690
doi:10.1017/S1431927620015548
Microsc. Microanal. 26 (Suppl 2), 2020

Learning Biology Through Puzzle-solving: Unbiased Automatic Understanding of Microscopy Images with Self-supervised Learning

Alex Lu[1], Oren Kraus[2], Sam Cooper[2] and Alan Moses[1]

[1]University of Toronto, Toronto, Ontario, Canada, [2]Phenomic AI, Toronto, Ontario, Canada

To extract information about biology from microscopy images, researchers rely on features that measure relevant image properties, like the shape and size of cells, or the intensity of fluorescent markers[1,2]. However, developing a set of features that robustly represents the biology of interest is challenging. A good representation usually involves either extensive engineering by experts to produce manually-designed features[3,4], or annotating large labeled training datasets to enable supervised deep learning[5,6]. Both options are laborious, creating a bottleneck in computational analysis.

To address this problem, we investigated self-supervised learning. Self-supervised learning methods train deep learning models to solve autonomous proxy tasks[7]. The proxy tasks do not need to produce useful outputs, and are only meant to teach the models transferable skills and perceptions of data: self-supervised proxy tasks often resemble puzzles, like solving jigsaw puzzles or determining how an image has been rotated. We created a self-supervised method[8], called "paired cell inpainting", designed to learn representations of protein biology from multi-channel fluorescent microscopy images (Figure 1). Given one cell, the model is asked to produce a synthetic image of inferred protein expression for another cell from the same well. As our proxy task can be defined using the structure of microscopy data alone, our models do not require any manual labelling efforts to train.

We show that our self-supervised models learn effective representations of protein biology, that outperform other feature representations when purposed for analyses like classifying protein subcellular localization in images of single cells. Our method is highly general: we learned similarly high-quality representations for proteome-wide image screens[9–12] originating from different labs employing different imaging modalities and fluorescent markers (Figure 2), including two technically-challenging datasets that have never been analyzed computationally previously.

Self-supervised methods learn representations unbiased by expert pre-conceptions of biology, as they learn through problem-solving on data directly. Consequentially, our representations can be used to identify biologically-relevant subclasses in high-throughput image screens, which are not as evident in other representations trained using expert labels for pre-defined classes. This property makes our method especially useful for exploratory analysis: I will demonstrate how our representations can be analyzed with unsupervised clustering methods to discover novel hypotheses.

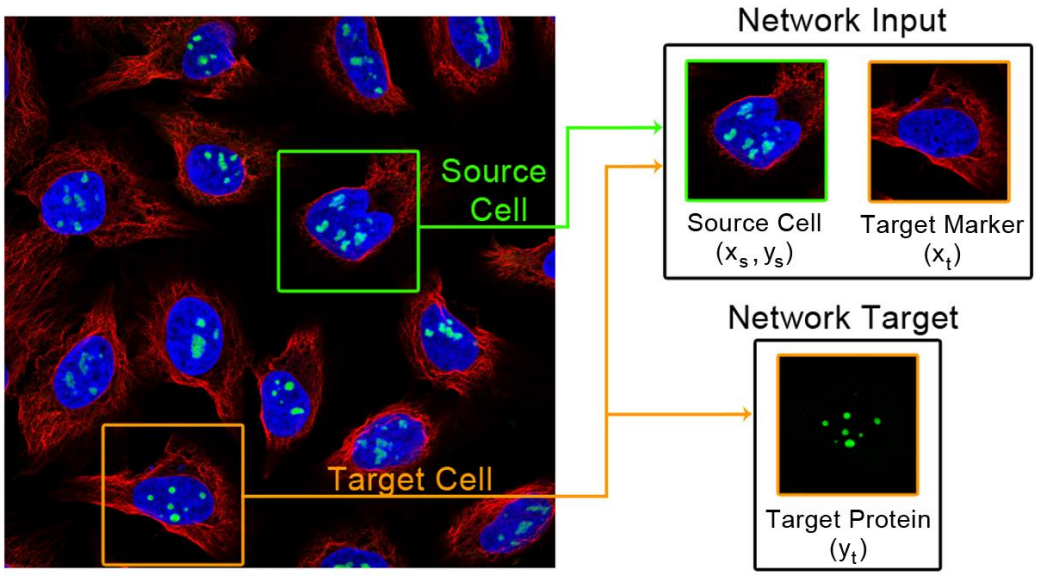

Figure 1. Training inputs and targets. We crop a source cell (green box) and a target cell (orange box) from the same image. Then, given all channels for the source cell, and the shape of the target cell (in this dataset, given by the nucleus and the microtubule channels), a convolutional neural network is trained to create an image of the protein channel in the target cell. Images shown are of human cells, with the nucleus colored blue, microtubules colored red, and a specific protein colored green.

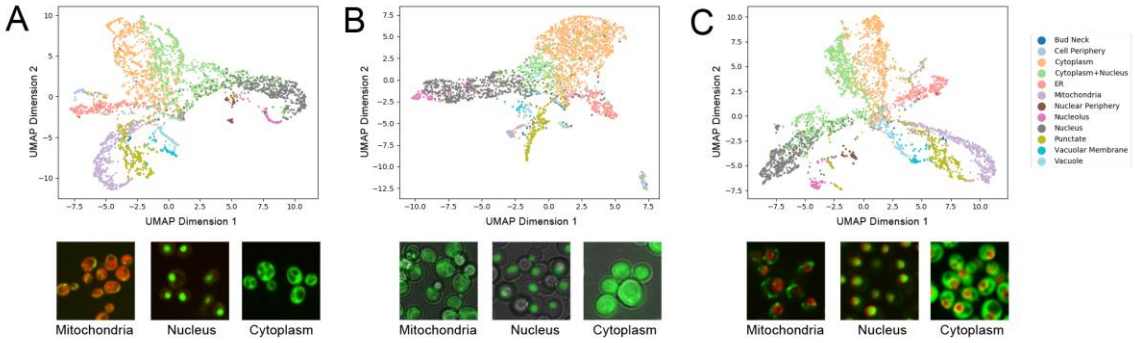

Figure 2. UMAP scatterplots of protein-level paired cell inpainting representations for three independent yeast image datasets: A) the CyCLOPS dataset, B) a brightfield dataset from YeastRGB, C) a nuclear pore dataset published by Tkach et al. (2012). We generate UMAPS with the same parameters (Euclidean distance, 20 neighbors, minimum distance of 0.3). We show representative images from each dataset. In all images, protein expression is shown in green; each image shows a distinct fluorescently tagged protein.

References

1. Uchida, S. Image processing and recognition for biological images. *Dev. Growth Differ.* **55**, 523 (2013).
2. Bengio, Y., Courville, A. & Vincent, P. Representation Learning: A Review and New Perspectives. (2012).
3. Handfield, L.-F., Chong, Y. T., Simmons, J., Andrews, B. J. & Moses, A. M. Unsupervised clustering of subcellular protein expression patterns in high-throughput microscopy images reveals protein complexes and functional relationships between proteins. *PLoS Comput. Biol.* **9**, e1003085 (2013).

https://doi.org/10.1017/S1431927620015548 Published online by Cambridge University Press

4. Li, Y., Majarian, T. D., Naik, A. W., Johnson, G. R. & Murphy, R. F. Point process models for localization and interdependence of punctate cellular structures. *Cytom. Part A* **89**, 633–643 (2016).
5. Kraus, O. Z. *et al.* Automated analysis of high-content microscopy data with deep learning. *Mol. Syst. Biol.* **13**, (2017).
6. Sullivan, D. P. *et al.* Deep learning is combined with massive-scale citizen science to improve large-scale image classification. *Nat. Biotechnol.* **36**, 820–828 (2018).
7. Jing, L. & Tian, Y. Self-supervised Visual Feature Learning with Deep Neural Networks: A Survey. (2019).
8. Lu, A. X., Kraus, O. Z., Cooper, S. & Moses, A. M. Learning unsupervised feature representations for single cell microscopy images with paired cell inpainting. *PLOS Comput. Biol.* **15**, e1007348 (2019).
9. Chong, Y. T. *et al.* Yeast Proteome Dynamics from Single Cell Imaging and Automated Analysis. *Cell* **161**, 1413–1424 (2015).
10. Tkach, J. M. *et al.* Dissecting DNA damage response pathways by analysing protein localization and abundance changes during DNA replication stress. *Nat. Cell Biol.* **14**, 966–76 (2012).
11. Dubreuil, B. *et al.* YeastRGB: comparing the abundance and localization of yeast proteins across cells and libraries. *Nucleic Acids Res.* (2018) doi:10.1093/nar/gky941.
12. Thul, P. J. *et al.* A subcellular map of the human proteome. *Science (80-.).* **356**, eaal3321 (2017).

https://doi.org/10.1017/S1431927620015548 Published online by Cambridge University Press