Embedding cell state dynamics via contrastive learning of representations of 3D dynamic imaging datasets

Eduardo Hirata-Miyasaki[1,*], Soorya Pradeep[1,*], Ziwen Liu[1,*], Alishba Imran[1,2,*]
Taylla Milena Theodoro[1], Ivan E. Ivanov[1], Sudip Khadka[1],
See-Chi Lee [1], Michelle Grunberg [1], Hunter Woosley[1], Madhura Bhave[1],
Carolina Arias[1], Shalin B. Mehta[1,†]

[1] Chan Zuckerberg Biohub San Francisco, San Francisco, CA 94158, USA
[2] University of California Berkeley, Berkeley, CA 94720, USA

Abstract

Robust and scalable profiling of cell state dynamics from large-scale 3D live cell imaging data is an open challenge. We propose a self-supervised method for embedding cell state **dyna**mics via **c**ontrastive **l**earning of **r**epresentations (DynaCLR) to address this need. DynaCLR integrates single-cell tracking and time-aware contrastive sampling to learn robust, temporally regularized representations of morphological dynamics. This pretext task leads to an embedding space in which distances encode transitions in cell state dynamics. DynaCLR embeddings generalize to out-of-distribution imaging experiments, and can be used for multiple downstream tasks with sparse human annotations. DynaCLR embeddings enabled robust classification of cell infection and division, and clustering of heterogeneous cell migration behaviors. DynaCLR is a generalist method for comparative analyses of dynamic cellular responses to pharmacological, microbial, and genetic perturbations. We provide a PyTorch-based implementation of the method and a model library(VisCy) trained with 3D and 2D time-lapse datasets.

1 Introduction

Learning biologically interpretable representations of the cell morphology and architecture from 100 TB-scale dynamic imaging datasets is an outstanding need in basic biology and therapeutic discovery. The dynamic responses of organelles and cells to perturbations such as infection, gene expression modulation, or pharmacological treatment can reveal biomarkers of health and disease, and establish causal links between the cell morphology and function. Supervised approaches for analyzing dynamic cell morphology are suboptimal because categorical labeling of continuous changes in cell and organelle morphology is hard. Self-supervised methods that use biologically and experimentally relevant pretext tasks have the potential to learn robust embeddings of cell and organelle dynamics that generalize across experimental conditions, disambiguate the relationships between complex perturbations and cellular responses, and enable the discovery of rare cell states. Current self-supervised embedding methods are not designed to encode multi-channel 3D time-lapse datasets or allow for flexible definition of pretext tasks based on prior knowledge of (dis-)similarity of cell morphologies.

We report a method to learn embeddings of cell state **Dyna**mics via **C**ontrastive **L**earning of **R**epresentations (DynaCLR). DynaCLR combines single-cell tracking with cell and time-aware

*equal contribution
†correspondence: shalin.mehta@czbiohub.org

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: The 3rd Workshop on Imageomics: Discovering Biological Knowledge from Images Using AI.

contrastive sampling to learn embeddings of cell and organelle dynamics from multi-channel 3D time-lapse microscopy data. Using the evolution of cell morphology in time-lapse datasets as a form of augmentation, DynaCLR learns temporally-regularized embeddings that robustly model cell state dynamics. DynaCLR models generalize to out-of-distribution data acquired with diverse imaging systems and cell types, making the learned embeddings useful for robust cell state analysis with few human annotations. We also share a scalable PyTorch implementation for training models on GPU clusters (VisCy).

We evaluate the accuracy of the visual representation learned by our method using metrics specific to the downstream task and metrics agnostic to the downstream task.

2 Background and related work

Representation learning of images of cells [He et al., 2021, Kraus et al., 2023] is accelerating our ability to learn biological relationships from images. In parallel, learning visual representations of objects and scenes from videos [Wang and Gupta, 2015, Denton, 2017, Sermanet et al., 2018, Qian et al., 2021, Dave et al., 2021] has been an active area of computer vision. Among the self-supervised learning approaches, contrastive learning [Hadsell et al., 2006] offers several advantages: it allows the introduction of prior knowledge of the relationships between the data points as a contrastive loss term [Chen et al., 2020, He et al., 2020], it can be used with deterministic or generative models [Aneja et al., 2021], and it enables joint embedding of diverse channels and modalities [Radford et al., 2021].

In cell biology, self-supervised models of time-lapse microscopy data have enabled diverse analyses, e.g., analysis of immune response [Wu et al., 2022, Shannon et al., 2024], profiling of cell lineages [Soelistyo et al., 2022, Ulicna et al., 2023], phenotyping of plant cells [Marin Zapata et al., 2021], and dense representations of cell dynamics [Gallusser et al., 2023]. In parallel, contrastive self-supervised models of static snapshots have enabled analyses of cell and organelle states, e.g., diversity of mitochondrial shapes [Natekar et al., 2023] in response to perturbations, detection of cell division [Zyss et al., 2024], and learning correlation between gene expression and morphology [Wang et al., 2024, Şenbabaoğlu et al., 2024]. Understanding the mechanisms of most dynamic cell state transitions requires time-resolved measurements [Shakarchy et al., 2024, Ulicna et al., 2023]. DynaCLR learns temporally regularized embeddings from trajectories of single cell images, and allows embedding of both snapshots or trajectories.

3 Method

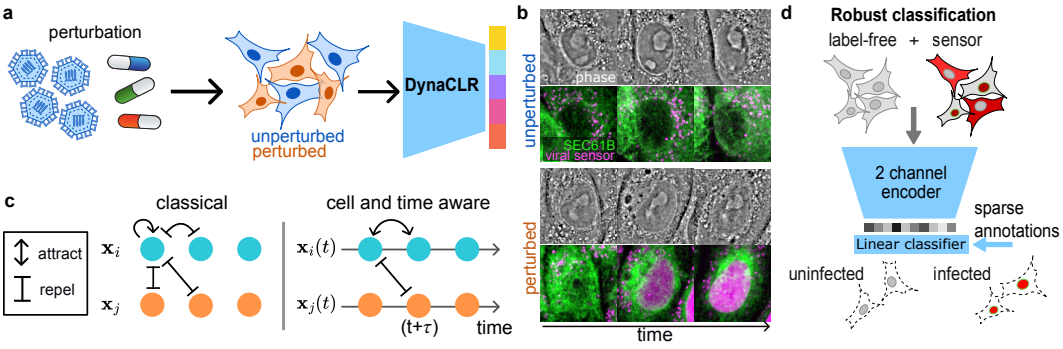


Figure 1: **Summary of DynaCLR:** (a) DynaCLR maps 3D multi-channel images of cells subjected to diverse perturbations to temporally regularized embeddings, (b) Single-cell tracks are used to train DynaCLR models. Illustrative patches of unperturbed and perturbed cells imaged with multiple channels are shown: phase (grayscale), viral sensor (magenta), and endoplasmic reticulum marker SEC61 (green). (c) Contrastive loss with two different sampling strategies, classical, and time-aware, is used to map multi-channel volumes to embedding vectors. (d) The learned embeddings enable robust classification of multiple cell states with efficient annotations (e.g., infection and cell division).

3.1 Time and cell-aware contrastive sampling

DynaCLR method is illustrated in Figure 1a-c, along with single-cell tracking and downstream analyses. We embed 3D multi-channel patches of single cells $\mathbf{x}_i(t) \in \mathbb{R}^{C \cdot Z \cdot Y \cdot X}$, where C denotes

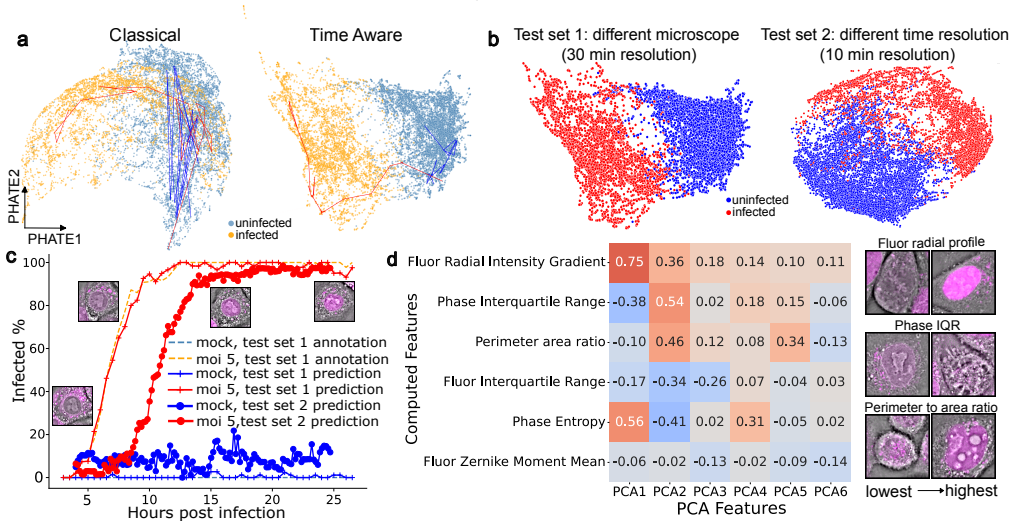


Figure 2: **DynaCLR embeddings generalize across cell types and microscopes:** (a) PHATE maps with classical vs the time-aware sampling. Infected (red) and uninfected (blue) tracks illustrate smoother embedding trajectories via time-aware sampling. (b) Embeddings generalize to a different microscope (1) and higher temporal sampling resolution (2) with clear separation of the infection states. (c) Infection progression over time shows a steady increase in the percentage of infected cells. (d) Correlation of morphological features with embedding principal components (PCs). Representative image patches highlight morphological differences associated with infection states.

channels and Z, Y, X are spatial dimensions. The cells are subjected to different perturbations, including the intrinsic perturbations of cell cycle (Figure 1b). DynaCLR method can be used with diverse channels, including fluorescence channels that report molecular architecture and label-free channels [3] that report physical architecture. The cells $\mathbf{x}_i$ are tracked across time $t_1, t_2, \ldots, t_n$ as they transition through different states, e.g., division, infection, death, and innate immune response. DynaCLR models are trained with a set $\{\mathbf{x}_i(t)\}$ of tracks using different contrastive sampling strategies, where i is the track ID and not a batch index. DynaCLR models (f) map the tracks in image space to temporally regularized embeddings $\mathbf{z}_i(t) = f[\mathbf{x}_i(t)]$. We trained DynaCLR models using either explicit negative sampling and triplet loss [Weinberger et al., 2005] or implicit negative sampling and NT-Xent loss [Chen et al., 2020].

We evaluate two sampling strategies (Figure 1c, Appendix A):

- **Classical sampling** follows the classical contrastive sampling of natural images without time or cell identity, treating each frame independently by forming the positive pairs from augmented views of the same image.
- **Time aware sampling** uses tracked images of the same cell at t and $t + \tau$ as positive pairs and images of other cells as negatives. This pretext task minimizes distances between temporally adjacent embeddings and maximizes distances between embeddings of unrelated cells, thereby regularizing trajectories in embedding space.

DynaCLR embeddings can enable multiple downstream analyses of cell states. Here, we evaluate using task-specific metrics for infection and division classification and task-agnostic metrics of temporal smoothness and dynamic range, which quantify the continuity and variability of trajectories in embedding space [Wu et al., 2022].

The model architecture, training, and data augmentations are described in the Appendix A and Table 1. The mathematical formulations of the sampling strategies and temporal regularization metrics (e.g, smoothness and dynamic range) are defined in Section A.2 and Section A.4. The details of the acquisition, preprocessing, and annotations are summarized in Appendix B, and the models are summarized in Table 2 and Table 3.

[3] Note that *label-free* in the context of biological microscopy implies the absence of fluorescent labeling of cells and not necessarily the absence of human annotations of cell states.

4 Experiments

4.1 Temporal regularization via time-aware contrastive sampling

We evaluated the effect of time-aware sampling on embedding smoothness and structure using the Dengue-infected cell dataset containing mock (MOI = 0) and infected (MOI = 5) conditions (Figure 2, Video 1). Models trained with time-aware sampling produced smoother trajectories and higher dynamic range than classical sampling (Table 4), as evident from PHATE projections of test cells (Figure 2a, Figure S2).

The same embeddings also enabled robust detection of cell division in uninfected and infected cells, showing smooth transitions from interphase to mitosis, whereas models without temporal regularization produced noisier embedding trajectories (Figure S2c and f).

4.2 Generalization and robust cell state classification

The embeddings were classified with a linear classifier trained only on a few annotations for infection state and cell division (Figure 2). Figure 2b shows the PHATE visualization of the embeddings of the test datasets from a different microscope and time resolution, with an overlay of the predicted class. The predicted infection percentages matched expert annotations for mock and MOI 5 conditions, rising and plateauing near 12 hours post-infection (HPI) (Section B.2). A similar trend was observed in the independent test data, where infections plateaued at 15 HPI (Figure 2c). Thus, the infection classification model trained with DynaCLR method demonstrated robust generalization across microscopes and multiple experiments.

We compared DynaCLR to ImageNet- and OpenPhenom-pretrained models, which are limited to a single channel (viral sensor or phase images). While these pretrained models achieve similar F1 scores (Figure S3) for infection state classification using the sensor channel where the phenotype is easy to see, they don't separate the cell states encoded in dense phase channel and produce embeddings that are temporally irregular.

To check that DynaCLR model trained on infected cells indeed learned biologically relevant phenotypes, we computed rank correlation between principal components (PCs) of the learned embeddings and select engineered features(Figure 2d). PC1 reflected radial redistribution of fluorescence reporter and increased phase roughness, while PC2 captured morphological changes (phase IQR, perimeter-to-area ratio, negative fluorescence IQR correlation), respectively (Figure 2d). In addition, visual inspection of cell patches along PC axes explain the variations in cell and organelle morphology in the embedding space (Figure 2d). Detailed correlations between principal components and image features are provided in (Figure S4). These correlations demonstrate that the model is sensitive to changes in image features relevant to biology, i.e., changes in the localization of viral sensor and roughness of cell density.

Temporally regularized DynaCLR embeddings also remained robust to tracking errors (Figure S5) and could improve downstream tracking across complex morphological changes. DynaCLR further generalized to diverse dynamic processes beyond infection: models trained on cell division (ALFI dataset) and microglial morphodynamics (Figure S1a–b) separated distinct morphological states across cell types and perturbations. Compared to prior temporally regularized VQ-VAE model [Wu et al., 2022], DynaCLR achieved clearer separation of heterogeneous dynamics (Figure S1).

Additional results on cell-cycle and migration dynamics are provided in the Supplement Figure S1

5 Limitations

The key limitations of DynaCLR method relative to the published self-supervised representation learning methods of time-lapse data are: 1) DynaCLR requires high enough time sampling so that the cells can be tracked over two neighboring frames, and 2) it does not include a decoder to reconstruct images from embeddings, which may limit the ability to interpret the learned embedding space.

6 Discussion and future work

DynaCLR learns biologically meaningful embeddings of dynamic cell morphology, enabling robust cell-state classification across modalities and cell types. Future work will extend these embeddings to downstream tasks such as robust tracking, event synchronization, and integration with -omics measurements (to be detailed in a forthcoming preprint).

References

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked Autoencoders Are Scalable Vision Learners, December 2021.

Oren Kraus, Kian Kenyon-Dean, Saber Saberian, Maryam Fallah, Peter McLean, Jess Leung, Vasudev Sharma, Ayla Khan, Jia Balakrishnan, Safiye Celik, Maciej Sypetkowski, Chi Vicky Cheng, Kristen Morse, Maureen Makes, Ben Mabey, and Berton Earnshaw. Masked Autoencoders are Scalable Learners of Cellular Morphology, November 2023.

Xiaolong Wang and Abhinav Gupta. Unsupervised learning of visual representations using videos. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2794–2802, 2015.

Emily L. Denton. Unsupervised learning of disentangled representations from video. *Advances in neural information processing systems*, 30, 2017.

Pierre Sermanet, Corey Lynch, Yevgen Chebotar, Jasmine Hsu, Eric Jang, Stefan Schaal, and Sergey Levine. Time-Contrastive Networks: Self-Supervised Learning from Video, March 2018.

Rui Qian, Tianjian Meng, Boqing Gong, Ming-Hsuan Yang, Huisheng Wang, Serge Belongie, and Yin Cui. Spatiotemporal Contrastive Video Representation Learning. *arXiv:2008.03800 [cs]*, April 2021.

Ishan Dave, Rohit Gupta, Mamshad Nayeem Rizve, and Mubarak Shah. TCLR: Temporal Contrastive Learning for Video Representation. *arXiv:2101.07974 [cs]*, April 2021.

R. Hadsell, S. Chopra, and Y. LeCun. Dimensionality Reduction by Learning an Invariant Mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1735–1742, June 2006. doi: 10.1109/CVPR.2006.100.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A Simple Framework for Contrastive Learning of Visual Representations. In *Proceedings of the 37th International Conference on Machine Learning*, pages 1597–1607. PMLR, November 2020.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum Contrast for Unsupervised Visual Representation Learning, March 2020.

Jyoti Aneja, Alex Schwing, Jan Kautz, and Arash Vahdat. A Contrastive Learning Approach for Training Variational Autoencoder Priors. In *Advances in Neural Information Processing Systems*, volume 34, pages 480–493. Curran Associates, Inc., 2021.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision, February 2021.

Zhenqin Wu, Bryant B. Chhun, Galina Popova, Syuan-Ming Guo, Chang N. Kim, Li-Hao Yeh, Tomasz Nowakowski, James Zou, and Shalin B. Mehta. DynaMorph: Self-supervised learning of morphodynamic states of live cells. *Molecular Biology of the Cell*, 33(6):ar59, May 2022. ISSN 1059-1524. doi: 10.1091/mbc.E21-11-0561.

Michael J. Shannon, Shira E. Eisman, Alan R. Lowe, Tyler F. W. Sloan, and Emily M. Mace. cellPLATO – an unsupervised method for identifying cell behaviour in heterogeneous cell trajectory data. *Journal of Cell Science*, 137(20):jcs261887, June 2024. ISSN 0021-9533. doi: 10.1242/jcs.261887.

Christopher J. Soelistyo, Giulia Vallardi, Guillaume Charras, and Alan R. Lowe. Learning biophysical determinants of cell fate with deep neural networks. *Nature Machine Intelligence*, 4(7):636–644, July 2022. ISSN 2522-5839. doi: 10.1038/s42256-022-00503-6.

Kristina Ulicna, Manasi Kelkar, Christopher J. Soelistyo, Guillaume T. Charras, and Alan R. Lowe. Learning dynamic image representations for self-supervised cell cycle annotation, May 2023.

Paula A Marin Zapata, Sina Roth, Dirk Schmutzler, Thomas Wolf, Erica Manesso, and Djork-Arné Clevert. Self-supervised feature extraction from image time series in plant phenotyping using triplet networks. *Bioinformatics*, 37(6):861–867, May 2021. ISSN 1367-4803, 1367-4811. doi: 10.1093/bioinformatics/btaa905.

Benjamin Gallusser, Max Stieber, and Martin Weigert. Self-supervised Dense Representation Learning for Live-Cell Microscopy with Time Arrow Prediction. In Hayit Greenspan, Anant Madabhushi, Parvin Mousavi, Septimiu Salcudean, James Duncan, Tanveer Syeda-Mahmood, and Russell Taylor, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*, pages 537–547, Cham, 2023. Springer Nature Switzerland. ISBN 978-3-031-43993-3. doi: 10.1007/978-3-031-43993-3_52.

Parth Natekar, Zichen Wang, Mehul Arora, Hiroyuki Hakozaki, and Johannes Schöneberg. Self-supervised deep learning uncovers the semantic landscape of drug-induced latent mitochondrial phenotypes. *bioRxiv*, 2023.

Daniel Zyss, Amritansh Sharma, Susana A. Ribeiro, Claire E. Repellin, Oliver Lai, Mary J. C. Ludlam, Thomas Walter, and Amin Fehri. Contrastive learning for cell division detection and tracking in live cell imaging data, August 2024.

Zitong Jerry Wang, Romain Lopez, Jan-Christian Hütter, Takamasa Kudo, Heming Yao, Philipp Hanslovsky, Burkhard Höckendorf, Rahul Moran, David Richmond, and Aviv Regev. Multi-ContrastiveVAE disentangles perturbation effects in single cell images from optical pooled screens, March 2024.

Yasin Şenbabaoğlu, Vignesh Prabhakar, Aminollah Khormali, Jeff Eastham, Evan Liu, Elisa Warner, Barzin Nabet, Minu Srivastava, Marcus Ballinger, and Kai Liu. MOSBY enables multi-omic inference and spatial biomarker discovery from whole slide images. *Scientific Reports*, 14(1): 18271, August 2024. ISSN 2045-2322. doi: 10.1038/s41598-024-69198-6.

Amit Shakarchy, Giulia Zarfati, Adi Hazak, Reut Mealem, Karina Huk, Tamar Ziv, Ori Avinoam, and Assaf Zaritsky. Machine learning inference of continuous single-cell state transitions during myoblast differentiation and fusion. *Molecular Systems Biology*, 20(3):217–241, March 2024. ISSN 1744-4292. doi: 10.1038/s44320-024-00010-3.

Kilian Q Weinberger, John Blitzer, and Lawrence Saul. Distance Metric Learning for Large Margin Nearest Neighbor Classification. In *Advances in Neural Information Processing Systems*, volume 18. MIT Press, 2005.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A ConvNet for the 2020s. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11966–11976, June 2022. doi: 10.1109/CVPR52688.2022.01167.

Vladislav Sovrasov. Ptflops: Flops counter for neural networks in pytorch framework, 2024.

Syuan-Ming Guo, Li-Hao Yeh, Jenny Folkesson, Ivan E Ivanov, Anitha P Krishnan, Matthew G Keefe, Ezzat Hashemi, David Shin, Bryant B Chhun, Nathan H Cho, Manuel D Leonetti, May H Han, Tomasz Nowakowski, and Shalin B Mehta. Revealing architectural order with quantitative label-free imaging and deep learning. *eLife*, 9:e55502, July 2020. ISSN 2050-084X. doi: 10.7554/eLife.55502.

Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization, January 2019.

M. Jorge Cardoso, Wenqi Li, Richard Brown, Nic Ma, Eric Kerfoot, Yiheng Wang, Benjamin Murrey, Andriy Myronenko, Can Zhao, Dong Yang, Vishwesh Nath, Yufan He, Ziyue Xu, Ali Hatamizadeh, Andriy Myronenko, Wentao Zhu, Yun Liu, Mingxin Zheng, Yucheng Tang, Isaac Yang, Michael Zephyr, Behrooz Hashemian, Sachidanand Alle, Mohammad Zalbagi Darestani, Charlie Budd, Marc Modat, Tom Vercauteren, Guotai Wang, Yiwen Li, Yipeng Hu, Yunguan Fu, Benjamin Gorman, Hans Johnson, Brad Genereaux, Barbaros S. Erdal, Vikash Gupta, Andres Diaz-Pinto, Andre Dourson, Lena Maier-Hein, Paul F. Jaeger, Michael Baumgartner, Jayashree Kalpathy-Cramer, Mona Flores, Justin Kirby, Lee A. D. Cooper, Holger R. Roth, Daguang Xu, David Bericat, Ralf Floca, S. Kevin Zhou, Haris Shuaib, Keyvan Farahani, Klaus H. Maier-Hein, Stephen Aylward, Prerna Dogra, Sebastien Ourselin, and Andrew Feng. MONAI: An open-source framework for deep learning in healthcare, November 2022.

HuggingFace. Huggingface/pytorch-image-models. Hugging Face, May 2024.

Laura Antonelli, Federica Polverino, Alexandra Albu, Aroj Hada, Italia A. Asteriti, Francesca Degrassi, Giulia Guarguaglini, Lucia Maddalena, and Mario R. Guarracino. ALFI: Cell cycle phenotype annotations of label-free time-lapse imaging data from cultured human cells. *Scientific Data*, 10(1):677, October 2023. ISSN 2052-4463. doi: 10.1038/s41597-023-02540-1.

Felix Pahmeier, Christopher J. Neufeldt, Berati Cerikan, Vibhu Prasad, Costantin Pape, Vibor Laketa, Alessia Ruggieri, Ralf Bartenschlager, and Mirko Cortese. A Versatile Reporter System To Monitor Virus-Infected Cells and Its Application to Dengue Virus and SARS-CoV-2. *Journal of Virology*, 95(4):e01715–20, January 2021. ISSN 1098-5514. doi: 10.1128/JVI.01715-20.

Ziwen Liu, Eduardo Hirata-Miyasaki, Soorya Pradeep, Johanna V. Rahm, Christian Foley, Talon Chandler, Ivan E. Ivanov, Hunter O. Woosley, See-Chi Lee, Sudip Khadka, Tiger Lao, Akilandeswari Balasubramanian, Rita Marreiros, Chad Liu, Camille Januel, Manuel D. Leonetti, Ranen Aviner, Carolina Arias, Adrian Jacobo, and Shalin B. Mehta. Robust virtual staining of landmark organelles with Cytoland. *Nature Machine Intelligence*, pages 1–15, June 2025. ISSN 2522-5839. doi: 10.1038/s42256-025-01046-2. URL `https://www.nature.com/articles/s42256-025-01046-2`. Publisher: Nature Publishing Group.

Jordão Bragantini, Ilan Theodoro, Xiang Zhao, Teun APM Huijben, Eduardo Hirata-Miyasaki, Shruthi VijayKumar, Akilandeswari Balasubramanian, Tiger Lao, Richa Agrawal, and Sheng Xiao. Ultrack: pushing the limits of cell tracking across biological scales. *Nature Methods*, pages 1–14, 2025. URL `https://www.nature.com/articles/s41592-025-02778-0`. Publisher: Nature Publishing Group US New York.

Arthur Edelstein, Nenad Amodaj, Karl Hoover, Ron Vale, and Nico Stuurman. Computer Control of Microscopes Using μManager. *Current Protocols in Molecular Biology*, 92(1):14.20.1–14.20.17, 2010. ISSN 1934-3647. doi: 10.1002/0471142727.mb1420s92.

Appendix

A Model architecture, training, and metrics

A.1 Model architecture and training

The model architecture has three main components: a spatial projection stem, an encoder backbone, and a multi-layer perceptron (MLP) head. The stem begins with a convolution layer with a kernel size of $(5, 4, 4)$ for 3D datasets and $(1, 4, 4)$ for 2D datasets and a stride of $(5, 4, 4)$ for 3D datasets and $(1, 4, 4)$ for 2D datasets, followed by a reshaping operation. This reshaping maps the down-sampled axial dimension to channels, efficiently projecting the anisotropic 3D input into a 2D feature map for encoding. The encoder backbone is adapted from the ConvNeXt Tiny architecture [Liu et al., 2022]. The stem and head modules from ConvNeXt are removed, and the backbone outputs a 768-dimensional embedding vector $\mathbf{z}$. The 768-dimensional vector $\mathbf{z} \in \mathbb{R}^{768}$ is projected onto a lower 32-dimensional vector $\mathbf{p} \in \mathbb{R}^{32}$ through a 2-layer MLP head, which helps speed up training [Chen et al., 2020]. To estimate the computational cost, we profiled the full DynaCLR encoder using the ptflops package [Sovrasov, 2024]. For input patches of size 2×15×256×256, the forward pass requires approximately 754 GFLOPs. Each training step, including the the gradient computations, is estimated at 2.26 TFLOPs. The total training cost per model is approximately 0.5–1 PFLOPs for 100K iterations.

A.2 Sampling and augmentation of patches of single cells

A.2.1 Classical sampling

In the classical contrastive setting, temporal or cell identity information is not used. The anchor $\mathcal{A}_1[\mathbf{x}_i]$ and positive pair $\mathcal{A}_2[\mathbf{x}_i]$ are created through random augmentations $\mathcal{A}$ of the same cell at a given time point, while negatives are augmented views of random cells $\mathcal{A}_3[\mathbf{x}_j]$ sampled from any other time point.

A.2.2 Cell- and time-aware sampling

This strategy uses tracking to define positives from the same cell at consecutive time points. Specifically, images of the same tracked cell at t and $t + \tau$ form positive pairs $\{(\mathcal{A}_1[\mathbf{x}_i(t)], \mathcal{A}_2[\mathbf{x}_i(t + \tau)])\}$, while an image of a different cell $\mathcal{A}_3[\mathbf{x}_j(t + \tau)]$ at time $t + \tau$ serves as a negative.

The pretext task minimizes the embedding distance between temporally adjacent views of the same cell and maximizes the distance to embeddings of other cells. For explicit negative sampling, the triplet loss is computed over the batch

$$\mathcal{B}_{\text{triplet}} = \{(\mathcal{A}_1[\mathbf{x}_i(t)], \mathcal{A}_2[\mathbf{x}_i(t + \tau)], \mathcal{A}_3[\mathbf{x}_j(t + \tau)]) \mid i \neq j\},$$

while the NT-Xent loss is computed over positive pairs

$$\mathcal{B}_{\text{NT-Xent}} = \{(\mathcal{A}_1[\mathbf{x}_i(t)], \mathcal{A}_2[\mathbf{x}_i(t + \tau)])\}.$$

The time offset τ is selected according to the temporal resolution of the dataset, typically adjacent frames. The time offset τ is a hyperparameter empirically chosen based on the time scales of the dynamic process and the time resolution of imaging. For this paper, the positive pair is sampled from adjacent frames.

A.2.3 Data sampling

3D imaging volumes are cropped around the centroids of the tracking nodes to form single-cell patches. We normalize the input image to reduce variability from experimental conditions. We rescale the viral sensor channel so that the median intensity is 0, and the 99th percentile intensity is 1. This normalization is more robust to extreme intensities in the fluorescence image, as well as variation in background fluorescence levels. The quantitative phase channel is normalized so that each field-of-view (FOV) has zero mean and unit standard deviation. The phase image is already normalized during reconstruction [Guo et al., 2020], and this extra standardization step ensures proper input numerical range for the model. We use a larger initial crop to ensure no padding is included in the final input patch after spatial augmentations. We apply extensive augmentations (Table 1) at training time to simulate variations induced by the imaging system and other non-biological conditions. The input patch size after augmentations is optimized for reducing the influence from background and neighboring cells while focusing on the peri-nuclear region of the cell, where the majority of infection-related changes, such as viral sensor re-localization and ER remodeling, are captured.

The models summarized in Table 2 were trained with a mini-batch size of 256, using the AdamW optimizer [Loshchilov and Hutter, 2019], and a learning rate of 2×10^{-5}. We used the HPC cluster on-premises using 2-4 GPUs with the distributed data parallel (DDP) strategy. A temperature of 0.3 was optimized for ALFI models to prevent overfitting with NT-Xent loss, as it is a small dataset. Other models with NT-Xent loss were trained with a temperature of 0.5. The margin of 0.5 was used in computational experiments that used the triplet loss. The time for model training depends on the size of the dataset, varying from an hour for the cell cycle model with the ALFI dataset to around 48 hours for the infection and organelle remodeling models.

Table 1: Augmentations applied to image patches. Parameters are supplied to respective MONAI [Cardoso et al., 2022] transforms, where α denotes scaling factor, θ denotes rotation (radians), s denotes shearing, γ denotes gamma value, σ denotes the standard deviation of the Gaussian distribution, and p denotes the probability of applying the random transform.

Augmentation Type	**Parameters**
Random spatial scaling	$\alpha_x, \alpha_y \in [-0.3, 0.3], p = 0.8$
Random rotation	$\theta_z \in [0, \pi], p = 0.8$
Random shearing	$s_x, s_y \in [0, 0.01], p = 1.0$
Random contrast adjustment	$\gamma \in [0.8, 1.2], p = 0.5$
Random intensity scaling	$\alpha \in [-0.5, 0.5]$, $p_{\text{Phase}} = 0.5$, $p_{\text{Sensor}} = 0.7$
Gaussian smoothing	$\sigma_x, \sigma_y \in [0.25, 0.75], p = 1.0$
Gaussian noise addition	$\sigma_{\text{Phase}} \in [0, 0.2]$, $\sigma_{\text{Sensor}} \in [0, 0.5], p = 0.5$

A.3 Model library

The paper reports multiple models (DynaCLR-*) trained with three datasets summarized in Appendix B, depending on the biological prediction task. We have organized the models based on the training and test data (Table 2). The DynaCLR-DENV-* models were trained using time-lapse data acquired with 30 min interval between frames and using time-lapse datasets acquired with 10 min interval. The patch sizes and z-ranges used for different models are listed in Table 3. We use the ImageNet [HuggingFace, 2024] and OpenPhenom-S [Kraus et al., 2023] pretrained models as natural vision baselines (Figure S3, Table 4).

Table 2: **Summary of models:** DynaCLR models organized by training and test data

Model name	**Training data**	**Test data**	**Results shown in**
DynaCLR-ALFI	U2OS (from ALFI)	HeLa + RPE1 (from ALFI)	Fig. S1a,
DynaCLR-microglia	Microglia (IL-17, IF-β)	Microglia (glioblastoma)	Fig. S1b
DynaCLR-DENV-VS+Ph	Phase+Viral Sensor	Phase+Viral sensor	Fig. 2

Table 3: **Model input specifications:** The table provides a summary of input specifications for DynaCLR models. The input channels, patch size, and z-range for models are listed.

Model name	**Input channels**	**Patch size (YX)**	**Z range**
DynaCLR-ALFI	DIC	128×128	[0–1]
DynaCLR-microglia	Phase	96×96	[0–1]
DynaCLR-DENV-VS+Ph	Phase + sensor	160×160	[15–45]

A.4 Metrics

To characterize the temporal continuity and variability of a tracked cell i in the embedding space, we analyze its trajectory $\mathbf{z}_i(t) \in \mathbb{R}^d$ via cosine distance. We measure the distance between the embeddings of two cells via cosine distance:

$$D_{ij}(t_a, t_b) = 1 - \frac{\mathbf{z}_i(t_a) \cdot \mathbf{z}_j(t_b)}{\|\mathbf{z}_i(t_a)\| \cdot \|\mathbf{z}_j(t_b)\|}, \quad \text{for } t = 1, \ldots, T-1. \tag{1}$$

We assess the effect of the contrastive sampling method and the loss functions on the temporally regularized embedding space using the pairwise cosine distance between random and adjacent timepoints t with the following metrics:

Smoothness: Smoothness quantifies how much short-term variation exists relative to overall variation in the embedding space. We compute the ratio of the mean distance between adjacent timepoints in each trajectory to the mean distance between randomly sampled timepoints from the same trajectory:

$$\text{Smoothness} = \frac{D_{adj}}{D_{rand}} = \frac{\text{mean}_{i,\,t}\,[D_i(t, t+1)]}{\text{mean}_{i,\,(t_a,t_b)}\,[D_i(t_a, t_b)]} \tag{2}$$

where $D_i(t_a, t_b)$ is the cosine distance between embeddings at a randomly sampled pair of embeddings at timepoints t_a and t_b, and $t_a \neq t_b$. A lower value indicates a temporally smooth embedding space.

Dynamic Range (DR): The dynamic range quantifies how much variation is captured in the embedding space over time. It is defined as the difference between the peaks of the embedding distance distributions for randomly selected frame pairs and adjacent frame pairs, computed over all tracks in the datasets. The peaks were identified using Gaussian KDE.

Table 4: **Performance of DynaCLR-DENV-VS+Ph models:** F1 score of linear classification of pairwise distance of adjacent and random frames, dynamic range, and smoothness for infection state classification for different losses, and sampling strategies. For OpenPhenom and ImageNet, only the viral sensor channel was used as input.

Experiments	F1 score ↑	Smoothness ↓	DR ↑
Time Aware + NT-Xent	**98.40**	**0.15**	**1.33**
Time Aware + triplet	96.56	0.16	1.07
Cell aware + triplet	98.24	0.24	1.07
Classical + NT-Xent	**98.41**	0.32	1.23
Classical + triplet	98.07	0.23	1.03
ImageNet pretrained	97.82	0.47	0.74
OpenPhenome-S/16	95.2	0.32	1.18
Supervised semantic segmentation model	83	-	-

B Datasets

B.1 Data and annotations

We explore the performance and applications of DynaCLR with three distinct time-lapse datasets: (1) a 5D dataset representing both infection and cell cycle dynamics, (2) a previously published 2D dataset capturing cell cycle dynamics [Antonelli et al., 2023], (3) a previously published 2D dataset of perturbed microglia [Wu et al., 2022].

B.2 Infected cells: 3D label-free and fluorescence movies of Dengue-infected A549 cells

We used 5D time-lapse datasets of A549 cells infected with live Dengue virus to evaluate DynaCLR's ability to disambiguate dynamic morphological states. The data were acquired using spinning disk confocal and light-sheet microscopes at two temporal resolutions (10 min and 30 min), under both mock (no virus, MOI 0) and infected (MOI 5) conditions. Each movie included a quantitative phase channel and a fluorescence channel encoding infection via a genetically engineered mCherry-NLS sensor [Pahmeier et al., 2021]. A549 cells with an ER marker, SEC61, infected with Dengue virus, were used to develop the methods for analyzing the organelle remodeling due to the infection.

Cells were segmented using virtual staining [Liu et al., 2025] and tracked with Ultrack [Bragantini et al., 2025].

B.2.1 Image acquisition and processing

We acquired 5D image datasets (time series of 3D volumes with phase and fluorescence channels) of A549 cells infected with Dengue virus at an MOI of 0 and 5, using:

- A spinning disk confocal microscope with 30 min temporal resolution and 0.25 um z-resolution and
- A light-sheet microscope with 10-minute and 30-minute temporal resolutions and 0.7 um z-resolution.

Mock wells served as controls. Imaging was performed for up to 24 hours in multi-well plates. Image acquisition was automated using `Micro-Manager` [Edelstein et al., 2010], and the resulting OME-TIFF files were converted to OME-Zarr using `iohub` for scalable I/O and downstream processing.

Phase images were reconstructed from brightfield z-stacks using Köhler illumination [Guo et al., 2020], and normalized per field-of-view to zero mean and unit variance. Fluorescence images were normalized per field-of-view centered around the median and scaled to range between the 50th and 99th percentile, effectively centering the background at zero while preserving signal dynamic range.

Virtual staining of cell nuclei was performed using a deep learning model [Liu et al., 2025], followed by segmentation and tracking via Ultrack [Bragantini et al., 2025]. 3D image patches of single cells were cropped based on track centroids. We applied extensive augmentations at training time (Table 3) to simulate imaging variability and improve generalization.

Together, these steps encoded the intrinsic perturbations of the cell cycle, the extrinsic perturbations of the infection cycle, and the response of organelles in a dataset of movies suitable for DynaCLR model training.

B.2.2 Annotations of infection and cell division

The cell division and infection states were identified by point annotations placed on the nuclei of cells. The points were matched to embeddings by assigning the points to centroids of closest nuclei. We validated the annotations and predictions by overlaying them on the projected embeddings. We also tested the model on independent test data to assess its generalization to new data.

Cell division is captured from cell tracking by Ultrack [Bragantini et al., 2025] and revised manually. The cell division is indicated by a parent track splitting into two daughter tracks with the same parent track IDs. The last time-point of the parent track is considered the division event. The human annotator proofread and corrected the cell division events through visual inspection of the tracks in Ultrack GUI.

B.3 ALFI: 2D label-free movies with cell cycle annotations

We used label-free Differential Interference Contrast (DIC) microscopy movies of three cell types (HeLa, RPE1, and U2OS) from the ALFI dataset by Antonelli et al. [2023]. In this dataset, bounding boxes of a subset of cells were tracked and annotated by human experts, with each time point labeled according to the corresponding cell cycle stage (mitosis or interphase). We evaluate the ability of DynaCLR models to discriminate between mitosis and interphase cell cycle stages, which generalizes to unseen cell types. The models were trained with perturbed and unperturbed U2OS cells and tested on unperturbed HeLa and RPE1 cells. All movies were acquired in 2D with a time resolution of 7 min.

B.4 Microglia: 2D label-free movies of pharmacologically perturbed human microglia

This dataset consists of label-free movies of human microglia cells subjected to pharmacological modulators of immune activity (IL-17, IF-β, extract of brain tumor glioblastoma) acquired with quantitative phase imaging (QPI) that was used to develop Dynamorph models [Wu et al., 2022]. Dynamorph method used temporally-regularized VQ-VAE and provided a useful baseline to evaluate the generalization, smoothness, and dynamic range of DynaCLR embeddings. For training, we selected movies of cells treated with IL-17, IF-β, and untreated control conditions, while the glioblastoma supernatant-treated condition was held out for testing.

Each condition contains nine non-overlapping fields of view (FOV), each containing approximately 250 cells. The raw data were acquired with 1 μm z-steps every 9 minutes using a Leica DMI-8

microscope with a 20× objective using QLIPP method [Guo et al., 2020]. Although the dataset contained bright-field, quantitative phase, and retardance channels, we only used the quantitative phase channel to perform virtual staining of nuclei with CytoLand [Liu et al., 2025] and joint segmentation and tracking with Ultrack [Bragantini et al., 2025], and subsequent model evaluation.

C Supplementary Figures

C.1 Embedding cell division and cell migration dynamics

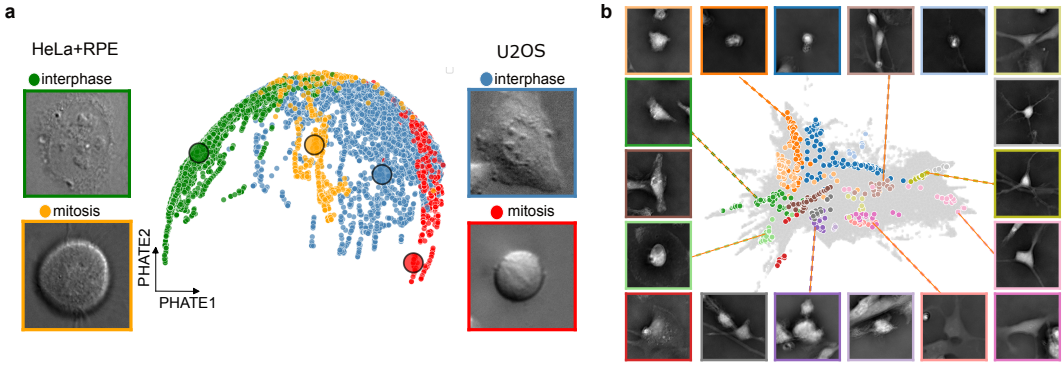


Figure S1 **Temporally regularized embeddings of cell division and cell migration:** (a) Training (U2OS cells) and test (HeLa + RPE1 cells) sets jointly embedded in a PHATE map to show the clustering of cells based on cell cycle state and cell size. (b) PHATE map of microglia morphotypes from a brain tumor environment (glioblastoma). Color-coded tracks with sampled cell images illustrate consistent embedding of similar morphologies.

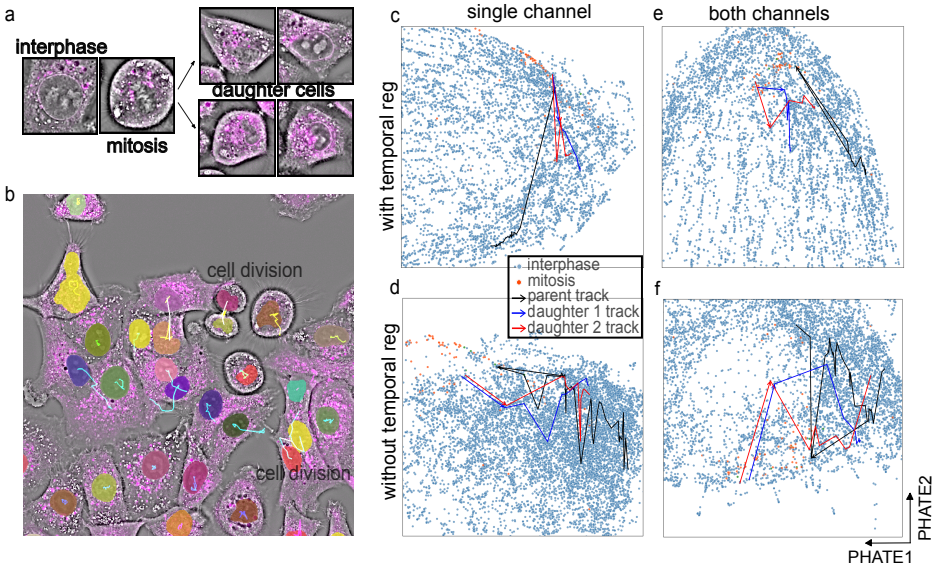


Figure S2 **Detection of rare events, e.g., cell division**:(a) The morphology of the cell changes over time during the transition between interphase and mitosis. (b) Ultrack tracks the cell over time and captures mitosis. White tracks indicate cell divisions. (c–f) The trajectory of one parent cell (black track) dividing into two daughter cells (blue and red tracks) overlaid on the PHATE from models using phase channel (single channel) and a combination of phase and viral sensor channels (both channels), and with and without temporal regularization, illustrates that temporal regularization leads to smooth trajectories and better clustering with just the phase channel.

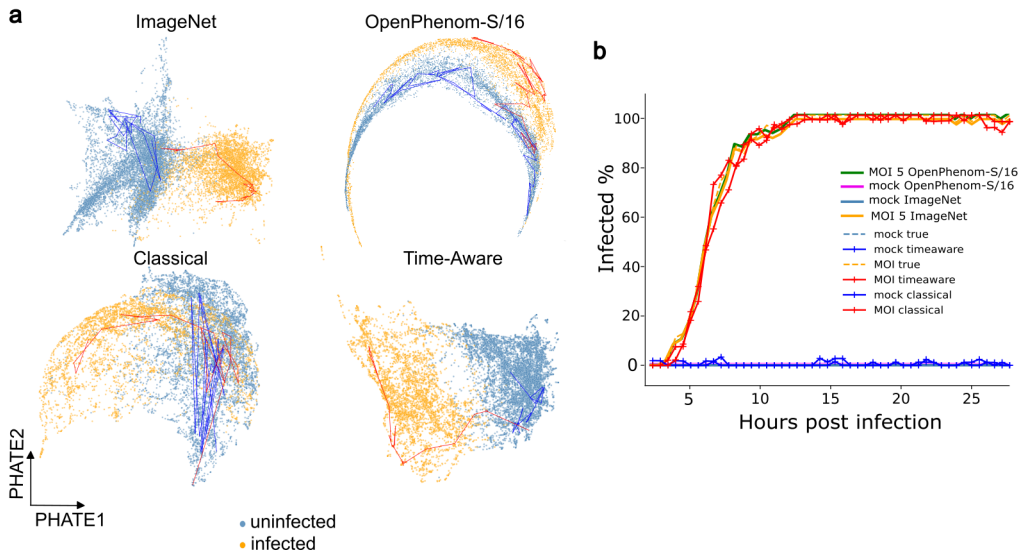


Figure S3 **DynaCLR classical and time-aware models perform comparably to pre-trained at infection state classification** a) PHATE embeddings reveals clear clustering of infected and uninfected cells based on features from the ImageNet pre-trained model, OpenPhenom-S/16 pre-trained model, DynaCLR Classical and Time-aware models using a linear classifier trained with sparse annotations. b) Infection percentage over time shows an exponential increase, consistent with expected infection dynamics. OpenPhenom and ImageNet models take only the viral sensor channel as input.

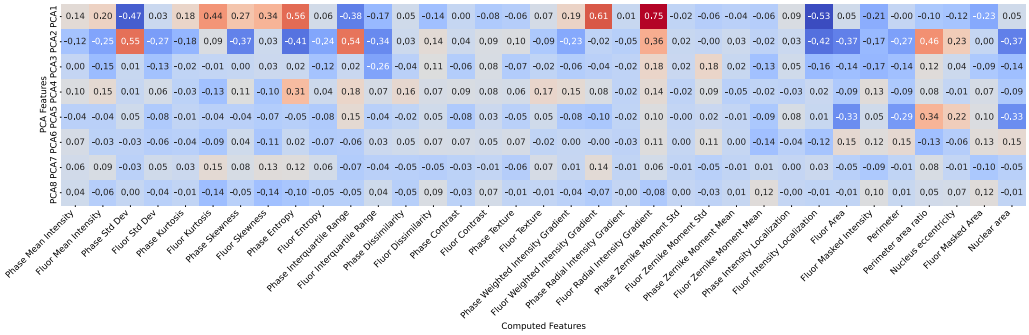


Figure S4 **Principal components vs computed features for the viral sensor and phase model**: Principal components correlate with interpretable image features such as radial intensity profile, area of fluorescence, and phase texture statistics, suggesting that the model captures biologically relevant variation.

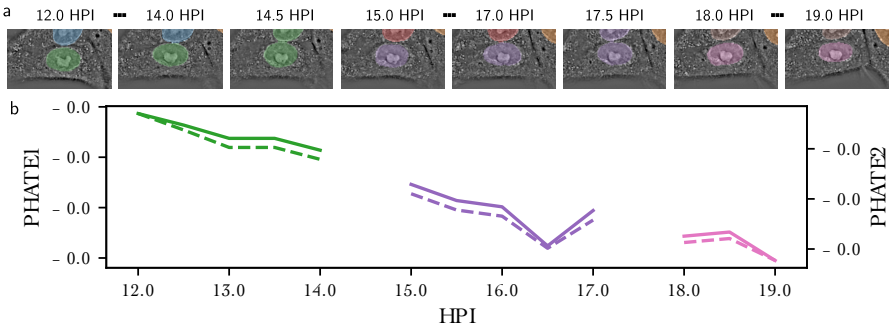


Figure S5 **DynaCLR cell and time-aware embeddings are smooth even when tracking is erroneous:** (a) snapshots of a cell and its tracking labels over time. Note that the false fusion in 14.5 and 17.5 HPI frames caused subsequent false division and identity jump of the cell. (b) PHATE components 1 (solid line) and 2 (dashed line) over time for the falsely assigned tracks. The gaps correspond to false fusion events which shifts the centroid of the track towards the edge of the FOV, resulting in invalid patches. The PHATE components are smoothly transitioning over time, even though they are assigned to different tracks.

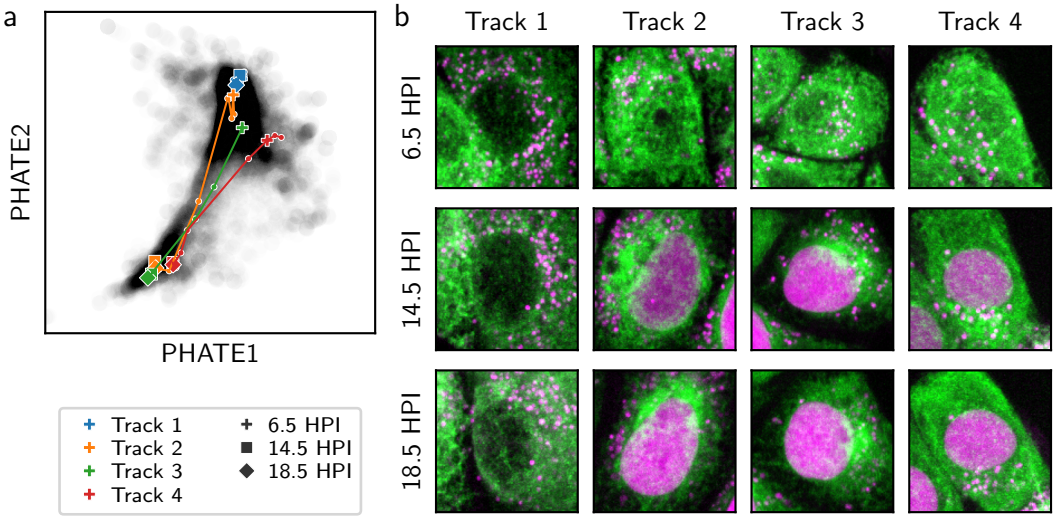
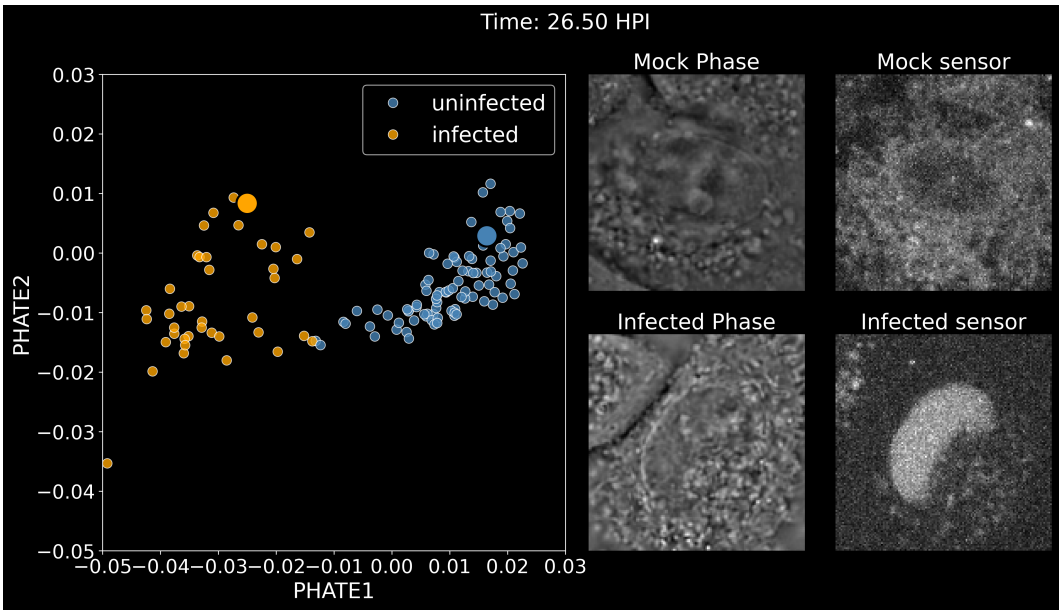


Figure S6 **Learned representation of the phase and viral sensor channels help exploration of organelle remodeling during infection.** (a) PHATE of learned features computed for mock and Dengue infected cells in the independent test dataset where the ER of cells is labeled with a fluorescent protein (SEC61-GFP). Track 1 from the mock well and tracks 2-4 from the Dengue infected well are highlighted. Cells other than the example tracks are marked in gray. (b) Snapshots from example tracks in (a), showing max-intensity projection of ER (green) and the viral sensor (magenta). In some of the infected cells (tracks 2 and 3), ER forms transient condensation.

D Videos



Video 1: **Infection dynamics in DynaCLR embedding space**: Evolving dynamics of infection in unseen test data with time from a different microscope, colored by model prediction. Images show representative cells from mock and MOI 5 infected conditions.

Data and code availability

The model architecture, training, and prediction code for the DynaCLR method is available at `https://github.com/mehta-lab/viscy`. The napari plugin for visualization of data, tracking results, embedding predictions, and performing human annotation is available at `https://github.com/czbiohub-sf/napari-iohub`. VisCy is built on PyTorch Lightning, MONAI libraries, and OME-Zarr data format. We used to convert image data into OME-Zarr format and to load data for training and inference. We used the development version of https://github.com/royerlab/ultrack for single-cell tracking. We used reconstruction algorithms of `https://github.com/mehta-lab/waveorder` to compute 3D phase from 3D brightfield volumes.

Acknowledgments and Disclosure of Funding

We thank Talon Chandler, CZ Biohub SF, and Sandra Schmid, CZ Biohub SF, for critical feedback on the manuscript. The Chan Zuckerberg Initiative funded this research through the Chan Zuckerberg Biohub, San Francisco. All authors are supported by the intramural program of the Chan Zuckerberg Biohub, San Francisco.