Learning representations of cell populations for image-based profiling using contrastive learning

Anonymous Author(s)

1 Image-based single-cell profiling

High-throughput assays enable quantifying cellular responses at a large scale. Image-based assays are among the most accessible and inexpensive technologies for this, and offer single-cell resolution. In these assays, cell populations are perturbed with compounds or genetic perturbations, stained, and then imaged. By extracting large amounts of quantitative morphological data from these microscopy images, a profile can be created that describes that cell population's phenotype. The profiles of different cell populations can be compared to predict previously unrecognized cell states induced by different experimental perturbations of interest. This method, called image-based cell profiling, is a powerful tool that can be used for drug discovery, functional genomics, and disease phenotyping [1]. Among other applications, image-based profiling has already been used to find drugs for SARS-CoV-2 [2], work towards label-free leukemia detection [3], and predict the impacts of particular gene mutations [4].

2 Introduction

Image-based cell profiling shows great potential, but many steps in its pipeline can still be improved [5]. One of the main challenges is to create a profile that summarizes the many features of a cell population while capturing their natural variations and subpopulations. Cell populations are known to be heterogeneous [6], [7] and recent studies have yielded many insights into its mechanisms and importance, particularly in cancer cells [8]–[11]. Capturing that heterogeneity could improve a profile's information content, and thus utility in various applications.

Nevertheless, so-called population-averaged profiling, where all single-cell features are averaged using either the mean or the median, has remained the most commonly-used approach in the field of image-based profiling, regardless of the type of features or the profile's post-processing [12]. Average profiling is a simple way of summarizing a cell population (henceforth referred to as a sample) into a vector (a sample's profile) with only one value per measured feature. It decreases the data size (as there are typically thousands of cells per well, hundreds of wells per plate, and multiple plates per experiment) and makes downstream analysis simpler.

However, by using average profiling, information on cell subpopulations is lost. This can result in identical average profiles despite cell populations having various subpopulation configurations. In that case, two profiles can be indistinguishable even though one is created from a sample that contains multiple subpopulations while the other sample does not contain any of those subpopulations. Additionally, not taking subpopulations into account can lead to a quantitatively incorrect interpretation. For example, two cell populations can show correlations among certain features when averaged but show completely different relations when compared after grouping the cells, i.e., Simpson's paradox [13]. Finally, by averaging a sample, the assumption that the joint distribution of the measured features is unimodal can lead to artifacts if violated.

Submitted to Learning Meaningful Representations of Life (LMRL) Workshop at NeurIPS 2022. Do not distribute.

3 Previous research

Several methods have been proposed to capture the heterogeneity of cell populations into their corresponding profiles. The most straightforward solution is to incorporate the cell population's dispersion for each feature and concatenate these values with the average profile. However, this approach comes with its own limitations and only leads to minor improvements over average profiling alone [14], [15]. A different approach involves first clustering cells either in an unsupervised or supervised way and then calculating the profiles based on their subpopulations [16], [17]. These methods capture more information about subpopulations rather than only incorporating their dispersions. However, they can lead to incomparable profiles across samples due to a varying number of subpopulations and many cell phenotypes are better described with a continuous rather than a discrete scale [12]. Moreover, these methods also did not significantly improve upon average profiling. In fact, a comparison study of profiling methods found that population means performed just as well as those that took advantage of cell heterogeneity [15].

Recently, however, the performance of average profiling was beaten by fusing features' averages, dispersion, and covariances [12]. This method provided 20% better performance in predicting a compound's mechanism of action and a gene's pathway, showing that capturing cell population heterogeneity can improve profile strength. However, this method has two major limitations. First, it only captures the first and second order moments of the data. Second, because it produces a similarity matrix rather than an embedding, it requires recomputing the pairwise similarities among all profiles each time a new profile is included in the dataset. Thus, a more accessible method for capturing single-cell heterogeneity is required to increase profile strength in practice. We propose a novel learning-based method that addresses both these limitations and automatically finds an effective way to aggregate single-cell data to improve the information content of sample profiles.

4 Method

We propose a weakly-supervised contrastive learning approach that uses information naturally available in profiling experiments. Specifically, perturbation ids are used as labels for learning a latent feature space. In this feature space, profiles of replicates of the same perturbation should be close to each other and different perturbations far away. This type of labeling frames the issue as a multiple-instance learning problem [18], which assumes that the replicate wells consist of cells with similar feature distributions and that different compounds produce populations with different feature distributions. Both of these are approximations of reality, but could potentially produce a feature embedding that captures biologically important variations in morphology.

The data is considered to be a collection of sets of cells, where each sample corresponds to one set (using the mathematical definition of "set"). A function that aggregates cells from such a sample into a profile requires a few properties. First, the function should be able to handle arbitrarily sized samples as an input. Second, because cells within a sample by definition have no order, the function should be permutation invariant. There are a few methods that have been developed for analyzing this type of data [19], [20], but a general formulation for solving this type of problem is known as Deep Sets [21]. Zaheer et al. [21] show that a universal approximator on sets has a fixed form, which provides a backbone for building neural networks to process them. In this study, the Deep Sets formulation is used to learn the best way of aggregating single-cell feature data into a profile that allows for better prediction of a compound's mechanism of action compared to averaged profiles. This is achieved by applying weakly-supervised contrastive learning in a multiple instance learning setting.

5 Results

We tested this method on the cpg0000 dataset [22], from the JUMP consortium [23], available from the Cell Painting Gallery on the Registry of Open Data on AWS (https://registry.opendata.aws/cellpainting-gallery/). This dataset consists of 90 different compound perturbations with 4 replicates per plate. On this dataset, our proposed model provides a more accessible and better performing method for aggregating single-cell feature data than previously published strategies and the average profiling baseline. Based on an interpretability analysis, it is likely that the model achieves this by performing some form of quality control by filtering out noisy cells and prioritizing less noisy cells. Remarkably, the model could also mitigate batch effects, even though this was not part of the training objective. This shows that the learned latent representation of the model prioritizes biological signal over technical variance, both on the cell level and the plate level. The model cannot be directly transferred to unseen batch data; however, it can readily be used by training on new data and inferring the improved profiles directly after because the labels required for training are naturally available in cell profiling experiments.

6 Future work

Our current and future work focuses on validating these conclusions on a much larger dataset which consists of 1.258 different compound perturbations, namely the LINCS dataset [24]. If successful, the method could improve the effectiveness of future cell profiling studies with the investment of additional computation time.

References

[1] J. C. Caicedo, S. Singh, and A. E. Carpenter, "Applications in image-based profiling of perturbations," Curr. Opin. Biotechnol., vol. 39, pp. 134–142, Jun. 2016.

[2] C. Mirabelli et al., "Morphological cell profiling of SARS-CoV-2 infection identifies drug repurposing candidates for COVID-19," Proc. Natl. Acad. Sci. U. S. A., vol. 118, no. 36, Sep. 2021, doi: 10.1073/pnas.2105815118.

[3] M. Doan et al., "Label-Free Leukemia Monitoring by Computer Vision," Cytometry A, vol. 97, no. 4, pp. 407–414, Apr. 2020.

[4] J. C. Caicedo et al., "Cell Painting predicts impact of lung cancer variants," Mol. Biol. Cell, vol. 33, no. 6, p. ar49, May 2022.

[5] S. N. Chandrasekaran, H. Ceulemans, J. D. Boyd, and A. E. Carpenter, "Image-based profiling for drug discovery: due for a machine-learning upgrade?," Nat. Rev. Drug Discov., vol. 20, no. 2, pp. 145–159, Feb. 2021.

[6] S. J. Altschuler and L. F. Wu, "Cellular heterogeneity: do differences make a difference?," Cell, vol. 141, no. 4, pp. 559–563, May 2010.

[7] K. A. Janes, "Single-cell states versus single-cell atlases - two classes of heterogeneity that differ in meaning and method," Curr. Opin. Biotechnol., vol. 39, pp. 120–125, Jun. 2016.

[8] A. Marusyk and K. Polyak, "Tumor heterogeneity: causes and consequences," Biochim. Biophys. Acta, vol. 1805, no. 1, pp. 105–117, Jan. 2010.

[9] D. Deb et al., "Combination Therapy Targeting BCL6 and Phospho-STAT3 Defeats Intratumor Heterogeneity in a Subset of Non-Small Cell Lung Cancers," Cancer Res., vol. 77, no. 11, pp. 3070–3081, Jun. 2017.

[10] L. Keller and K. Pantel, "Unravelling tumour heterogeneity by single-cell profiling of circulating tumour cells," Nat. Rev. Cancer, vol. 19, no. 10, pp. 553–567, Oct. 2019.

[11] J. Goveia et al., "An Integrated Gene Expression Landscape Profiling Approach to Identify Lung Tumor Endothelial Cell Heterogeneity and Angiogenic Candidates," Cancer Cell, vol. 37, no. 1, pp. 21–36.e13, Jan. 2020.

[12] M. H. Rohban, H. S. Abbasi, S. Singh, and A. E. Carpenter, "Capturing single-cell heterogeneity via data fusion improves image-based profiling," Nat. Commun., vol. 10, no. 1, p. 2082, May 2019.

[13] C. Trapnell, "Defining cell types and states with single-cell genomics," Genome Res., vol. 25, no. 10, pp. 1491–1498, Oct. 2015.

[14] B. Wang et al., "Similarity network fusion for aggregating data types on a genomic scale," Nat. Methods, vol. 11, no. 3, pp. 333–337, Mar. 2014.

[15] V. Ljosa et al., "Comparison of methods for image-based profiling of cellular morphological responses to small-molecule treatment," J. Biomol. Screen., vol. 18, no. 10, pp. 1321–1329, Dec. 2013.

[16] L.-H. Loo, H.-J. Lin, R. J. Steininger 3rd, Y. Wang, L. F. Wu, and S. J. Altschuler, “An approach for extensibly profiling the molecular states of cellular subpopulations,” Nat. Methods, vol. 6, no. 10, pp. 759–765, Oct. 2009.

[17] F. Fuchs et al., “Clustering phenotype populations by genome-wide RNAi and multiparametric imaging,” Mol. Syst. Biol., vol. 6, p. 370, Jun. 2010.

[18] Maron and Lozano-Pérez, “A Framework for Multiple-Instance Learning,” Adv. Neural Inf. Process. Syst., Jun. 1997, [Online]. Available: https://proceedings.neurips.cc/paper/1997/file/82965d4ed8150294d4330ace00821d77-Paper.pdf

[19] H. Edwards and A. Storkey, “Towards a Neural Statistician,” arXiv [stat.ML], Jun. 07, 2016. [Online]. Available: http://arxiv.org/abs/1606.02185

[20] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, “PointNet++: Deep Hierarchical Feature Learning on Point Sets in a Metric Space,” arXiv [cs.CV], Jun. 07, 2017. [Online]. Available: http://arxiv.org/abs/1706.02413

[21] M. Zaheer, S. Kottur, S. Ravanbakhsh, B. Poczos, R. Salakhutdinov, and A. Smola, “Deep Sets,” arXiv [cs.LG], Mar. 10, 2017. [Online]. Available: http://arxiv.org/abs/1703.06114

[22] S. N. Chandrasekaran et al., “Three million images and morphological profiles of cells treated with matched chemical and genetic perturbations,” bioRxiv, p. 2022.01.05.475090, Jan. 05, 2022. doi: 10.1101/2022.01.05.475090.

[23] A. Mullard, “Machine learning brings cell imaging promises into focus,” Nat. Rev. Drug Discov., vol. 18, no. 9, pp. 653–655, Sep. 2019.

[24] Way, Gregory P., Ted Natoli, Adeniyi Adeboye, Lev Litichevskiy, Andrew Yang, Xiaodong Lu, Juan C. Caicedo, et al. 2021. “Morphology and Gene Expression Profiling Provide Complementary Information for Mapping Cell State.” bioRxiv. https://doi.org/10.1101/2021.10.21.465335.