# MultiCPA: Multimodal Compositional Perturbation Autoencoder

**Kemal Inecik** [1 2] **Andreas Uhlmann** [3] **Mohammad Lotfollahi** [1 *] **Fabian Theis** [1 3 *]

## Abstract

Single-cell multimodal profiling provides a high-resolution view of cellular information. Recently, multimodal profiling approaches have been coupled with CRISPR technologies to perform pooled screens of single or combinatorial perturbations. This opens the possibility of exploring the massive space of combinatorial perturbations and their regulatory effects computationally from the extrapolation of a few experimentally feasible combinations. Here, we propose *MultiCPA*, an end-to-end generative architecture to predict multimodal perturbation response at single cell level. Two mixing strategies to integrate multiple modalities are introduced and compared with existing methods. MultiCPA was also shown to accurately predict unseen combinatorial perturbation responses for multiple modalities. The code to reproduce the results is available on *GitHub*, theislab/multicpa.

## 1. Introduction

Single-cell multiomics (Teichmann & Efremova, 2020) datasets are routinely generated to capture the cellular heterogeneity (Stephenson et al., 2021; Yao et al., 2021) with higher resolution through simultaneous quantification of transcriptome and surface proteins in the same cell (Stoeckius et al., 2017). Recently, multimodal technologies have been combined with CRISPR-compatible cellular indexing of transcriptomes and epitopes to profile cells under single (Mimitou et al., 2019; Frangieh et al., 2021) or combinatorial (Wessels et al., 2022) genetic perturbation. However, a comprehensive experimental investigation of combinatorial perturbations is challenging due to massive exploration space of possible combinations. Thus, computational methods are required to *in silico* predict cellular responses to a perturbation to navigate the large perturbation space and facilitate the experimental design.

Computational models based on representation learning (Lotfollahi et al., 2019; Ji et al., 2021) have been successfully applied to predict gene expression response to disease and chemical perturbation at the single-cell level. Recently, compositional perturbation autoencoder (CPA) (Lotfollahi et al., 2021) was proposed to predict gene expression response to combinatorial drug or genetic perturbations. Yet, CPA predictions are limited to a single modality, the gene expression. However, perturbation response prediction across multiple modalities helps to obtain a more holistic view of cellular behavior. Existing approaches such as Total Variational Inference (totalVI) (Gayoso et al., 2021) have been shown to efficiently model CITE-seq data by modeling biological and technical factors in the data. totalVI has been applied to perform counterfactual prediction to impute unmeasured surface proteins for single-cell RNA-seq data. However, the model is unable predict the combinatorial perturbation responses.

To address these challenges, we present multimodal compositional perturbation autoencoder, *MultiCPA*, an end-to-end generative model to exploit paired measurement of RNA and surface proteins to learn perturbation responses across both modalities at single-cell level. We demonstrate MultiCPA can efficiently model highly multiplexed multimodal CRISPR screens to predict unseen single and combinatorial perturbations. Furthermore, MultiCPA learns a probabilistic representation of the data while accounting for biological and technical factors.

## 2. Methods

### 2.1. Integrating multimodal perturbation profiles

To evaluate the relative successes of two different mixture models, Product-of-Expert (PoE) (Lee & van der Schaar, 2021) and concatenation, two variational autoencoders (VAEs) (Kingma & Welling, 2013) were built as shown in Figure 1. In concatenation based model architecture, MultiCPA (concat), joint feature vectors are constructed by concatenating observed data from both modalities, proteins ($x_P$) and genes ($x_G$). Joint embedding, $Z_{joint}$, is sam-

[*]Equal contribution [1]Institute of Computational Biology, Helmholtz Zentrum München, German Research Center for Environmental Health, 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: Fabian Theis, Mohammad Lotfollahi < fabian.theis@helmholtz-muenchen.de, mohammad.lotfollahi@helmholtz-muenchen.de>.

*The 2022 ICML Workshop on Computational Biology*. Baltimore, Maryland, USA, 2022. 

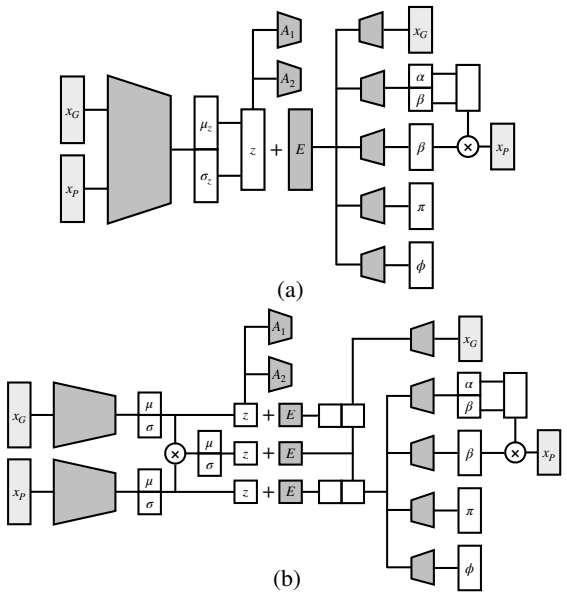



*Figure 1.* Overview of proposed MultiCPA architectures, where $A_i$ denotes adversarial discriminator networks, $E$ denotes separate perturbation and covariate embeddings. Multimodal integration by a) concatenation mixture module. b) PoE mixture module.

pled via the reparameterization trick (Kingma et al., 2015) from joint posterior, $q(Z_{joint}|x_G, x_P)$, which is estimated by a shared encoder from the concatenated feature vector. The PoE mixture model architecture, MultiCPA (PoE), estimates independent marginal latent distributions $q(Z_P|x_P)$ and $q(Z_G|x_G)$ by the respective encoder of each modality. Joint embedding, $Z_{joint}$, is estimated similarly from joint posterior $q(Z_{joint}|x_G, x_P)$, the product of the conditional marginal posteriors calculated using PoE framework. The loss function of the mixture module $\mathcal{L}_{\mathrm{M},1}$ for MultiCPA (concat) is given in Equation 1, while arithmetic mean of all KL divergences is used for MultiCPA (PoE).

$$\mathcal{L}_{\mathrm{M},1} = \mathrm{KL}(\mathcal{N}(\mu_{joint}, \sigma_{joint})||\mathcal{N}(0,1)). \tag{1}$$

The information about perturbations and covariates in the joint embedding is disentangled by using an adversarial network as implemented in CPA framework (Lotfollahi et al., 2021). Auxiliary cross entropy losses implemented for two adversarial discriminator networks, which are trained to predict perturbation and cell covariates from $Z_{joint}$, forces the encoders to produce the basal cellular state, $Z_{basal}$. The adversarial loss for MultiCPA (concat) is given in Equation 2, while the adversarial loss for MultiCPA (PoE) is defined by the arithmetic mean of $\mathcal{L}_{\mathrm{A}}(Z_P)$, $\mathcal{L}_{\mathrm{A}}(Z_G)$ and $\mathcal{L}_{\mathrm{A}}(Z_{joint})$.

$$\begin{aligned}\mathcal{L}_{\mathrm{A},1}(Z) &= \mathrm{CrossEntropy}(A_1(Z), \mathrm{perturbation})\\ \mathcal{L}_{\mathrm{A},2}(Z) &= \mathrm{CrossEntropy}(A_2(Z), \mathrm{covariate})\end{aligned} \tag{2}$$

$Z_{basal}$ is then composed with each of perturbation and covariate embeddings separately in the latent space and forwarded to the decoder networks. To reconstruct the observed data from the latent representation of the observations, the conditional latent space embeddings are entailed to learn perturbations and covariates. Extracting expressions (gene and protein), perturbations, and covariates as disentangled embeddings in the latent space allows to predict counterfactual scenarios such as unseen perturbation combinations.

### 2.2. Modality-specific data reconstruction

Joint latent space embedding feeds into the modality-specific decoders trying to reconstruct the corresponding input data. Inspired by totalVI concepts (Gayoso et al., 2021) integrating genes and protein modalities, the decoder network in MultiCPA architectures consists of five neural networks. All encoders, decoders, and adversarials of both MultiCPA models were built from fully-connected blocks. The gene data is decoded using a single decoder to reconstruct the observed data utilizing negative binomial loss function (Equation 3, indices are dropped for readability), where the distribution is specified by the mean $\mu_G$ and the inverse dispersion $\theta_G$.

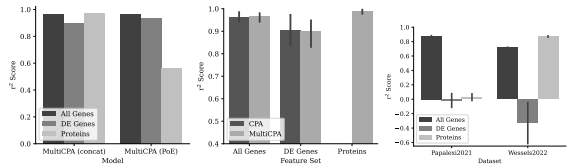
$$\begin{aligned}\mathcal{L}_{\mathrm{G}} &= \mathrm{NB}(x; \mu, \theta)\\ &= \frac{\Gamma(x+\theta)}{\Gamma(x+1)\Gamma(\theta)}\left(\frac{\theta}{\theta+\mu}\right)^{\theta}\left(\frac{\mu}{\theta+\mu}\right)^{x}\end{aligned} \tag{3}$$

On the other hand, four separate decoders for background mean $\mu_b$, foreground mean $\mu_f$, protein mixing $\pi_P$ of background and foreground, and protein dispersion $\theta_P$ are used to decode protein data. Negative binomial mixture loss comparing the decoded protein signal components with observed protein data was implemented to guide reconstruction procedure (Equation 4, indices are dropped for readability). Additionally, a KL divergence term, which utilizes a protein specific prior for the background mean learned during training, is computed.

$$\begin{aligned}\mathcal{L}_{\mathrm{P}} &= \pi\mathrm{NB}(x; \mu_b, \theta) + (1-\pi)\mathrm{NB}(x; \mu_f, \theta)\\ \mathcal{L}_{\mathrm{M},2} &= \mathrm{KL}(\mathcal{N}(\alpha, \beta)||\mathcal{N}(\alpha_{prior}, \beta_{prior}))\end{aligned} \tag{4}$$

Decoding procedure during training is very similar for both models, except that MultiCPA (PoE) uses multiple respective latent space embedding instead. Optimization of the models during model training process repeats two successive steps. A batch from observed data passes through encoder networks to compute joint embedding $Z_{joint}$, which feeds into perturbation and covariate discriminator adversarial networks. The stability of adversarial networks are improved by the implementation of gradient penalty to prevent gradients with large norm values. In the next iteration, joint embedding $Z_{joint}$ is combined with perturbation and covariate embeddings in the latent space and then decoded through multiple decoders. Here, the total loss for backpropagation is given by Equation 5, where $w_i$ are model hyperparameters.

$$\begin{aligned}\mathcal{L}_{\mathrm{MultiCPA}} = \mathcal{L}_{\mathrm{G}} + \mathcal{L}_{\mathrm{P}} w_1 + (\mathcal{L}_{\mathrm{M},1} + \mathcal{L}_{\mathrm{M},2}) w_2\\ - (\mathcal{L}_{\mathrm{A},1} + \mathcal{L}_{\mathrm{A},2}) w_3\end{aligned} \tag{5}$$




(a) (b) (c)

*Figure 2.* a) Prediction performances of the best models after hyperparameter tuning for the two MultiCPA models on *test* split. b) Comparison of MultiCPA and CPA in terms of *OOD* split prediction. c) Baseline expectation under random model behaviour.

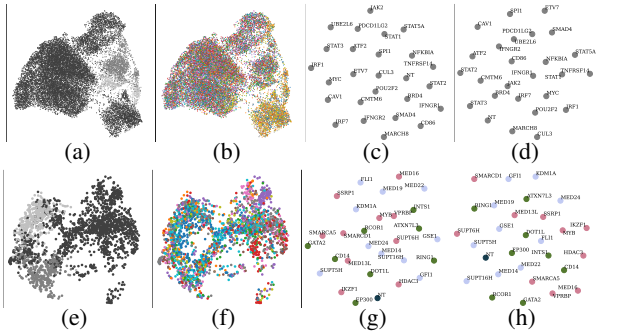
### 2.3. Hyperparameter tuning and datasets

Best hyperparameters of both models were extensively searched in a large hyperparameter space using the experiment management tool *sacred* (Greff et al., 2017) and a *MongoDB* (Gyorodi et al., 2015) experiment database on a large computer cluster with iterative sweeps. A certain set of hyperparameters were selected for each dataset and model combination, which optimize the dataset-specific counterfactual prediction accuracy in later analyses. Best models after hyperparameter tuning were manually examined in terms of reconstruction and adversarial losses in *train*, *test* and *OOD* (out-of-distribution) splits for both datasets to identify any possible overfitting issues. Two CITE-seq datasets of THP-1 human monocytic cells were used for model training and subsequent analyses. First dataset, named as *Wessels2022* (Wessels et al., 2022) has 30707 cells with 16920 genes and 24 proteins for 28 single gene knock-out perturbations. Second dataset, named as *Papalexi2021*, (Papalexi et al., 2021) has 20729 cells with 18649 genes and 4 proteins for 26 single gene knock-out perturbations, but no double perturbation. The datasets were quality checked, visualized and preprocessed using *scanpy* (Wolf et al., 2018). 5000 highly variable genes (HVGs) were selected for training the models. For each perturbation in each dataset, 20 differentially expressed (DE) genes were calculated to assess the model performances at test time.

## 3. Results

### 3.1. Choosing the best model architecture

Two VAE architectures with alternative mixture models were compared on the *Papalexi2021* dataset. Approximately 5000 models for each of the proposed architectures were trained on a high performance computing cluster to find best hyperparameter combination. The best models were chosen in terms of counterfactual perturbation prediction accuracy. Here, coefficient of determination, $R^2$, is determined as the metric when comparing observed data with model predictions. The results in Figure 2a show that MultiCPA model with concatenation mixture module outperformed PoE based model in predicting protein data (0.98 vs. 0.57), although the reconstruction and adversarial losses were comparable.



(a) (b) (c) (d)

(e) (f) (g) (h)

*Figure 3.* UMAP visualizations of input datasets and perturbation embeddings. First row is for *Papalexi2021*, second row is for *Wessels2022* for which only single perturbations are plotted for easier visual comparison. a, e) Input dataset UMAPs, where colors represent division phases G1, G2M, S, from darkest to lightest. b, f) Input dataset UMAPs, where colors represent single perturbations. c, g) Perturbation embeddings UMAPs for MultiCPA. d, h) Perturbation embeddings UMAPs for CPA.

MultiCPA (concat) was hence chosen as the best model in terms of total accuracy in counterfactual prediction, and MultiCPA (PoE) was omitted from subsequent analyses.

### 3.2. MultiCPA outperforms CPA leveraging additional modalities

MultiCPA was compared with CPA model in terms of the prediction accuracy on out-of-distribution (*OOD*) split. Both models were thus trained with *Wessels2022* dataset, where six perturbation combinations had been completely removed from the dataset and labelled as *OOD*. Figure 2b shows that MultiCPA model predicts counterfactual gene expression with slightly higher accuracy ($0.96 \pm 0.02$ vs. $0.95 \pm 0.02$). Nevertheless, for all genes and differentially expressed (DE) genes in the dataset, both CPA and MultiCPA performed robustly ($0.89 \pm 0.07$ vs. $0.88 \pm 0.05$). This suggests both MultiCPA and CPA not only extract and learn the individual effects of perturbations from *train* split, but also successfully combine these information in the latent space to predict the effect of unseen perturbation combinations. Additionally, MultiCPA predicts protein data with a very high accuracy for unseen perturbation combinations, leading to position itself as a multimodal extension of CPA.

Perturbation latent space in both datasets was inspected in order to assess whether perturbation embeddings could give clues about the similarities of perturbations' mode of action. CPA has been previously shown to be competent in grouping perturbations together which are associated via cellular regulatory, metabolic or signaling pathways. It was hypothesized that additional protein information integrated with the MultiCPA model should result in a better resolution and biologically more meaningful groupings in perturbation latent space. Based on manual annotation of perturbations considering underlying biological mechanisms, MultiCPA

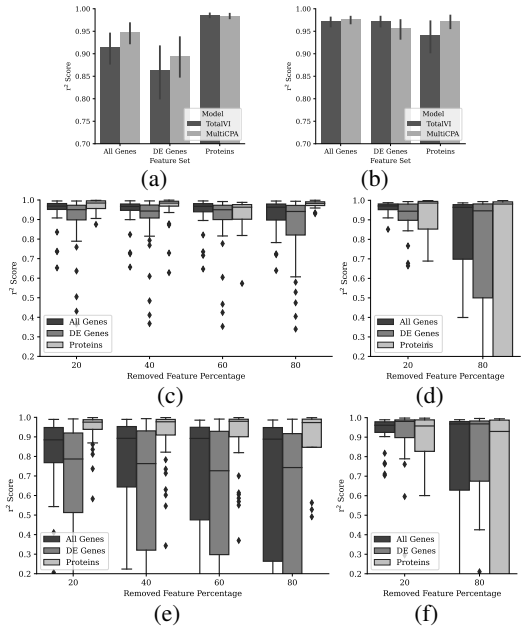



*Figure 4.* Counterfactual perturbation prediction performances on *test* split, respectively for *Wessels2022* and *Papalexi2021* datasets at each row. a, b) Prediction performances of totalVI and MultiCPA models. c, d) Prediction performances of MultiCPA model with noise added datasets. e, f) Prediction performances of totalVI model with noise added datasets.

and also CPA were not able to group the perturbations in both datasets in a consistent fashion, suggesting a dataset-specific variation or potential issues that still need to be accounted for (Figure 3). On the other hand, both models could successfully extract perturbation information from the input datasets although the datapoints of each perturbation do not cluster together on input feature set based UMAPs.

### 3.3. Comparison with existing deep learning models

MultiCPA and totalVI are compared to assess their relative advantages in integration of a multimodal single-cell datasets to make counterfactual predictions. Learning the effect of individual perturbations from the input dataset and combining learned information to make a counterfactual combinational perturbation prediction is not possible using totalVI. The comparative analysis is thus conducted on *test* split only, but not *OOD* split. Considering each perturbation as a different batch, *batch transformation* method of totalVI was applied to unperturbed cells in the *test* set into the perturbation of interest to obtain model predictions of perturbation effects. Analogously, MultiCPA predictions were made only using the unperturbed cells. It was observed that MultiCPA compares slightly favorably to the existing totalVI method (Figure 4a and 4b).

With the idea of testing the robustness of the model to the noise found in the data, certain percentages of the input features were randomly selected from a random quarter of the perturbations in both datasets, and were replaced with zero values. Both totalVI and MultiCPA models were then trained with modified datasets using the same hyperparameters tuned for the complete datasets, and counterfactual prediction accuracies were calculated as usual. totalVI model was considerably affected by such an intervention while MultiCPA still retains high prediction accuracies even with the most severe scenario for *Wessels2022* dataset (Figure 4). The responds to the intervention were comparable for *Papalexi2021* dataset. These results suggest that the information of each perturbation effect could be learned via untouched perturbations for MultiCPA but not totalVI, which is then used in test time to predict intervened perturbations in the dataset, as *Wessels2022* dataset contains many combined perturbations. However, *Papalexi2021* dataset contains only single perturbations, making it unlikely to learn intervened perturbations for both models.

## 4. Discussion

Harnessing the strength of multiple modalities in extracting the information regarding cellular effect of perturbations helps to acquire a broader insight into perturbation effects. Here, we showed that MultiCPA can provide a more comprehensive characterization of cellular phenotypes and perturbations in an unbiased manner through a joint multimodal data representation and improves the prediction of unseen perturbation combinations with higher accuracy. Moreover, MultiCPA is the first method that learns the effect of individual perturbations on the surface protein data and uses the information to predict unseen perturbation combinations.

Furthermore, the proposed generative deep learning model outperformes the existing totalVI model for multimodal single-cell data integration with regards to overall prediction accuracy in the *test* split. MultiCPA learns individual perturbation and covariate information from combinations and performs more robustly than totalVI in response to the noise in high-dimensional input data. While totalVI is not devised to learn and combine individual perturbations, MultiCPA exploits learned latent space embeddings to predict unseen combinations in the training data for both modalities. Additionally, relative advantages of two mixture models has been tested in this context, where MultiCPA (concat) was shown to outperform MultiCPA (PoE) in terms of counterfactual prediction accuracy of surface protein data.

We anticipate MultiCPA to guide exploration of the perturbation space, leading to the development of novel therapeutics (Brochado et al., 2018), and facilitating the discovery of the general principles of a cellular machinery (Muscato et al., 2022) in biological research. As future directions, we aim to extend our model with other cellular modalities, such as chromatin accessibility data by scATAC-seq (Lareau et al., 2019; Lotfollahi et al., 2022) technology.

## References


Brochado, A. R., Telzerow, A., Bobonis, J., Banzhaf, M., Mateus, A., Selkrig, J., Huth, E., Bassler, S., Zamarreño Beas, J., Zietek, M., et al. Species-specific activity of antibacterial drug combinations. *Nature*, 559 (7713):259–263, 2018.

Frangieh, C. J., Melms, J. C., Thakore, P. I., Geiger-Schuller, K. R., Ho, P., Luoma, A. M., Cleary, B., Jerby-Arnon, L., Malu, S., Cuoco, M. S., et al. Multimodal pooled perturb-cite-seq screens in patient models define mechanisms of cancer immune evasion. *Nature genetics*, 53(3):332–341, 2021.

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, 2021.

Greff, K., Klein, A., Chovanec, M., Hutter, F., and Schmidhuber, J. The sacred infrastructure for computational research. In *Proceedings of the 16th python in science conference*, volume 28, pp. 49–56, 2017.

Gyorodi, C., Gyorodi, R., Pecherle, G., and Olah, A. A comparative study: Mongodb vs. mysql. In *2015 13th International Conference on Engineering of Modern Electric Systems (EMES)*, pp. 1–6. IEEE, 2015.

Ji, Y., Lotfollahi, M., Wolf, F. A., and Theis, F. J. Machine learning for perturbational single-cell omics. *Cell Systems*, 12(6):522–537, 2021.

Kingma, D. P. and Welling, M. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Kingma, D. P., Salimans, T., and Welling, M. Variational dropout and the local reparameterization trick. *Advances in neural information processing systems*, 28, 2015.

Lareau, C. A., Duarte, F. M., Chew, J. G., Kartha, V. K., Burkett, Z. D., Kohlway, A. S., Pokholok, D., Aryee, M. J., Steemers, F. J., Lebofsky, R., et al. Droplet-based combinatorial indexing for massive-scale single-cell chromatin accessibility. *Nature Biotechnology*, 37(8):916–924, 2019.

Lee, C. and van der Schaar, M. A variational information bottleneck approach to multi-omics data integration. In *International Conference on Artificial Intelligence and Statistics*, pp. 1513–1521. PMLR, 2021.

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

Lotfollahi, M., Susmelj, A. K., De Donno, C., Ji, Y., Ibarra, I. L., Wolf, F. A., Yakubova, N., Theis, F. J., and Lopez-Paz, D. Learning interpretable cellular responses to complex perturbations in high-throughput screens. *bioRxiv*, 2021.

Lotfollahi, M., Litinetskaya, A., and Theis, F. J. Multigrate: single-cell multi-omic data integration. *bioRxiv*, 2022.

Mimitou, E. P., Cheng, A., Montalbano, A., Hao, S., Stoeckius, M., Legut, M., Roush, T., Herrera, A., Papalexi, E., Ouyang, Z., et al. Multiplexed detection of proteins, transcriptomes, clonotypes and crispr perturbations in single cells. *Nature methods*, 16(5):409–412, 2019.

Muscato, J. D., Morris, H. G., Mychack, A., Rajagopal, M., Baidin, V., Hesser, A. R., Lee, W., Inecik, K., Wilson, L. J., Kraml, C. M., et al. Rapid inhibitor discovery by exploiting synthetic lethality. *Journal of the American Chemical Society*, 144(8):3696–3705, 2022.

Papalexi, E., Mimitou, E. P., Butler, A. W., Foster, S., Bracken, B., Mauck, W. M., Wessels, H.-H., Hao, Y., Yeung, B. Z., Smibert, P., et al. Characterizing the molecular regulation of inhibitory immune checkpoints with multimodal single-cell screens. *Nature genetics*, 53(3): 322–331, 2021.

Stephenson, E., Reynolds, G., Botting, R. A., Calero-Nieto, F. J., Morgan, M. D., Tuong, Z. K., Bach, K., Sungnak, W., Worlock, K. B., Yoshida, M., et al. Single-cell multi-omics analysis of the immune response in covid-19. *Nature medicine*, 27(5):904–916, 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, 2017.

Teichmann, S. and Efremova, M. Method of the year 2019: single-cell multimodal omics. *Nat. Methods*, 17(1):2020, 2020.

Wessels, H.-H., Méndez-Mancilla, A., Papalexi, E., Mauck, W. M., Lu, L., Morris, J. A., Mimitou, E., Smibert, P., Sanjana, N. E., and Satija, R. Efficient combinatorial targeting of rna transcripts in single cells with cas13 rna perturb-seq. *bioRxiv*, 2022.

Wolf, F. A., Angerer, P., and Theis, F. J. Scanpy: large-scale single-cell gene expression data analysis. *Genome biology*, 19(1):1–5, 2018.

Yao, C., Bora, S. A., Parimon, T., Zaman, T., Friedman, O. A., Palatinus, J. A., Surapaneni, N. S., Matusov, Y. P., Chiang, G. C., Kassar, A. G., et al. Cell-type-specific immune dysregulation in severely ill covid-19 patients. *Cell reports*, 34(1):108590, 2021.