Note the above only states an impossibility result when $T < d$. When $T \geq d$, it can be seen from the proof of Proposition 2.2 that the injectivity of the propensity score depends on the injectivity of the following expression in z:

$$f(z) = \begin{bmatrix} \log(p(z|t=1)) - \log(p(z|t=0)) \\ \vdots \\ \log(p(z|t=T)) - \log(p(z|t=0)) \end{bmatrix}, \tag{4}$$

which then depends on the latent process itself. If the above is non-injective, this is a fundamental indeterminacy that cannot be resolved without making strong assumptions on point-wise latent variable recovery, as we have already seen that the propensity score contains the maximal shared information, as shown in Proposition 2.1. Nonetheless, note that f in Equation (4) is injective if any of the subset of its entries are, clearly outlining the benefits of collecting data from a larger number of perturbations for matching.

3 METHODOLOGY

For the remainder, we will drop (e, t) from our notation, and use (x_i, t_i) to denote observations from modality 1, and (x_j, t_j) to denote observations from modality 2 whenever it does not cause confusion to do so. Given a multimodal dataset with observations $\{(x_i, t_i)\}_{i=1}^{n_1}$ and $\{(x_j, t_j)\}_{j=1}^{n_2}$, we wish to compute a matching matrix (or coupling) between the two modalities. We define a $n_1 \times n_2$ matching matrix M where M_{ij} represents the likelihood of x_i being matched to x_j. We always perform matching only within observations with the same value of t, so that in practice we obtain a matrix M_t for each t.

Our method approximates the propensity scores by training separate classifiers that predicts t given x for each modality. We denote the estimated propensity score by π_i and π_j respectively, where

$$\pi_i \approx \pi(x_i) = P(T = t \mid X_i^{(e,t)} = x_i). \tag{5}$$

This results in the transformed datasets $\{\pi_i\}_{i=1}^{n_1}$ and $\{\pi_j\}_{j=1}^{n_2}$, where π_i, π_j are in the T dimensional simplex. We use this correspondence to compute a cross-modality distance function:

$$d(x_i, x_j) := d'(\pi_i, \pi_j), \tag{6}$$

where in practice, we typically compute the Euclidean distance in $\mathbb{R}^T$ of the logit-transformed classification scores. Given this distance function, we rely on existing matching techniques to constructing a matching matrix. In our experiments, we found OT matching gave the best performance, but we also evaluated Shared Nearest Neighbour matching; details can be found in Appendix B.

Optimal Transport Matching The propensity score distance allows us to easily compute a cost function associated with transporting mass between modalities, $c(x_i, x_j) = d'(\pi_i \pi_j)$. Let p_1, p_2 denote the uniform distribution over $\{\pi_i\}_{i=1}^{n_1}$ and $\{\pi_j\}_{j=1}^{n_2}$ respectively. Discrete OT aims to solve the problem of optimally redistributing mass from p_1 to p_2 in terms of incurring the lowest cost. Let $C_{ij} = c(x_i, x_j)$ denote the $n_1 \times n_2$ cost matrix. The Kantorovich formulation of optimal transport aims to solve the following constrained optimization problem:

$$\min_M \sum_i^{n_1} \sum_j^{n_2} C_{ij} M_{ij}, \quad M_{ij} \geq 0, \qquad M\mathbf{1} = p_1, \quad M^\top \mathbf{1} = p_2. \tag{7}$$

This is a linear program, and for $n_1 = n_2$, it can be shown that the optimal solution is a bipartite matching between $\{\pi_i\}_{i=1}^{n_1}$ and $\{\pi_j\}_{j=1}^{n_2}$. We refer to this as exact OT; in practice we add an entropic

regularization term, resulting in a soft matching, that ensures smoothness and uniqueness, and can be solved efficiently using Sinkhorn's algorithm. Entropic OT takes the following form:

$$\min_M \sum_i^{n_1} \sum_j^{n_2} C_{ij} M_{ij} - \lambda H(M), \qquad M_{ij} \geq 0, \quad M\mathbf{1} = p_1, \quad M^\top \mathbf{1} = p_2, \qquad (8)$$

where $H(M) = -\sum_{i,j} M_{ij} \log(M_{ij})$, the entropy of the joint distribution implied by M. This approach regularizes towards a higher entropy solution, which has been shown to have statistical benefits (Genevay et al., 2018), but nonetheless for small enough λ serves as a computationally appealing approximation to exact OT.

Cross-modality prediction Given a matching matrix M, we can interpret this as the probability that each sample, i, from modality (1) is matched to sample j in modality (2); that is $M_{i,j} = P(x_j^{(2)}|x_i^{(1)})$. As a downstream task, and as an evaluation metric for M, we can use this matching to estimate a cross-modal prediction model, f_θ, that maps from modality (2) to (1) by minimizing the following loss,

$$\mathcal{L}(\theta) := \sum_i (x_i^{(1)} - M_i f_\theta(x_j^{(2)}))^2. \qquad (9)$$

However, this requires evaluating f_θ for all n_2 examples from modality (2) for each of the n_1 examples in modality (1). Of course, we can avoid this cost with stochastic gradient descent by sampling from modality (2) via $M_{i\cdot}$ for each training example (1), but to get an unbiased estimate of $\nabla_\theta \mathcal{L}$, we need two samples from modality (2) for each sample from modality (1),

$$\nabla \mathcal{L}(\theta) \approx -2 \left(x_i^{(1)} - f_\theta(\dot{x}_j^{(2)}) \right) \nabla_\theta f_\theta(\ddot{x}_j^{(2)}) \qquad \dot{x}_j^{(2)}, \ddot{x}_j^{(2)} \sim P(x_j^{(2)}|x_i^{(1)}). \qquad (10)$$

By taking two samples as in equation (10), we get an unbiased estimator of $\nabla \mathcal{L}(\theta)$, whereas a single sample would have resulted in optimizing an upper-bound on equation (9); for details, see Hartford et al. (2017) where a similar issue arises in the gradient of their causal effect estimator.

4 EXPERIMENTS

We evaluate our proposed methodology on two datasets. The first is a synthetic interventional image dataset generated satisfying the assumptions of Equation (1). The second is a real-world single-cell CITE-seq data obtained from the NeurIPS 2021 Multimodal single-cell data integration competition (Lance et al., 2022), which provides a ground truth matching by allowing for a small number of cell surface proteins to be measured simultaneously to RNA sequencing. Note the CITE-seq dataset is not interventional—we use the cell type as the classification target t instead. All details are made available in the Appendix. In both cases the ground truth matching is known to make evaluation possible, but hidden during training.

Evaluation Metrics We report three evaluation metrics. The trace metric computes the average mass that the matching matrix places on the true matches (higher is better), and FOSCTTM reports the Fraction Of Samples Closer Than the True Match (FOSCTTM) ((Demetci et al., 2022), (Liu et al., 2019)) (lower is better, 0.5 corresponds to random guessing). For synthetic images where we know the ground true latent, we report the MSE to the true latent after matching. Full details of these metrics are given in Appendix F.4.

Modality Prediction We also consider one metric that examines whether matched samples are useful for downstream tasks. For this, we chose the cross-modality prediction task of predicting protein levels from RNA expression in the CITE-seq dataset. To do this, we train a supervised learning model (in our case, a 2-layer MLP with either MSE loss or the unbiased gradients loss in Section 3) using samples generated by different matching matrices M. As baselines, we also train on pairs from the ground truth matching ($M_{ii} = 1$) and random sampling ($M_{ij} = 1/n$). In view of the performance observed in Table 3 and Table 1, we only considered M resulting from matching with OT using the propensity score and VAE embeddings.

---

[1]This refers to training the predictor with the unbiased MSE (Equation (9)). The evaluation is still using R^2 as defined by the standard MSE.

[2]The standard and unbiased MSE are equivalent in this case.

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	0.0612
	(0.3760, 0.3822)	(0.0588, 0.0634)
VAE+OT	0.3953	0.0814
	(0.3912, 0.4045)	(0.0777, 0.8895)
PS+SNN	0.3126	0.0941
	(0.3121, 0.3160)	(0.0880, 0.0989)
PS+OT	**0.3049**	**0.1163**
	(0.3008, 0.3078)	**(0.1093, 0.1250)**

Table 1: Results on CITE-seq data. Metrics are computed on a held out test set averaged over a total of 45 cell types (groups) with a varying amount of observations per group. As such, the average trace can be difficult to interpret (for example, the maximum trace is near 1, corresponding to a near-perfect matching within a group with only 7 cells in the test set), but OT matching on propensity scores outperformed the other methods within each group as well as on average.

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

Table 2: Modality prediction results on CITE-seq data. Entries represent R^2 computed on a held-out test set consisting of ground-truth pairs, which is computed as $1 - \mathrm{MSE}(\hat{x}, x)/\mathrm{MSE}(\bar{x}, x)$, where $\hat{x}$ are the predicted protein levels, and $\bar{x}$ is the mean value of the validation set. Note random sampling, especially with the unbiased gradient loss, achieves a positive R^2 against $\bar{x}$ because it samples randomly *within* t, which results in learning group-level statistics.

Results We checkpoint the classifier and VAEs at the lowest validation loss and report their metrics on a held out test set over 10 random seeds. Note that this does not necessarily select the embedding model that exhibits optimal matching, but instead the best model to select without validating against ground truth matching metrics, see Figure 3. scGLUE is trained according to the public implementation, which uses learning rate reduction and early stopping strategies. SCOT is a non-iterative approach and thus we directly report its results.

We found that OT matching on propensity scores consistently outperforms other methods on all metrics, typically followed by SNN matching on propensity scores, or OT matching on VAE embeddings. Furthermore, both the VAE and propensity score, which leverage experimental label information, tend to perform better. This suggests that using the propensity score as embeddings for matching, and using OT to perform the final matching both independently improve matching performance. Curiously, SCOT performs well on the MSE metric for image data, but only places slightly above random in the trace metric. This indicates that it matches scenes with similar latent coordinates, without placing any significant weight on the exact ground truth. This suggests that exact matches may not be necessary for a method to be useful for downstream tasks.

Surprisingly, using the propensity score embeddings with OT matching appears to improve generalization in the modality prediction task (Table 2), when using the unbiased MSE over a model trained on ground truth pairings (indeed, we observed that the ground truth model had a lower training loss, but higher test loss). This reveals an unexpected benefit of (soft) matching: we can sample from the conditional to minimize the loss Equation (9), which, with a suitable M, results in improved generalization compared to the naive MSE. This hinges strongly on the quality of M—the M resulting from the VAE embeddings results in worse generalization, and in practice the quality of the matching, at least in absolute terms, remains hidden.

5 CONCLUSION

This work presents a simple algorithm for aligning unpaired data from different modalities. The method is both very general—only requiring a classifier to be trained on each modality—and highly effective for matching, which we show both theoretically and empirically. As a downstream task, we demonstrate the effectiveness of the resulting matchings for cross modality prediction, which leads to *better* generalization than the ground truth matching on the dataset we evaluated. We suspect that this improved generalization is the result of implicitly enforcing invariance to modality specific information, but more work is needed to evaluate the conditions under which this improved generalization occurs.

REFERENCES

Ahuja, K., Mahajan, D., Wang, Y., and Bengio, Y. Interventional causal representation learning. In *ICML*, 2023.

Almahairi, A., Rajeshwar, S., Sordoni, A., Bachman, P., and Courville, A. Augmented cyclegan: Learning many-to-many mappings from unpaired data. In *ICML*, 2018.

Amodio, M. and Krishnaswamy, S. Magan: Aligning biological manifolds. In *ICML*, 2018.

Buchholz, S., Rajendran, G., Rosenfeld, E., Aragam, B., Schölkopf, B., and Ravikumar, P. Learning linear causal representations from interventions under general nonlinear mixing. In *NeurIPS*, 2023.

Butler, A., Hoffman, P., Smibert, P., Papalexi, E., and Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature Biotechnology*, 36(5):411–420, 2018.

Cao, K., Gong, Q., Hong, Y., and Wan, L. A unified computational framework for single-cell data integration with optimal transport. *Nature Communications*, 13(1):7419, 2022.

Cao, Z.-J. and Gao, G. Multi-omics single-cell data integration and regulatory inference with graph-linked embedding. *Nature Biotechnology*, 40(10):1458–1466, 2022.

Demetci, P., Santorella, R., Sandstede, B., Noble, W. S., and Singh, R. Scot: single-cell multi-omics alignment with optimal transport. *Journal of Computational Biology*, 29(1):3–18, 2022.

Flamary, R., Courty, N., Gramfort, A., Alaya, M. Z., Boisbunon, A., Chambon, S., Chapel, L., Corenflos, A., Fatras, K., Fournier, N., Gautheron, L., Gayraud, N. T., Janati, H., Rakotomamonjy, A., Redko, I., Rolet, A., Schutz, A., Seguy, V., Sutherland, D. J., Tavenard, R., Tong, A., and Vayer, T. Pot: Python optimal transport. *Journal of Machine Learning Research*, 22(78):1–8, 2021.

Foster, A., Vezér, Á., Glastonbury, C. A., Creed, P., Abujudeh, S., and Sim, A. Contrastive mixture of posteriors for counterfactual inference, data integration and fairness. In *ICML*, 2022.

Genevay, A., Peyré, G., and Cuturi, M. Learning generative models with sinkhorn divergences. In *AISTATS*, 2018.

Gossi, F., Pati, P., Chouvardas, P., Martinelli, A. L., Kruithof-de Julio, M., and Rapsomaniki, M. A. Matching single cells across modalities with contrastive learning and optimal transport. *Briefings in Bioinformatics*, 24(3), 2023.

Gresele, L., Rubenstein, P. K., Mehrjou, A., Locatello, F., and Schölkopf, B. The incomplete rosetta stone problem: Identifiability results for multi-view nonlinear ica. In *UAI*, 2020.

Gutmann, M. and Hyvärinen, A. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *AISTATS*, 2010.

Hartford, J., Lewis, G., Leyton-Brown, K., and Taddy, M. Deep iv: A flexible approach for counterfactual prediction. In *ICML*, 2017.

Khemakhem, I., Kingma, D. P., Monti, R. P., and Hyvärinen, A. Variational autoencoders and nonlinear ICA: A unifying framework. In *AISTATS*, 2020.

Korsunsky, I., Millard, N., Fan, J., Slowikowski, K., Zhang, F., Wei, K., Baglaenko, Y., Brenner, M., Loh, P.-r., and Raychaudhuri, S. Fast, sensitive and accurate integration of single-cell data with harmony. *Nature Methods*, 16(12):1289–1296, 2019.

Lance, C., Luecken, M. D., Burkhardt, D. B., Cannoodt, R., Rautenstrauch, P., Laddach, A., Ubingazhibov, A., Cao, Z.-J., Deng, K., Khan, S., et al. Multimodal single cell data integration challenge: Results and lessons learned. In *NeurIPS 2021 Competitions and Demonstrations Track*, pp. 162–176, 2022.

Liu, J., Huang, Y., Singh, R., Vert, J.-P., and Noble, W. S. Jointly Embedding Multiple Single-Cell Omics Measurements. In *19th International Workshop on Algorithms in Bioinformatics (WABI 2019)*, 2019.

Liu, M.-Y., Breuel, T., and Kautz, J. Unsupervised image-to-image translation networks. *NeurIPS*, 2017.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

Rubin, D. B. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688, 1974.

Schiebinger, G., Shu, J., Tabaka, M., Cleary, B., Subramanian, V., Solomon, A., Gould, J., Liu, S., Lin, S., Berube, P., et al. Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming. *Cell*, 176(4):928–943, 2019.

Spivak, M. *Calculus on Manifolds: a Modern Approach to Classical Theorems of Advanced Calculus*. CRC press, 2018.

Squires, C., Seigal, A., Bhate, S. S., and Uhler, C. Linear causal disentanglement via interventions. In *ICML*, 2023.

Sturma, N., Squires, C., Drton, M., and Uhler, C. Unpaired multi-domain causal representation learning. In *NeurIPS*, 2023.

Tong, A., Huang, J., Wolf, G., Van Dijk, D., and Krishnaswamy, S. Trajectorynet: A dynamic optimal transport network for modeling cellular dynamics. In *ICML*, 2020.

van den Oord, A., Li, Y., and Vinyals, O. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Villani, C. *Optimal Transport: Old and New*, volume 338. Springer, 2009.

von Kügelgen, J., Besserve, M., Wendong, L., Gresele, L., Kekić, A., Bareinboim, E., Blei, D. M., and Schölkopf, B. Nonparametric identifiability of causal representations from unknown interventions. In *NeurIPS*, 2023.

Xi, Q. and Bloem-Reddy, B. Indeterminacy in generative models: Characterization and strong identifiability. In *AISTATS*, 2023.

Yang, K. D., Belyaeva, A., Venkatachalapathy, S., Damodaran, K., Katcoff, A., Radhakrishnan, A., Shivashankar, G., and Uhler, C. Multi-domain translation between single-cell imaging and sequencing data using autoencoders. *Nature Communications*, 12(1):31, 2021.

Zhu, J.-Y., Park, T., Isola, P., and Efros, A. A. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pp. 2223–2232, 2017.

A   RELATED WORK

Unpaired Data   Dealing with unpaired data is important in translation tasks, either for images (Liu et al., 2017; Zhu et al., 2017; Almahairi et al., 2018), or more recently for biological modalities (Amodio & Krishnaswamy, 2018; Yang et al., 2021). In particular, Yang et al. (2021) formalize this setting as being generated by a shared underlying latent variable, a useful modelling assumption for theoretical identification in multi-view or multi-modal representation learning (Gresele et al., 2020; Sturma et al., 2023). Our work also exploits this modelling assumption for theoretical justification, but, unlike previous works, our method does not depend on performing inference over the latent space. Specifically, since we do not require modality translation, we do not require an encoder-decoder structure (Yang et al., 2021), which embeds modality-specific reconstruction information in the latent space that can be detrimental for the simpler matching task.

Optimal Transport Matching   Optimal transport is commonly used to solve a similar problem in single-cell biology known as cell trajectory inference. In this case, $X^{(1)}$ to $X^{(2)}$ are random variables measured at different time points in a shared (metric) space, and OT matching can be achieved by minimizing the metric (Schiebinger et al., 2019; Tong et al., 2020). Recent work (Demetci et al., 2022) tackles our multi-modal alignment problem, when each modality is observed in a (separate) metric space, by using the Gromov-Wasserstein distance, which computes the difference between the metric evaluated within pairs of points from each modality (Demetci et al., 2022). In addition to this "pure" OT approach, Gossi et al. (2023) use OT on contrastive learning representations, though this approach requires matched pairs for training, while Cao et al. (2022) use OT in the latent space of a multi-modal VAE.

Perturbations and Heterogeneity   Many methods treat observation-level heterogeneity as another dimension to globally integrate, even when cluster labels are observed (Butler et al., 2018; Korsunsky et al., 2019; Foster et al., 2022). This is sensible when clusters correspond to noise rather than the signal of interest. However, it is well known in causal representation learning that heterogeneity—particularly heterogeneity arising from different perturbations—has theoretical benefits (Khemakhem et al., 2020; Squires et al., 2023; Ahuja et al., 2023; Buchholz et al., 2023; von Kügelgen et al., 2023). There, the benefits (weakly) increase with the number of perturbations, which is also true of our setting (Proposition 2.2). In the context of aligning unpaired data, only Yang et al. (2021) leverage this heterogeneity in their method. Specifically, they require VAE representations to classify experimental labels in addition to reconstructing modalities, while our method only requires a representation to classify labels. Yang et al. (2021) treat the classifier objective as a regularizer, but our theory suggests that it is responsible for their matching performance. In our experiments, we found that requiring reconstruction led to worse matching performance with equivalent architectures.

B   SHARED NEAREST NEIGHBOURS MATCHING

Using the propensity score distance, we can compute nearest neighbours both within and between the two modalities. We follow Cao & Gao (2022) and compute the normalized shared nearest neighbours (SNN) between each pair of observations as the entry of the matching matrix. For each pair of observations $(\pi_i^{(1)}, \pi_j^{(2)})$, we define four sets:

- $\texttt{11}_{ij}$: the k nearest neighbours of $\pi_i^{(1)}$ amongst $\{\pi_i^{(1)}\}_{i=1}^{n_1}$. $\pi_i^{(1)}$ is considered a neighbour of itself.

- $\texttt{12}_{ij}$: the k nearest neighbours of $\pi_j^{(2)}$ amongst $\{\pi_i^{(1)}\}_{i=1}^{n_1}$.

- $\texttt{21}_{ij}$: the k nearest neighbours of $\pi_i^{(1)}$ amongst $\{\pi_j^{(2)}\}_{j=1}^{n_2}$.

- $\texttt{22}_{ij}$: the k nearest neighbours of $\pi_j^{(2)}$ amongst $\{\pi_j^{(2)}\}_{j=1}^{n_2}$. $\pi_j^{(2)}$ is considered a neighbour of itself.

Intuitively, if $\pi_i^{(1)}$ and $\pi_j^{(2)}$ correspond to the same underlying propensity score, their nearest neighbours amongst observations from each modality should be the same. This is measured as a set difference between $\texttt{11}_{ij}$ and $\texttt{12}_{ij}$, and likewise for $\texttt{21}_{ij}$ and $\texttt{22}_{ij}$. Then, a modified Jaccard index

is computed as follows. Define

$$J_{ij} = |\texttt{11}_{ij} \cap \texttt{12}_{ij}| + |\texttt{21}_{ij} \cap \texttt{22}_{ij}|, \tag{11}$$

the sum of the number of shared neighbours measured in each modality. Then, we compute the following Jaccard distance to populate the unnormalized matching matrix:

$$\tilde{M}_{ij} = \frac{J_{ij}}{4k - J_{ij}}, \tag{12}$$

where notice that $4k = |\texttt{11}_{ij}| + |\texttt{12}_{ij}| + |\texttt{21}_{ij}| + |\texttt{22}_{ij}|$, since each set contains k distinct neighbours, and thus $0 \le \tilde{M}_{ij} \le 1$, as with the standard Jaccard index. Then, we normalize each row to produce the final matching matrix:

$$M_{ij} = \frac{\tilde{M}_{ij}}{\sum_{i=1}^{n_1} \tilde{M}_{ij}}. \tag{13}$$

Note M_{ij} is always well defined because $\pi_i^{(1)}$ and $\pi_j^{(2)}$ are always considered neighbours of themselves.

Lemma B.1. $\tilde{M}_{ij}$ *has at least one non-zero entry in each of its rows and columns for any number of neighbours $k \ge 1$.*

Proof. We prove that $J_{ij} > 0$ for at least one j in each i, which is equivalent to $\tilde{M}_{ij} > 0$. Fix an arbitrary i. $\texttt{21}_{ij}$ by definition is the same set for every j. By the assumption of $k \ge 1$ it is non-empty, so there exists $\pi_{j^*}^{(2)} \in \texttt{21}_{ij}$. Since $\pi_{j^*}^{(2)}$ is a neighbour of itself, we have $\pi_{j^*}^{(2)} \in \texttt{22}_{ij^*}$, showing that $J_{ij^*} > 0$. The same reasoning applied to $\texttt{11}$ and $\texttt{12}$ also shows that J_{ij} for at least one i in each j. $\qquad\square$

C OMITTED PROOFS

C.1 PROOF OF PROPOSITION 2.1

Proof. Let $x^{(e,t)}$ denote the observed modality and $z^{(t)}, u^{(e)}$ be the unique corresponding latent values. By injectivity,

$$
\begin{aligned}
\pi(x^{(e,t)}) &= P(t|X^{(e,t)} = x^{(e,t)}) \\
&= P(t|Z^{(t)} = z^{(t)}, U^{(e)} = u^{(e)}) \\
&= P(t|Z^{(t)} = z^{(t)}) = \pi(z^{(t)}),
\end{aligned}
\tag{14}
$$

for $e = 1, 2$, since we assumed $U^{(e)} \perp\!\!\!\perp t \mid Z^{(t)}$ in Equation (1). Since this holds pointwise, it shows that $\pi(X^{(1,t)}) = \pi(X^{(2,t)}) = \pi(X^{(t)}) = \pi(Z^{(t)})$ as random variables. Now, a classical result of Rubin (1974) gives that $Z^{(t)} \perp\!\!\!\perp t \mid \pi(Z^{(t)})$, and that for any other function b (a *balancing score*) such that $Z^{(t)} \perp\!\!\!\perp t \mid b(Z^{(t)})$, we have $\pi(Z^{(t)}) = g(b(Z^{(t)}))$. The first property written in information theoretic terms yields,

$$I(t, Z^{(t)} \mid \pi(Z^{(t)})) = I(t, Z^{(t)} \mid \pi(X^{(t)})) = 0, \tag{15}$$

since $\pi(X^{(t)}) = \pi(Z^{(t)})$ as random variables, as required. $\qquad\square$

C.2 PROOF OF PROPOSITION 2.2

Proof. In what follows, we write π to be the restriction to its domain where it is strictly positive. The i-th dimension of the propensity score can be written as

$$(\pi(z))_i = p(t = i|z) = \frac{p(z|t = i)p(t = i)}{\sum_{i=0}^{T} p(z|t = i)p(t = i)}, \tag{16}$$

which, when restricted to be strictly positive, maps to the relative interior of the T-dimensional probability simplex. Consider the following transformation:

$$h(\pi(z))_i = \log\left(\frac{(\pi(z))_i}{(\pi(z))_0}\right) \tag{17}$$

$$= \log(p(z|t=i)) - \log(p(z|t=0)) + C, \tag{18}$$

where $C = \log(p(t=i)) - \log(p(t=0))$ is constant in z, and that $h(\pi(z))_0 \equiv 0$. Ignoring the constant first dimension, we can view h as an invertible map to $\mathbb{R}^T$. Under this convention, the map $h \circ \pi : \mathbb{R}^d \to \mathbb{R}^T$ is smooth (log is smooth, and the densities are smooth by assumption). Since it is smooth, it cannot be injective if $T < d$ (Spivak, 2018). Finally, since h is bijective, this implies that π cannot be injective. $\qquad\square$

D   DISCUSSION OF ASSUMPTIONS

Our matching results rely on two key modelling assumptions,

- (A1): t has a non-trivial effect on Z, but does not affect $u^{(e)}$, implying that interventions are able to target the common underlying process without changing modality-specific properties.
- (A2): Injectivity of $f^{(e)}$ in the propensity score derivation, $f^{(e)}(z, u) = f^{(e)}(z', u') \Rightarrow (z, u) = (z', u')$, which implies each modality contains complete information about the underlying process.

(A1) is justifiable insofar as modalities represent different measurements of an isolated system, such as in biological studies where Z might correspond to an underlying cell state and modalities refer to different single-cell measurements. However, in practice, even different measurement devices may be more or less sensitive to the biological variation implied by t. We discuss how to weaken this assumption in Appendix D.1 below.

(A2) requires that no information about the latent state is lost in the observation process. Note that the injectivity is in the sense of f as a function of both u and z, which allows observations that have a shared z but differ by their value in u, and the function remains injective. For example, rotated images with the exact same content can have a shared z, but remain injective due to the rotation being captured in u.

D.1   RELAXING ASSUMPTION 1

Consider the propensity score

$$\pi(x^{(e,t)}) = P(t|X^{(e,t)} = x^{(e,t)}) \tag{19}$$

where we do not necessarily require $U^{(e)} \perp\!\!\!\perp t \mid Z^{(t)}$, and thus we obtain

$$\pi(x^{(1,t)}) = P(t|Z^{(t)} = z^{(t)}, U^{(1)} = u^{(1)}) \neq P(t|Z^{(t)} = z^{(t)}, U^{(2)} = u^{(2)}) = \pi(x^{(2,t)}), \tag{20}$$

see the proof of Proposition 2.1 for details.

Suppose that the two observed modalities are indeed generated by a shared $\{z_i\}_{i=1}^n$, but where the indices of modality 2 are potentially permuted, and with values differing by modality specific information:

$$\{x_i^{(1,t)} = f^{(1)}(z_i, u_i^{(1)})\}_{i=1}^n, \{x_j^{(2,t)} = f^{(2)}(z_2, u_j^{(2)})\}_{j=1}^n, \tag{21}$$

where $j = \pi(i)$ denotes a permutation of the sample indices. Under (A1), we would be able to find some j such that $\pi(x_i^{(1,t)}) = \pi(x_j^{(2,t)})$ for each i.

Matching via OT can allow us to relax (A1) in a very particular way. Consider the simple case where $t \in \{0, 1\}$, so that π can be written in a single dimension, e.g., $P(t = 1|X^{(e,t)} = x^{(e,t)}) \in [0, 1]$. In this case, exact OT is equivalent to sorting $\pi(x_i^{(1,t)})$ and $\pi(x_j^{(2,t)})$, and matching the sorted versions 1-to-1. Under (A1), the sorted versions will be exactly equal. A relaxed version of (A1) that would
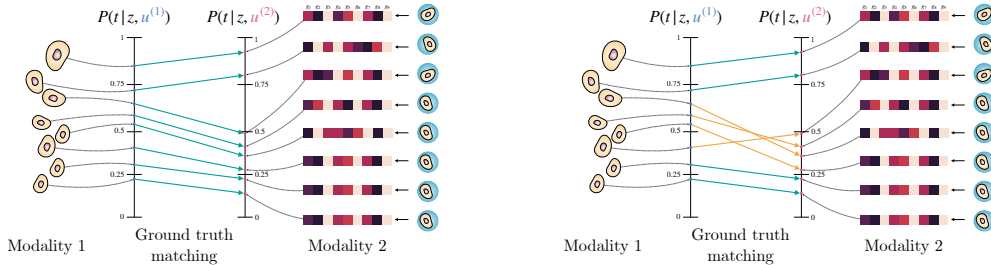
Figure 2: OT matching allows for t to have different effects on the modality specific information, here $u_i^{(1)}$ and $u_i^{(2)}$, as long as they can be written as transformations that preserve the relative order within modalities. Exact OT in 1-d always matches according to the relative ordering, and thus exhibits this type of "no crossing" behaviour shown in the figure on the left. The figure on the right shows a case where we would fail to correctly match across modalities because of the crossing shown in orange.

still result in the correct ground truth matching is to assume that t affects $U^{(1)}$ and $U^{(2)}$ differently, but that the difference is order preserving, or monotone. Denote $(\pi(x_i^{(1,t)}), \pi(x_i^{(2,t)})) = (\pi_i^{(1)}, \pi_i^{(2)})$ as the true pairing, noting that we use the same index i. We require the following:

$$(\pi_{i_1}^{(1)} - \pi_{i_2}^{(1)})(\pi_{i_1}^{(2)} - \pi_{i_2}^{(2)}) \geq 0, \quad \forall i_1, i_2 = 1, \ldots, n. \tag{22}$$

This says that, even if $\pi_i^{(1)} \neq \pi_i^{(2)}$, that their relative orderings will still coincide. Then, exact OT will still recover the ground truth matching. See Figure 2 for a visual example of this type of monotonicity. For example, suppose that t is a chemical perturbation of a cell, and thus $\pi_i^{(1)}, \pi_j^{(2)}$ can be seen as a measure of biological response to the perturbation, e.g., in a treated population, $\pi_{i_1} > \pi_{i_2}$ indicates samples i_1 had a stronger response than sample i_2, as perceived by the first modality indexed by i. Then, this monotonocity states that we should see the same $\pi_{j_1} > \pi_{j_2}$ in the other modality as well, if the samples i_1 and i_2 truly corresponded to j_1 and j_2.

We can see the monotonicity requirement Equation (22) as the monotonicity of the function with graph $(\pi_i^{(1)}, \pi_i^{(2)}) \in [0, 1]^2$. In higher dimensions, we require that the "graph" satisfies the following cyclic monotonicity property (Villani, 2009):

Definition D.1. The collection $\{(\pi_i^{(1)}, \pi_i^{(2)})\}_{i=1}^n$ is said to be c-cyclically monotone for some cost function c, if for any $n = 1, \ldots, N$, and any subset of pairs $(\pi_1^{(1)}, \pi_1^{(2)}), \ldots, (\pi_n^{(1)}, \pi_n^{(2)})$, we have

$$\sum_{n=1}^N c(\pi_n^{(1)}, \pi_n^{(2)}) \leq \sum_{n=1}^N c(\pi_n^{(1)}, \pi_{n+1}^{(2)}). \tag{23}$$

Importantly, we define $\pi_{n+1} = \pi_1$, so that the sequence represents a cycle.

Note in our setting, the OT cost function is the Euclidean distance, $c(x, y) = \|x - y\|_2$. It is known that the OT solution must satisfy cyclic monotonicity. Thus, if the true pairing is uniquely cyclically monotone, we can recover it with OT. However, unlike the 1-d case, we are not aware of more intuitive conditions under which we should expect cyclic monotonicity to hold.

E ADDITIONAL EXPERIMENTS

E.1 VALIDATION CROSS-ENTROPY MONITOR

When the assumptions are satisfied, the success of our propensity score method hinges on how well we can learn the conditional distributions $p(t \mid X_i^{(e,t)})$. Thus, the validation cross entropy can be used as a tractable metric that can serve as a model selection monitor for the matching performance, which would be unknown in practice. The effectiveness of such a monitor is validated in our experiments, even on real-world CITE-seq data, where we see that a lower validation loss typically
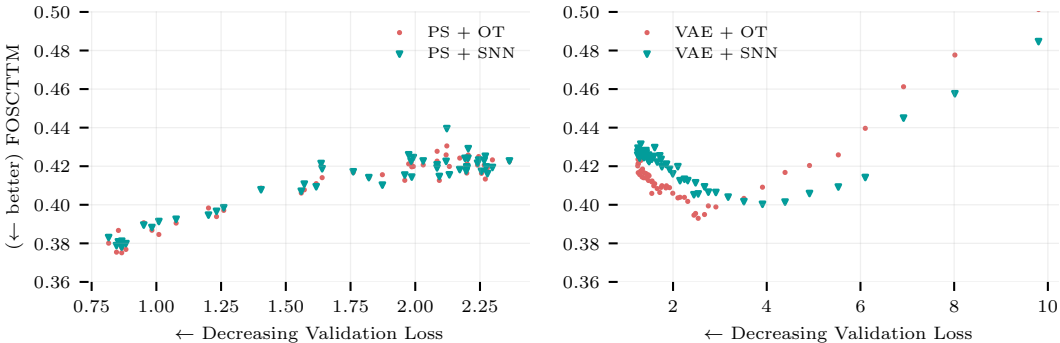
Figure 3: VAE and classifier validation metrics on the CITE-seq dataset. Notice that validation cross-entropy inversely tracks the ground truth matching metrics, and thus can be used as a proxy in practical settings where the ground truth is unknown. The same pattern does not hold for the VAE (Yang et al., 2021), which we suspect is because reconstruction is largely irrelevant for matching.

Method	MSE (Med (Q1, Q3))	Trace (Med (Q1, Q3)) $\times 10^{-3}$
SCOT	0.0354	0.5964
VAE+SNN	0.0622	3.116
	(0.0571, 0.0676)	(2.818, 3.213)
VAE+OT	0.0324	7.733
	(0.0316, 0.0350)	(7.473, 7.794)
PS+SNN	0.0552	7.924
	(0.0530, 0.0558)	(7.569, 9.504)
PS+OT	**0.0316**	**18.329**
	(0.0300, 0.0330)	**(17.068, 18.987)**
Rand	0.0709	N/A
	(0.0707, 0.0714)	

Table 3: Results on synthetic image data. Metrics are computed on a held out test set over 12 groups corresponding to interventions on the latent position, with approximately 1700 observations per group. A trace of 10 for example corresponds to a total weight of $10 \times 10^{-3} \times 1700 = 17$ on the true matching, out of 1700.

corresponds to higher matching performance which is not the case for the VAE (Figure 3). We suspect that this is precisely due to the additional requirement that the VAE minimize a reconstruction loss, in addition to the classification cross-entropy.

E.2 SYNTHETIC EXPERIMENTS

The results of the synthetic experiments with known ground truth are shown in Table 3. They were consistent with what we found on real data.

F EXPERIMENTAL DETAILS

F.1 BASELINES

In terms of the problem setting, the most closely related method is Gromov-Wasserstein optimal transport (SCOT) (Demetci et al., 2022), which explicitly computes a cross-modality cost for optimal transport matching by leveraging the local geometry within each modality. For the CITE-seq data, graph-linked VAE (scGLUE), which has access to specific biological metadata that connects the

biological modalities, can also be used for matching.[3] However, neither of the preceding approaches use the experimental label information, which is a key component of our method. In fact, the only method to our knowledge that leverages such information is the VAE approach developed in Yang et al. (2021), which also minimizes a classification loss from the latent space, which we will consider our main baseline. For all methods besides SCOT, we use both SNN and OT to perform the final matching in the latent space. Note that when combined with OT, our VAE baseline also resembles Cao et al. (2022).

F.2 MODEL DETAILS

For our method and the VAE, we use the general architecture of a linear classification head on top of a suitable encoder. The architecture of this encoder is always the same between the VAE and our method. For the image dataset, we use the convolutional neural network from Yang et al. (2021) as the encoder, and for the CITE-seq dataset, we use fully-connected MLPs for both RNA (top 200 PC's) and protein data. Optimization was performed using Adam with a one-cycle learning rate scheduler. We use the standard cross-entropy loss to train our classifiers for propensity score estimation. For other methods, we use existing implementations with suggested default settings.

F.3 MATCHING DETAILS

Both SNN and OT use the Euclidean distance function to determine neighbours and compute the cost matrix, respectively. SCOT uses the correlation distance by default, and we found that this resulted in better performance than Euclidean distance. We use only a single neighbour for SNN matching, which interestingly resulted in the best performance. Both SCOT and OT solve the entropic regularized OT, for these we use a regularization parameter of 0.05.

F.4 EVALUATION METRICS

We can evaluate how well samples are matched using the ground truth provided by our datasets. In these cases, the dataset sizes are necessarily balanced, so that $n = n_1 = n_2$. In each case, the metric is a function of a $n \times n$ matching matrix M, which recall we compute within samples with the same t. Our reported results are then averaged over each cluster. We randomize the order of the datasets before performing the matching to avoid pathological cases.

Trace Metric Assuming the given indices correspond to the true matching, we can compute the average weight on correct matches, which is the normalized trace of M:

$$\frac{1}{n}\text{Tr}(M) = \frac{1}{n}\sum_{i=1}^{n} M_{ii}. \tag{24}$$

As a baseline, notice that a uniformly random matching that assigns $M_{ij} = 1/n$ for each cell yields $\text{Tr}(M) = 1$ and hence will obtain a metric of $1/n$. This metric however does not capture potential failure modes of matching. For example, exactly matching one sample, while adversarially matching dissimiliar samples for the remainder also yields a trace of $1/n$, which is equal to that of a random matching.

Latent MSE Because the image dataset is synthetic, we have access to the ground truth latent values that generated the images, $\mathbf{z} = \{z_i\}_{i=1}^{n}$. We compute matched latents as $M\mathbf{z}$, the mean projection according to the matching matrix. Then, to evaluate the quality of the matching, we compute the MSE:

$$\text{MSE}(M) = \frac{1}{n}\|\mathbf{z} - M\mathbf{z}\|_2^2. \tag{25}$$

FOSCTTM For the CITE-seq dataset, we can use the Fraction Of Samples Closer Than the True Match (FOSCTTM) (Demetci et al., 2022; Liu et al., 2019) as an alternative matching metric. First,

---

[3] A variant of scGLUE was the top entrant in the NeurIPS 2021 Multi-modal single-cell data integration competition, from which we source our data. However, ground truth matchings were made available in the training set in that case, which fundamentally changes the nature and difficulty of the problem.

we distribute the mass of $\mathbf{x}^{(2)} = \{x_j\}_{j=1}^n$ to $\mathbf{x}^{(1)} = \{x_i\}_{i=1}^n$ as $\hat{\mathbf{x}}^{(1)} = M\mathbf{x}^{(2)}$, resulting in a projection of the first modality $\mathbf{x}^{(1)}$ in the space of the second modality. Then, we can compute a cross-modality distance as follows. For each point in $\hat{\mathbf{x}}^{(1)}$, we compute the Euclidean distance to each point in $\mathbf{x}^{(1)}$, and compute the fraction of samples in $\mathbf{x}^{(1)}$ that are closer than the true match. We also repeat this for each point in $\mathbf{x}^{(1)}$, computing the fraction of samples in $\hat{\mathbf{x}}^{(1)}$ in this case. That is, assuming again that the given indices correspond to the true matching, we compute:

$$\text{FOSCTTM}(M) =$$

$$\frac{1}{2n}\Bigg[\sum_{i=1}^n \bigg(\frac{1}{n}\sum_{j\neq i}\mathbb{1}\{d(\hat{\mathbf{x}}_i^{(1)}, \mathbf{x}_j^{(1)}) < d(\hat{\mathbf{x}}_i^{(1)}, \mathbf{x}_i^{(1)})\}\bigg) \tag{26}$$

$$+ \sum_{j=1}^n \bigg(\frac{1}{n}\sum_{i\neq j}\mathbb{1}\{d(\mathbf{x}_j^{(1)}, \hat{\mathbf{x}}_i^{(1)}) < d(\mathbf{x}_j^{(1)}, \hat{\mathbf{x}}_j^{(1)})\}\bigg)\Bigg], \tag{27}$$

where notice that this is a function of M through the computation $\hat{\mathbf{x}}^{(1)} = M\mathbf{x}^{(2)}$. As a baseline, we should expect a random matching, when distances between points are randomly distributed, to have an FOSCTTM of 0.5.

F.5 Models

In this section we describe experimental details pertaining to the propensity score and VAE (Yang et al., 2021). SCOT (Demetci et al., 2022) and scGLUE (Cao & Gao, 2022) are used according to tutorials and recommended default settings by the authors.

Loss Functions The propensity score approach minimizes the standard cross-entropy loss for both modalities, as implemented in PyTorch 2.0.1. The VAE includes, in addition to the standard ELBO loss (with parameter λ on the KL term), two cross-entropy losses based on classifiers from the latent space: one, weighted by a parameter α to classify t as in the propensity score, and another, weighted by a parameter β, that classifies which modality the latent point belongs to.

Hyperparameters and Optimization We use the Adam optimizer with learning rate 0.0001 and one cycle learning rate scheduler. We follow Yang et al. (2021) and set $\alpha = 1$, $\beta = 0.1$, but found that $\lambda = 10^{-9}$ (compared to $\lambda = 10^{-7}$ in Yang et al. (2021)) resulted in better performance. We used batch size 256 in both instances and trained for either 100 epochs (image) or 250 epochs (CITE-seq).

Architecture For the synthetic image dataset, we use an 5-layer convolutional network (channels $= 32, 54, 128, 256, 512$) with batch normalization and leaky ReLU activations, with linear heads for classification (propensity score and VAE) and posterior mean and variance estimation (VAE). For the VAE, the decoder consists of convolutional transpose layers that reverse those of the encoder. For the CITE-seq dataset, we use a 5-layer MLP with constant hidden dimension 1024, with batch normalization and ReLU activations (adapted from the fully connected VAE in Yang et al. (2021)) as both the encoder and VAE decoder. We use the same architecture for both modalities, RNA-seq (as we process the top 200 PCs) and protein.

Optimal Transport We used POT (Flamary et al., 2021) to solve the entropic OT problem, using the log-sinkhorn solver, with regularization strength $\gamma = 0.05$.

F.6 Data

Synthetic Data We follow the data generating process Equation (1) to generate coloured scenes of two simple objects (circles, or squares) in various orientations and with various backgrounds. The position of the objects are encoded in the latent variable z, which is perturbed by a do-intervention (setting to a fixed value) randomly sampled for each t. Each object has an x and y coordinate, leading to a 4-dimensional z, for which we consider 3 separate interventions each, leading to 12 different settings. The modality then corresponds to whether the objects are circular or square, and a fixed transformation of z, while the modality-specific noise U controls background distortions. Scenes are generated using a rendering engine from PyGame as $f^{(e)}$. Example images are given in Figure 4.

Figure 4: Example pair of synthetic images with the same underlying z.

CITE-seq Data

We also use the CITE-seq dataset from Lance et al. (2022) as a real-world benchmark (obtained from GEO accession GSE194122). These consist of paired RNA-seq and surface level protein measurements, and their cell type annotations over 45 different cell types. We used scanpy, a standard bioinformatics package, to perform PCA dimension reduction on RNA-seq by taking the first 200 principal components. The protein measurements (134-dimensional) was processed in raw form. For more details, see Lance et al. (2022).