# Auto-encoders with fibered latent spaces: A geometric approach to batch correction

**Tariq Daouda** [1 2 3 *] **Reda Chhaibi** [4 *] **Prudencio Tossou** [5 6] **Alexandra-Chloe Villani** [1 2 3]

## Abstract

This work introduces a geometric framework and a novel network architecture for creating correspondences between samples of different conditions. Under this formalism, the latent space is a fiber bundle stratified into a *base space* encoding conditions, and a *fiber space* encoding the variations within conditions. The correspondences between conditions are obtained by minimizing an energy functional, resulting in diffeomorphism flows between fibers. We illustrate this approach using MNIST and apply it to the batch-correction of single cell RNA sequencing datasets.

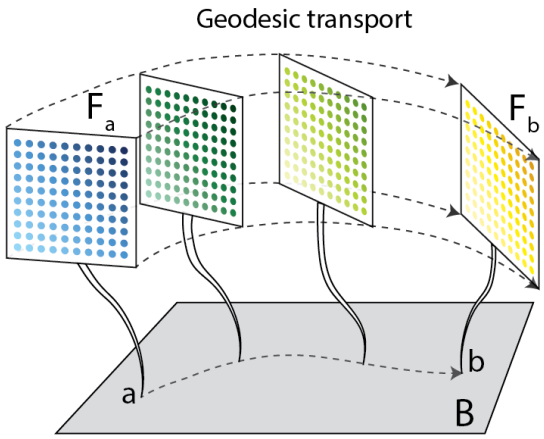


*Figure 1.1.* Conceptual sketch of the proposed method. We stratify the learned latent space $M$ into a base space $B$ and a fiber space $F$. Under this representation, samples can be formally transported from $F_a$ to $F_b$, and geodesic interpolations between latent spaces $F_a$ and $F_b$ can be generated.

## 1. Introduction

Single cell RNA-seq (scRNA-seq) sequencing measures the gene expression of each cell individually. The result is a matrix where each cell is represented by a vector of gene expressions. However, differences in sample handling and technical platforms leave strong imprints that uniquely mark each batch and overshadow the biology. These imprints are commonly referred to as *batch effects*. The process of *batch correction* refers to the integration of batches together while preserving relevant biological signal, such as cell types that can be inferred through the expression of distinct gene modules. Several methods have been recently proposed for correcting batch effects (Tran et al., 2020), and some of the most effective methods use deep neural networks (Lotfollahi et al., 2019; Lopez et al., 2018). In the context of batch correction, datasets are naturally stratified into two spaces, the batch (or condition) space and the gene expression space. Theoretically, these two spaces are considered independent, as the biological signal should be independent from technical artifacts. It naturally follows that an effective latent representation for a batch correcting model should enforce the disentanglement of the batch and biological signal.

However, most neural networks methods proposed for batch correction fail to consider or neglect this disentanglement (Lotfollahi et al., 2019; Lopez et al., 2018; Amodio et al., 2019). For example, the distribution of a variable $x$ stratified by a condition $c$ is often modelled using latent variables $z$ as $p(x|z, c)$ without any considerations for the dependencies between $z$ and $c$. Unlike $\beta$-VAE frameworks (Kimmel, 2020; Higgins et al., 2016), our disentanglement is geometrically built in the formalism and reflects supervised labeling.

Indeed, in this work, we propose a geometric formalism that explicitly stratifies the latent space $M$. Here, $M$ is taken as a Riemannian fiber bundle $M \subset \mathbb{R}^{\dim M}$, stratified into a base space $B$, encoding batch (or conditions), and a fiber space $F$, encoding variations within conditions (biological signal). Within this formalism samples are encoded using two coordinates $(f, b) \in B \times F$ (see Fig. 1.1).

[*]Equal contribution [1]Massachusetts General Hospital, Boston, MA [2]Harvard Medical School, Boston, MA [3]Broad Institute, Cambridge, MA [4]Institut de Mathématiques de Toulouse (IMT), France [5]InvivoAI, Montreal, QC [6]Université Laval, Québec, QC. Correspondence to: Reda Chhaibi <reda.chhaibi@math.univ-toulouse.fr>, Tariq Daouda <tariq.daouda@broadinstitute.org>.

*Proceedings of the $37^{th}$ International Conference on Machine Learning*, Vienna, Austria, PMLR 108, 2020. 

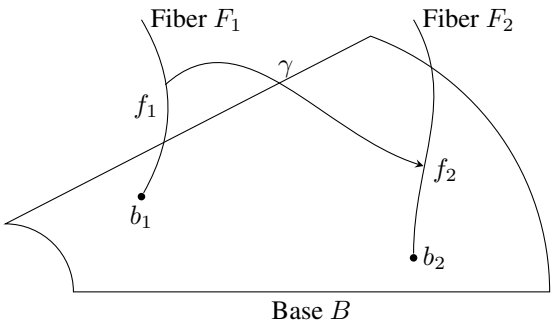


*Figure 2.1.* Illustration of geodesic transport with curve $\gamma$ from the fiber $F_1$ to $F_2$.

More importantly, this formalism allows us to compute correspondences between conditions using a natural Riemannian structure on $M$. This allow us to reframe the batch correction problem to the problem of translating samples between two conditions. We achieve this by finding the geodesics (shortest paths) in $M$ linking their corresponding fibers . We achieve this by minimizing an energy functional, which allow us to calculate diffeomorphism flows between fibers.

## 2. Fibered Auto-Encoders

Our goal is to construct the shortest curves between fibers in $M$. This is illustrated in Fig. 2.1.We assume the existence of a generating process

$$\Psi_\theta : M \to (\mathcal{X}, \|\cdot\|_2) \ . \tag{2.1}$$

A Fibered Auto-encoder (FAE) is essentially a standard auto-encoder whose latent space is stratified into base space $B$ and a fiber space $F$. However, to enforce the disentanglement between $B$ and $F$ and to improve the quality of generated samples, we have added auxiliary objectives to the reconstruction loss: (i) to enforce the disentanglement between base and fiber spaces, (ii) to improve the quality of the reconstructions. The entire architecture can be seen in Fig. 2.2, the theoretical treatment as well as training and implementation are detailed in the full paper (Daouda et al., 2020).

## 3. Results

### 3.1. MNIST

Throughout this section we contrast naive transport between two fibers (using the identity map to transport from $F_a$ to $F_b$), with geodesic transport (finding the shortest path on $M$ linking $F_a$ to $F_b$). We first illustrate our method on the MNIST dataset. Here, the possible conditions are the digits $\{0, 1, \ldots, 9\}$, the base is $B = \mathbb{R}^2$ and the standard fiber is $F = [-1, 1]^2$. Fig. 3.1 shows manifold plots obtained from an evenly spaced grid on $F_4$ and $F_9$. We can see that the reconstructions are of high quality and show a high diversity of samples despite the small bottleneck size (2 units). We also see that the learned latent space is contiguous as any coordinate $f \in [-1, 1]^2$ yields a realistic digit. Finally, Fig. 3.1 shows that the learned space has an intrinsic organisation, as gradually moving on fibers gradually changes digit features. Manifolds for $F_4$ and $F_9$ exhibit similar structures, with points at the same coordinate having similar inclination and boldness. This shows that naive transport is capable of creating rather accurate correspondences between fibers of similar conditions. We do not give the manifold plot after geodesic transport as the difference with naive transport are barely perceivable. Fig. 3.3 gives an estimation of the diffeomorphism between fibers $F_4$ and $F_9$. The starting points in $F_4$ are the orange dots obtained using an evenly spaced grid. The blue dots are the endpoints in $F_9$ for the calculated geodesics. The diffeomorphism induced through geodesic transport is globally close to the identity. This shows that naive transport can give a good approximation. However, the correction applied by geodesic transport is more apparent as we get closer to the edges.

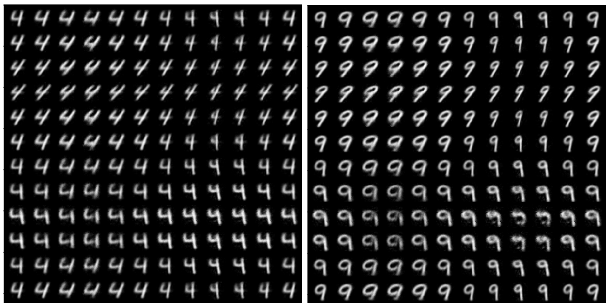

*Figure 3.1.* Manifold plots for the fibers $F_4$ and $F_9$ in MNIST. Images were generated using an evenly spaced grid in the standard fiber space $F = [-1, 1]^2$.

### 3.2. Batch-correcting scRNA-seq datasets

We use FAEs to represent batches as separate conditions, we then correct the batch effect by transporting all cells to a single reference batch. We benchmark our methods against the current state of the art in batch correction: Harmony (Korsunsky et al., 2019), and two neural networks developed to handle and batch correct scRNA-seq data: scGen (Lotfollahi et al., 2019) and SAUCIE (Amodio et al., 2019). As Harmony runs on principal components, we ran our benchmarks on PCA reduced data by using the first 20 principal components, in line with the methodology of (Tran et al., 2020).

We quantify batch correction quality using the prediction accuracies of three classifiers and the batch correction met-

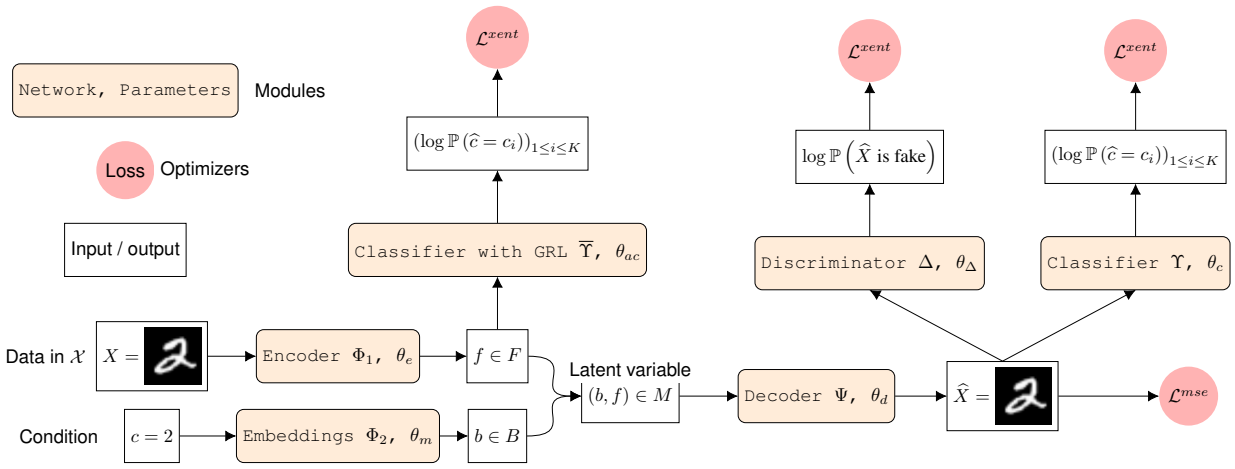


*Figure 2.2.* Network architecture. The general architecture is that of an auto-encoder receiving couples of samples and conditions $(X, c)$ and outputting a reconstruction $\widehat{X}$. The latent space is stratified into the fiber coordinate $f$ (output of the bottleneck layer), and the base coordinate $b$ encoding conditions. To the auto-encoder architecture we have added the classifier $\overline{\Upsilon}$ coupled with a GRL (Gradient Reversal Layer (Ganin & Lempitsky, 2014)) to disentangle $f$ from $b$, the GAN discriminator $\Delta$ to ensure reconstruction realism, and the condition classifier $\Upsilon$ to prevent mode collapses.

ric LISI (Korsunsky et al., 2019). Finally, we use Ward's variance decomposition to quantify how much changes in variance can be attributed to variations within groups, as opposed to in-between groups (Daouda et al., 2020) (Saporta, 2006)[p.258]. We benchmark all methods on two datasets. The first contains two batches of Peripheral Blood Mononuclear Cells (i.e., PBMCs): unstimulated and stimulated (with INF-b[1]) (Kang et al., 2018). The second is a compilation of 4 published pancreatic datasets generated by different groups using 4 distinct single-cell RNA sequencing experimental approaches.

A successful correction removes batch imprint and conserves cells biological identity. Therefore we report in Fig. 3.2 the accuracy on predicting the batch (i.e., lower is better) and the accuracy on predicting the cell type (i.e., higher is better). All methods performed well when it comes to removing the batch signal. Despite using a bottle-neck about 10 times lower [2], our transport method shows results close to to scGen (Lotfollahi et al., 2019). SAUCIE over-corrected the batch effect at the expense cell type identity. Compared to naive, geodesic transport increases both cell type and batch predictability. For the LISI scores, once again, naive and geodesic transport show results on par with scGen.

Ward's variance decomposition in Fig. 3.2, shows that transport is the only method capable of retaining a significant part of the original variance of the uncorrected dataset. This suggests that our method retains more of the biological signal. On the PBMC datasets, we observe increases in variance within batches and between cell types. Such increases could be due to the network imputing values for missing genes: due to single-cell experimental technical issues, gene expression matrices are very sparse (over 80% to 90%).

Fig. 3.4 3.5 shows UMAP plots (McInnes et al., 2018) for uncorrected data, transport and scGen. After correction, batches overlap and cells cluster by cell type. In contrast to scGen, transport kept cell types together, CD16 monocytes (purple) and CD16 monocytes (red), close to each other as in the original uncorrected dataset (see fig. 3.4). This suggest that FAEs are better at conserving relevant structures from datasets, and echoes variance decomposition results (Fig. 3.2). Finally, Fig. 3.5 shows that FAEs are able to integrate batches containing small numbers of cells.

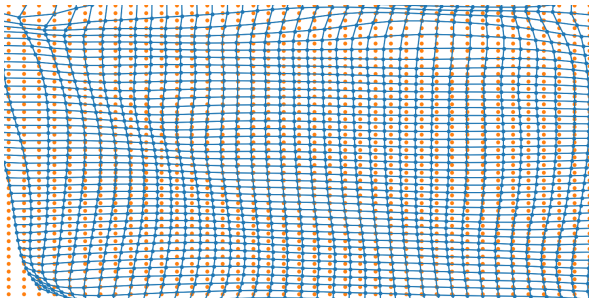

*Figure 3.3.* Diffeomorphism between $F_4$ and $F_9$. The orange dots represent the original coordinates in $F_4$, the blue dots are their corresponding images in $F_9$ computed through geodesic transport.

[1] Interferon-beta

[2] We used 10 for pancreas and 16 for PBMC vs 100 for scGen.

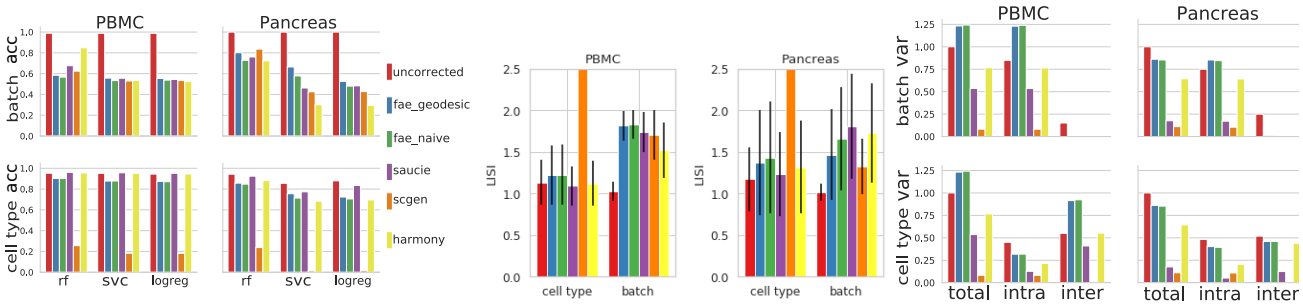


*Figure 3.2.* Left: Accuracies for uncorrected and batch correction methods on both dataset. Batch accuracy (lower is better) and, cell type accuracy (higher is better) are reported for random forest (rf), support vector classifier (svc) and logistic regression (logreg). Middle: LISI scores for uncorrected and batch correction methods on both datasets. LISI on cell type (closer to 1 is better), LISI on batch (higher is better). Error bars display the standard deviation. Right: Total variance and Ward's variance decomposition, for uncorrected data and batch correction methods. Total variance has been normalized to 1, on the uncorrected dataset.

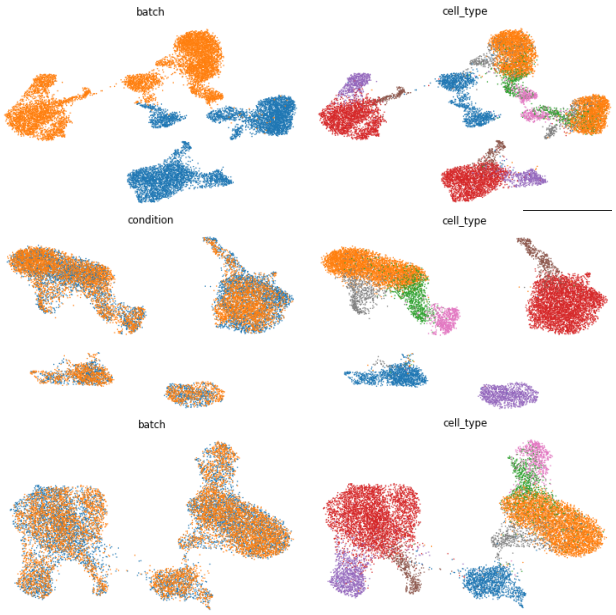


*Figure 3.4.* UMAP visualization of PBMC cells. Left column: cells colored by batch, Right: colored by cell types. From top to bottom: uncorrected data, scGen, geodesic transport (the plot of naive transport is very close to the naked eye). Transport conserves cell types relationships by keeping purple cells (i.e., CD16 monocytes) close to red cells (i.e., CD14 monocytes), which are related to each other.

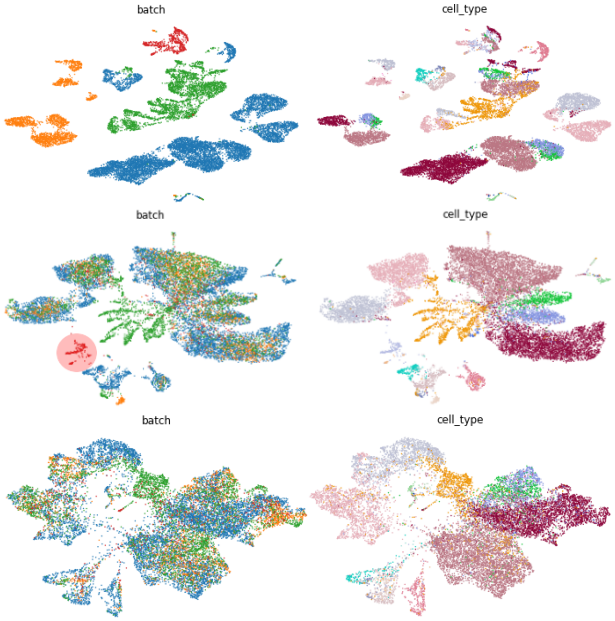


*Figure 3.5.* UMAP visualization of pancreas cells. Left column: cells colored by batch, right: colored by cell types. From top to bottom: uncorrected data, scGen (bottleneck size: 100), naive transport (bottleneck size: 16). Contrary to scGen, naive transport was able to integrate cells form the red batch despite the small sample size. This suggests that FAE are better at integrating datasets of small sample sizes

## 4. Conclusion

We proposed FAEs, a general learning framework capable of independently representing batch and biological signal and used FAEs to successfully correct batch effects in scRNA-seq. Here, We restricted ourselves to cases where conditions are of similar nature (e.g. datasets containing the same cell types). Consequently, naive and geodesic transport yield similar results. One could argue that this proximity morally measures similarity between conditions. Future work could explore applications to more dissimilar conditions as well as different layer types and loss functions. Finally, geodesics computation becomes more challenging as latent space size increases. This technical issue is mitigated by the fact that FAEs can learn accurate representations in very small latent spaces. Addressing it would be a natural follow-up.

## References


Amodio, M., Van Dijk, D., Srinivasan, K., Chen, W. S., Mohsen, H., Moon, K. R., Campbell, A., Zhao, Y., Wang, X., Venkataswamy, M., et al. Exploring single-cell data with deep multitasking neural networks. *Nature methods*, pp. 1–7, 2019.

Daouda, T., Chhaibi, R., Tossou, P., and Villani, A.-C. Geodesics in fibered latent spaces: A geometric approach to learning correspondences between conditions. *arXiv preprint arXiv:2005.07852*, 2020.

Ganin, Y. and Lempitsky, V. Unsupervised domain adaptation by backpropagation. *arXiv preprint arXiv:1409.7495*, 2014.

Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., and Lerchner, A. beta-vae: Learning basic visual concepts with a constrained variational framework. 2016.

Kang, H. M., Subramaniam, M., Targ, S., Nguyen, M., Maliskova, L., McCarthy, E., Wan, E., Wong, S., Byrnes, L., Lanata, C. M., et al. Multiplexed droplet single-cell rna-sequencing using natural genetic variation. *Nature biotechnology*, 36(1):89, 2018.

Kimmel, J. C. Disentangling latent representations of single cell rna-seq experiments. *bioRxiv*, 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*, pp. 1–8, 2019.

Lopez, R., Regier, J., Cole, M. B., Jordan, M. I., and Yosef, N. Deep generative modeling for single-cell transcriptomics. *Nature methods*, 15(12):1053–1058, 2018.

Lotfollahi, M., Wolf, F. A., and Theis, F. J. scgen predicts single-cell perturbation responses. *Nature methods*, 16 (8):715–721, 2019.

McInnes, L., Healy, J., and Melville, J. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

Saporta, G. *Probabilités, analyse des données et statistique*. Editions Technip, 2006.

Tran, H. T. N., Ang, K. S., Chevrier, M., Zhang, X., Lee, N. Y. S., Goh, M., and Chen, J. A benchmark of batch-effect correction methods for single-cell rna sequencing data. *Genome Biology*, 21(1):1–32, 2020.