Multigrate: single-cell multi-omic data integration

Mohammad Lotfollahi [*1 2] **Anastasia Litinetskaya** [*1 3] **Fabian J. Theis** [1 2 3]

Abstract

Single-cell multimodal omics technologies provide a holistic approach to study cellular decision making. Yet, learning from multimodal data is complicated because of missing and incomplete reference samples, non-overlapping features and batch effects between datasets. To integrate and provide a unified view of multi-modal datasets, we propose *Multigrate*. Multigrate is a generative multi-view neural network to build multimodal reference atlases. In contrast to existing methods, Multigrate is not limited to specific paired assays, and it compares favorably to existing data-specific methods on both integration and imputation tasks. We further show that Multigrate equipped with transfer learning enables mapping a query multi-modal dataset into an existing reference atlas.

1. Introduction

Recent advances in single-cell technologies allow us to quickly and efficiently measure several features of cells at the same time. For instance, CITE-seq (Stoeckius et al., 2017) measures gene expression levels and surface protein counts, and ATAC-seq (Buenrostro et al., 2015) measures transcriptome and chromatin openness in one cell. While RNA-seq data integration has become a well-studied problem, similar methods for multi-omics are still pending.

Several approaches have tackled the integration of paired single-cell multi-omic measurements such as CITE-seq or/and ATAC-RNA (Gayoso et al., 2021; Argelaguet et al., 2020; Hao et al., 2020). However, existing methods are limited to a specific paired technology or they use simple linear models and lack imputation mechanisms for missing modalities. Additionally, none of the existing methods allow mapping novel multi-omic query datasets (Lotfollahi et al., 2020) to reference atlases constructed from multiple multi-omic technologies. Finally, none of these models can robustly integrate datasets with non-matching measurements (Lopez et al., 2019; Lotfollahi et al., 2019).

Here, we present *Multigrate*, an unsupervised deep generative model to integrate multi-omic datasets and address these challenges. Multigrate learns a joint latent space combining information from multiple modalities from paired and unpaired measurements while accounting for technical biases within each modality. Combined with transfer learning, Multigrate can map novel multi-omic query datasets to a reference atlas and impute missing modalities. We first compare our model with state-of-the-art approaches on integration and imputation tasks and later demonstrate the multi-modal reference mapping feature of Multigrate.

2. Methods

We first define the observed data as $X = \{X^i\}_{i=1,\dots,n}$ for modalities $1, \dots, n$, where X^i denotes observations for modality i and can be empty in case of a missing modality. Let $S = \{S^i\}_{i=1,\dots,n}$ be the set of study labels (i.e. samples, experiments across labs or sequencing technologies), and let Z^i denote the conditional modality representation. We employ the Product of Experts (PoE) framework (Lee & van der Schaar, 2021) to calculate the joint distribution for data that comes from several modalities, also when some of the modalities are partially missing. Let $\phi = \{\phi_i\}_{i=1,\dots,n}$ be parameters of the posterior distributions q and let $\theta = \{\theta_i\}_{i=1,\dots,n}$ be parameters of the data distribution p. We denote the joint latent representation by Z^{joint} and the joint posterior by $q_\phi(Z^{joint}|X, S)$. We model the joint posterior as the product of the conditional marginal posteriors:

$$q_\phi(Z^{joint}|X, S) = \Pi_{i=1}^{n} q_{\phi_i}(Z^i|X^i, S^i), \qquad (1)$$

setting $q_{\phi_i}(Z^i|X^i, S^i)$ to 1 if modality i is missing.

Furthermore, modality encoder f_i outputs parameters of q_{ϕ_i} and modality decoder g_i outputs parameters of p_{θ_i}. We assume that $q_{\phi_i}(Z^i|X^i, S^i) = \mathcal{N}(Z^i|\mu_i, \sigma_i)$, where μ_i, σ_i are the output of the modality encoder $f_i(X^i, S^i)$. The

[*]Equal contribution. Both co-first authors have the right to list their name first in their CV. [1]Institute of Computational Biology, Helmholtz Center Munich, Neuherberg, Germany [2]School of Life Sciences Weihenstephan, Technical University of Munich, Munich, Germany [3]Department of Mathematics, Technical University of Munich, Munich, Germany. Correspondence to: Mohammad Lotfollahi <mohammad.lotfollahi@helmholtz-muenchen.de>.

The 2021 ICML Workshop on Computational Biology.

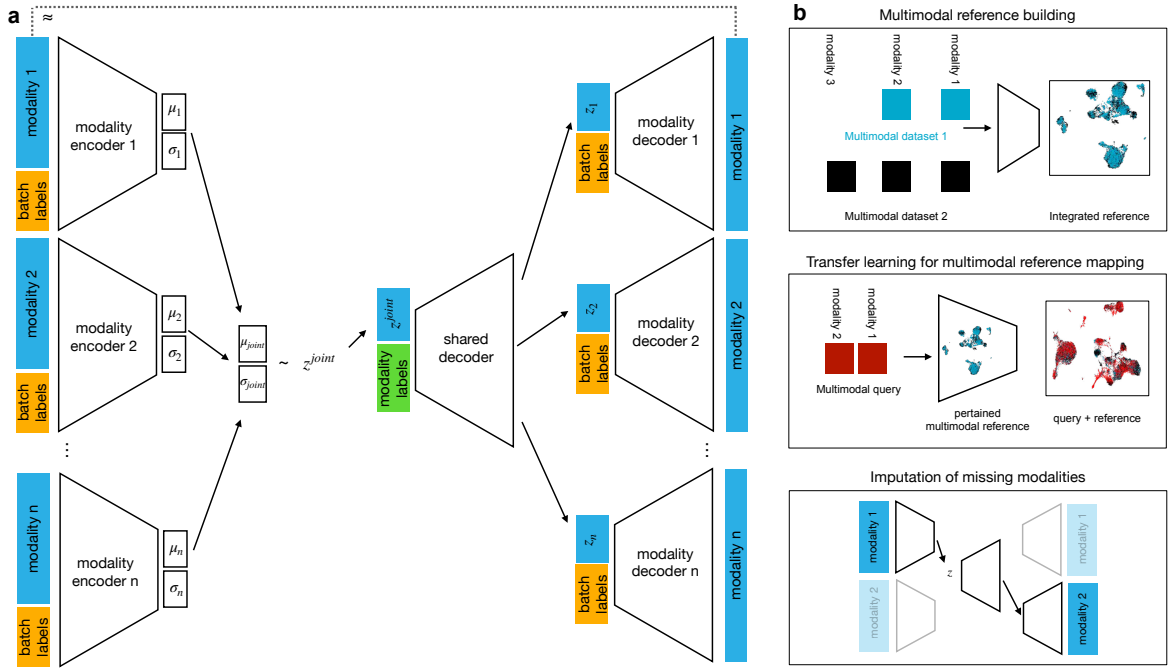

Figure 1. **(a-b)** Multigrate architecture and applications.

parameters of the joint distribution are calculated as

$$\mu_{joint} = (\mu_0\sigma_0^{-1} + \sum_{i=1}^{n} m_i\mu_i\sigma_i^{-1})(\sigma_0^{-1} + \sum_{i=1}^{n} m_i\sigma_i^{-1})^{-1},$$
$$\sigma_{joint} = (\sigma_0^{-1} + \sum_{i=1}^{n} m_i\sigma_i^{-1})^{-1}, \tag{2}$$

where μ_0 and σ_0 are the parameters of the prior $\mathcal{N}(\mu_0, \sigma_0)$, which in our case is standard normal, and m_i is 1 if modality i is present and is 0 otherwise.

We formulate the objective function for a specific dataset as

$$\begin{aligned}\mathcal{L}_{\text{AE}}(\phi, \theta, X_i, S_i, \alpha, \eta) = \\ \alpha\mathbb{E}_{q_\phi(Z^{joint}|X_i,S_i)}[\log p_\theta(X_i|Z^{joint}, S_i)] \\ - \eta\mathbb{KL}(q_\phi(Z^{joint}|X_i, S_i)||p_\theta(Z^{joint}|S_i)),\end{aligned} \tag{3}$$

where α and η are hyper-parameters. Finally, to ensure that different datasets are integrated well, we utilize the maximum mean discrepancy (MMD) loss. It allows to minimize the distance between two distributions and was previously shown to improve the performance of VAE-based models (Lotfollahi et al., 2019). We calculate the MMD loss between the joint representations for pairs of datasets. In the implementation, we use multi-scale radial basis kernels defined as $k(x, x') = \sum_{i=1}^{l} k(x, x', \gamma_i)$, where $k(x, x', \gamma_i) = \exp(\gamma_i||x - x'||^2)$ is a Gaussian kernel, x, x' are observations from two different distributions and $l, \gamma_1, \ldots, \gamma_l$ are hyper-parameters. Given d datasets $X_1, \ldots, X_d$ with study labels $S_1, \ldots, S_d$, the final loss function is defined as

$$\begin{aligned}\mathcal{L}_{\text{multigrate}} = \sum_{i=1}^{d} \mathcal{L}_{\text{AE}}(\phi, \theta, X_i, S_i, \alpha, \eta) \\ + \beta \sum_{\substack{i,j=0 \\ i<j}}^{d} \mathcal{L}_{\text{MMD}}(Z_i^{joint}, Z_j^{joint})\end{aligned} \tag{4}$$

where α, β, η are hyper-parameters.

The decoder part of the network consists of two parts (Figure 1a): z^{joint} is first fed into the shared decoder g that reintroduces modality variation to the joint to obtain modality-specific representations. Then modality decoders g_i take the modality-specific representations as input and output the parameters of p_{θ_i}. By default, a negative binomial loss is used for the RNA modality, in which case the distribution mean is output by the modality decoder, and the discrepancy parameter is learned per batch. For normalized protein counts and normalized binary chromatic peaks, we use the mean squared error loss. In this case, the output of the modality decoder can be seen as reconstructed data, hence we refer to this part of the loss function as reconstruction loss. The overall reconstruction loss is the sum over all modalities.

We also implement the single-cell architectural surgery approach introduced in (Lotfollahi et al., 2020) to allow the building of reference atlases and mapping of new query data into the reference atlas. When a new query data needs to be added to an existing reference atlas built with Multigrate,

we introduce a new set of batch labels S_{d+1} and fine-tune conditional weights in modality decoders f_i and modality encoders g_i.

To find optimal hyper-parameters, i.e. α, β and η, we used the grid search over the parameter space. To quantitatively access the performance of different methods, we use some of the metrics proposed in (Luecken et al., 2020). Adjusted rand index (ARI), normalized mutual information (NMI), average silhouette width (ASW) cell type and isolated label silhouette are biological conservation metrics that measure how much of biological variance was preserved after the integration. Graph connectivity and ASW batch are batch correction metrics (in cases where there are several batches present in the data) to assess how well batch effects were removed after integration. The final score was calculated as 0.4*batch correction + 0.6*bio conservation.

We tested Multigrate on several peripheral blood mononuclear cell (PBMC) datasets: Dataset 1 is a paired RNA-seq/ATAC-seq dataset (10x); Datasets 2, 3, and 4 (Hao et al., 2020; Kotliarov et al., 2020; Stephenson et al., 2021) are CITE-seq datasets. All datasets were quality controlled and preprocessed following the same pipeline. RNA-seq datasets were normalized to sum up to $10,000$ and $\log(x+1)$ transformed. Peaks in Dataset 1 were binarized and log-normalized as above. Protein counts were normalized using the centered log-ratio transformation (Stoeckius et al., 2017).

Figure 1a depicts the complete architecture of Multigrate and Figure 1b lists possible applications of our method such as multi-modal reference building, query to reference mapping, and imputations of missing modalities which we investigate in the following.

3. Results

3.1. Benchmarking multi-modal integration quality

We applied Multigrate on three datasets (1-3) to assess its ability to integrate a multi-modal single-cell dataset using paired single-cell measurements ranged from CITE-seq to joint ATAC-RNA. We compared our model against three other methods: MOFA+ (Argelaguet et al., 2020), Seurat v4 (Hao et al., 2020) and totalVI (Gayoso et al., 2021).

In the benchmarks experiments, totalVI was run with default parameters. In Seurat, we first calculated the weighted nearest neighbor (WNN) graph, and then to obtain embeddings in a latent space, we ran supervised PCA with default parameters. When multiple batches were present in the data, we first performed the integration for each modality separately using Seurat v4 and Signac (Stuart et al., 2020) and then ran the WNN analysis.

Figure 2a,b shows the UMAP of the integrated Dataset 3 and the individual metric scores for the same dataset. Multigrate performed well in both the batch correction metrics and the bio-conservation metrics and overall performed slightly better than all the other three methods. Figure 2c depicts the overall score for all three datasets demonstrating the robust performance of our method against existing approaches. Since totalVI is a method for CITE-seq integration, we benchmarked it only on the two CITE-seq datasets. Dataset 1 does not contain batches, therefore only the batch-independent metrics are reported. We observed that Multigrate compares favorably to the existing methods.

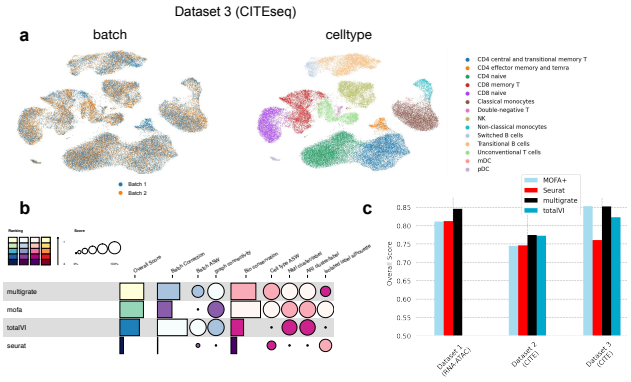

Figure 2. **(a)** UMAP embedding for the latent space of Multigrate for Dataset 3. **(b)** Integration quality metrics for Dataset 3. **(c)** Benchmarking against existing methods across three datasets.

3.2. Multi-modal reference building and mapping query

To demonstrate Multigrate's functionality to build reference atlases and map new query data, we first built a healthy blood cell atlas using healthy cells from Datasets 1, 2 and 4. In total, the reference atlas comprised around $160,000$ cells. The reference atlas incorporated measurements from three modalities: gene expressions, protein counts, and chromatin openness. Next, we mapped a new query dataset consisting of $50,000$ sampled diseased COVID-19 cells from Dataset 4 into the reference atlas by fine-tuning the new conditional weights using scArches (Lotfollahi et al., 2020). Figure 3a-c shows the UMAPs of the integrated reference and query data together across studies, cell types and conditions. We observe that the query was well integrated into the reference.

To transfer cell-type annotations from the reference data to the query, we trained a random forest classifier on the reference data and predicted cell types for the query data. The classifier achieves an overall accuracy of 79% over all cell types. Figure 3d shows a heatmap of the confusion matrix between the true and the predicted cell types. The cell types that were not correctly classified, e.g. ASDC or Treg, were present in the reference as very small populations

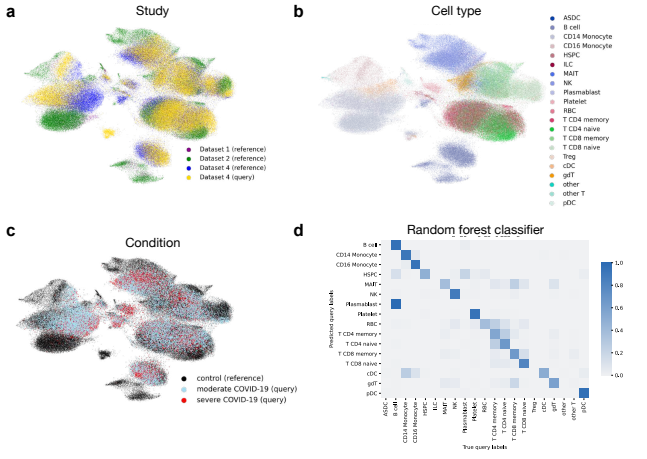

Figure 3. **(a-c)** UMAPs of the integrated reference and query across studies, cell types and conditions. **(d)** Heatmap of a confusion matrix between true cell types and predicted cell types for the query.

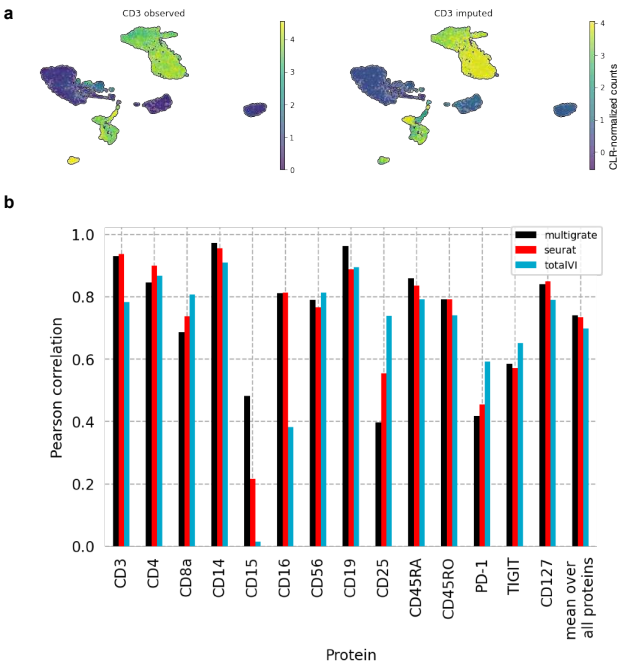

Figure 4. **(a)** UMAPs of observed expressions of CD3 (left) and imputed expressions by Multigrate (right). **(b)** Bar plot of Pearson's correlation coefficients across all proteins comparing Multigrate, Seurat and totalVI.

comprising less than 100 cells each. Overall, we observed that Multigrate can successfully build multi-modal reference atlases, update the atlas with new query datasets and transfer information from reference to query.

3.3. Imputation of missing modalities

Multigrate can also integrate unpaired datasets, for instance, CITE-seq data and RNA-seq data, and impute missing modalities as protein abundance in this case. To illustrate this functionality, we leveraged a PBMCs CITE-seq dataset (Gayoso et al., 2021), consisting of $15,000$ cell with both transcriptomic and 14 protein measurements. We first integrated $10,000$ paired observations and $5,000$ RNA-seq only observations, where we left out protein counts in the latter as ground truth. Then we imputed protein expression and calculated Pearsons's correlation coefficients between the predicted protein expressions and the ground truth. We compared the performance of Multigrate on this task to Seurat v4 and totalVI which were run with default parameters.

As an example, we observe the imputed CD3 protein agrees with the ground truth protein abundance (Figure 4a). Next, we evaluated the overall accuracy of the imputed measurements for individual proteins and overall average performance (see the last column of the barplot in Figure4b). These results demonstrate the generalization power and robustness of Multigrate in imputing missing proteins compared to state-of-the-art models as totalVI specifically designed for this task. On average, Multigrate slightly outperforms both of the other methods.

4. Conclusion

We introduced Multigrate, a scalable deep learning approach to learn a joint representation from multi-omic single-cell datasets. While Multigrate is generalizable to potentially any multi-omic technology, it compares favorably to existing integration approaches for specific paired measurements for both integrating and imputation tasks. Multigrate is also able to map multi-modal COVID-19 data onto a healthy reference atlas and transfer knowledge from reference to query.

We predict that the addition of regularization terms as cycle-consistency (Zhu et al., 2020) would improve imputation accuracy and unpaired data integration quality. Moreover, replacing one-hot modality labels with learnable embeddings (Lotfollahi et al., 2021) to induce modality effect will further help to decompose the explained variance for each modality and increase the interpretability of the model by comparing modality vectors.

With the increased availability of single-cell multi-omic datasets, we expect Multigrate to enable users to easily integrate and analyze these data, providing a holistic view of cells instead of looking through the lens of a single measurement with limited information.

The code to reproduce the results is available at https://bit.ly/3fOvupR.

References

Datasets - single cell multiome atac + gene exp. - official 10x genomics support. https://support.10xgenomics.com/single-cell-multiome-atac-gex/datasets/2.0.0/pbmc_granulocyte_sorted_10k.

Argelaguet, R., Arnol, D., Bredikhin, D., Deloro, Y., Velten, B., Marioni, J. C., and Stegle, O. Mofa+: a statistical framework for comprehensive integration of multi-modal single-cell data. *Genome Biology*, 21(1):111, May 2020. URL https://doi.org/10.1186/s13059-020-02015-1.

Buenrostro, J. D., Wu, B., Litzenburger, U. M., Ruff, D., Gonzales, M. L., Snyder, M. P., Chang, H. Y., and Greenleaf, W. J. Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature*, 523(7561):486–490, Jul 2015. URL https://doi.org/10.1038/nature14590.

Gayoso, A., Steier, Z., Lopez, R., Regier, J., Nazor, K. L., Streets, A., and Yosef, N. Joint probabilistic modeling of single-cell multi-omic data with totalvi. *Nature Methods*, 18(3):272–282, Mar 2021. URL https://doi.org/10.1038/s41592-020-01050-x.

Hao, Y., Hao, S., Andersen-Nissen, E., Mauck, W. M., Zheng, S., Butler, A., Lee, M. J., Wilk, A. J., Darby, C., Zagar, M., Hoffman, P., Stoeckius, M., Papalexi, E., Mimitou, E. P., Jain, J., Srivastava, A., Stuart, T., Fleming, L. B., Yeung, B., Rogers, A. J., McElrath, J. M., Blish, C. A., Gottardo, R., Smibert, P., and Satija, R. Integrated analysis of multimodal single-cell data. *bioRxiv*, 2020. URL https://www.biorxiv.org/content/early/2020/10/12/2020.10.12.335331.

Kotliarov, Y., Sparks, R., Martins, A. J., Mulè, M. P., Lu, Y., Goswami, M., Kardava, L., Banchereau, R., Pascual, V., Biancotto, A., Chen, J., Schwartzberg, P. L., Bansal, N., Liu, C. C., Cheung, F., Moir, S., and Tsang, J. S. Broad immune activation underlies shared set point signatures for vaccine responsiveness in healthy individuals and disease activity in patients with lupus. *Nature Medicine*, 26(4):618–629, Apr 2020. URL https://doi.org/10.1038/s41591-020-0769-8.

Lee, C. and van der Schaar, M. A variational information bottleneck approach to multi-omics data integration. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pp. 1513–1521, 13–15 Apr 2021. URL http://proceedings.mlr.press/v130/lee21a.html.

Lopez, R., Nazaret, A., Langevin, M., Samaran, J., Regier, J., Jordan, M. I., and Yosef, N. A joint model of unpaired data from scRNA-seq and spatial transcriptomics for imputing missing gene expression measurements. *arXiv:1905.02269 [cs, q-bio, stat]*, May 2019. URL http://arxiv.org/abs/1905.02269.

Lotfollahi, M., Naghipourfar, M., Theis, F. J., and Wolf, F. A. Conditional out-of-sample generation for unpaired data using trvae. *CoRR*, abs/1910.01791, 2019. URL http://arxiv.org/abs/1910.01791.

Lotfollahi, M., Naghipourfar, M., Luecken, M. D., Khajavi, M., Büttner, M., Avsec, Z., Misharin, A. V., and Theis, F. J. Query to reference single-cell integration with transfer learning. preprint, July 2020. URL http://biorxiv.org/lookup/doi/10.1101/2020.07.16.205997.

Lotfollahi, M., Susmelj, A. K., Donno, C. D., Ji, Y., Ibarra, I. L., Wolf, F. A., Yakubova, N., Theis, F. J., and Lopez-Paz, D. Compositional perturbation autoencoder for single-cell response modeling. *bioRxiv*, pp. 2021.04.14.439903, April 2021. URL https://www.biorxiv.org/content/10.1101/2021.04.14.439903v1.

Luecken, M., Büttner, M., Chaichoompu, K., Danese, A., Interlandi, M., Mueller, M., Strobl, D., Zappia, L., Dugas, M., Colomé-Tatché, M., and Theis, F. Benchmarking atlas-level data integration in single-cell genomics. *bioRxiv*, 2020. URL https://www.biorxiv.org/content/early/2020/05/27/2020.05.22.111161.

Stephenson et al. The cellular immune response to covid-19 deciphered by single cell multi-omics across three uk centres. *medRxiv*, 2021.

Stoeckius, M., Hafemeister, C., Stephenson, W., Houck-Loomis, B., Chattopadhyay, P. K., Swerdlow, H., Satija, R., and Smibert, P. Simultaneous epitope and transcriptome measurement in single cells. *Nature Methods*, 14 (9):865–868, Sep 2017. URL https://doi.org/10.1038/nmeth.4380.

Stuart, T., Srivastava, A., Lareau, C., and Satija, R. Multimodal single-cell chromatin analysis with signac. *bioRxiv*, 2020. URL https://www.biorxiv.org/content/early/2020/11/10/2020.11.09.373613.

Zhu, J.-Y., Park, T., Isola, P., and Efros, A. A. Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks. *arXiv:1703.10593 [cs]*, August 2020. URL http://arxiv.org/abs/1703.10593.