Automated, Multiparameter Cell Cycle Analysis

Mark Colasurdo[1], Rachael Sheridan[2], John Quinn[1], Josef Spidlen[1]

[1]BD Biosciences; [2]Van Andel Institute

Abstract

Introduction

- Traditional cell cycle analysis in flow cytometry has relied on single-parameter data and indeterminant modeling.
- Multiparameter analysis enables more accurate detection and discrete classification of phases.
- Machine learning can be used to automate this process, providing consistent, objective, and rapid results.

Methods

- 17 datasets: 7 3D (DAPI, EdU, PH3) and 10 time-series (DAPI, PE)
- Hierarchical Density Based Scanning in Applicstions with Noise (HDBSCAN) algorithm used to cluster multiscale clusters.
- Python, Scikit-Learn, Numpy, and Pandas

Results

- Algorithm identifies G1, S, G2M, G2, and M phases with >99% precision.
- Classification recall is a bit lower, >80%, due to classification of data as outliers.

Conclusion

- HDBSCAN provides a robust basis for automated cell cycle classification.

Methods

Data

- **2D/3D Dataset –** 7 independent experiences containing cell cycle data using DAPI (DNA marker), EdU (proliferation marker), and PH3 (mitosis marker).
- **Time-series Dataset –** 2 independent experiments with 5 timepoints each at 1, 4, 8, 16 and 24 hours. Used DAPI (DNA marker) and PE (proliferation marker).
- The time-series dataset was used to develop the algorithm while the 2D/3D datasets were used as the test dataset.
- Ground truth was established by expert-gating in BD FlowJo™.

Algorithm

- HDBSCAN, a multi-scale, density-based clustering algorithm, is used to cluster cell cycle populations.
- Data is first preprocessed by log transforming proliferation data and then centering and scaling the data to unit variance.
- Minimum cluster size, a key parameter in HDBSCAN, is dynamically determined based on the size of the dataset, n:

$$\text{Minimum Cluster Size} = \sqrt{n}$$

- After clustering, the clusters are then classified based on their position relative to the centroid of all the data.

Statistics

- Precision and Recall were calculated for classes G1, G2, M, G2M, and S.
- Precision represents the correctness of labels predicted as the target class.
- Recall represents the proportion of the target class that was captured by the prediction.

$$Precision = \frac{TP}{TP + FP} \qquad Recall = \frac{TP}{TP + FN}$$
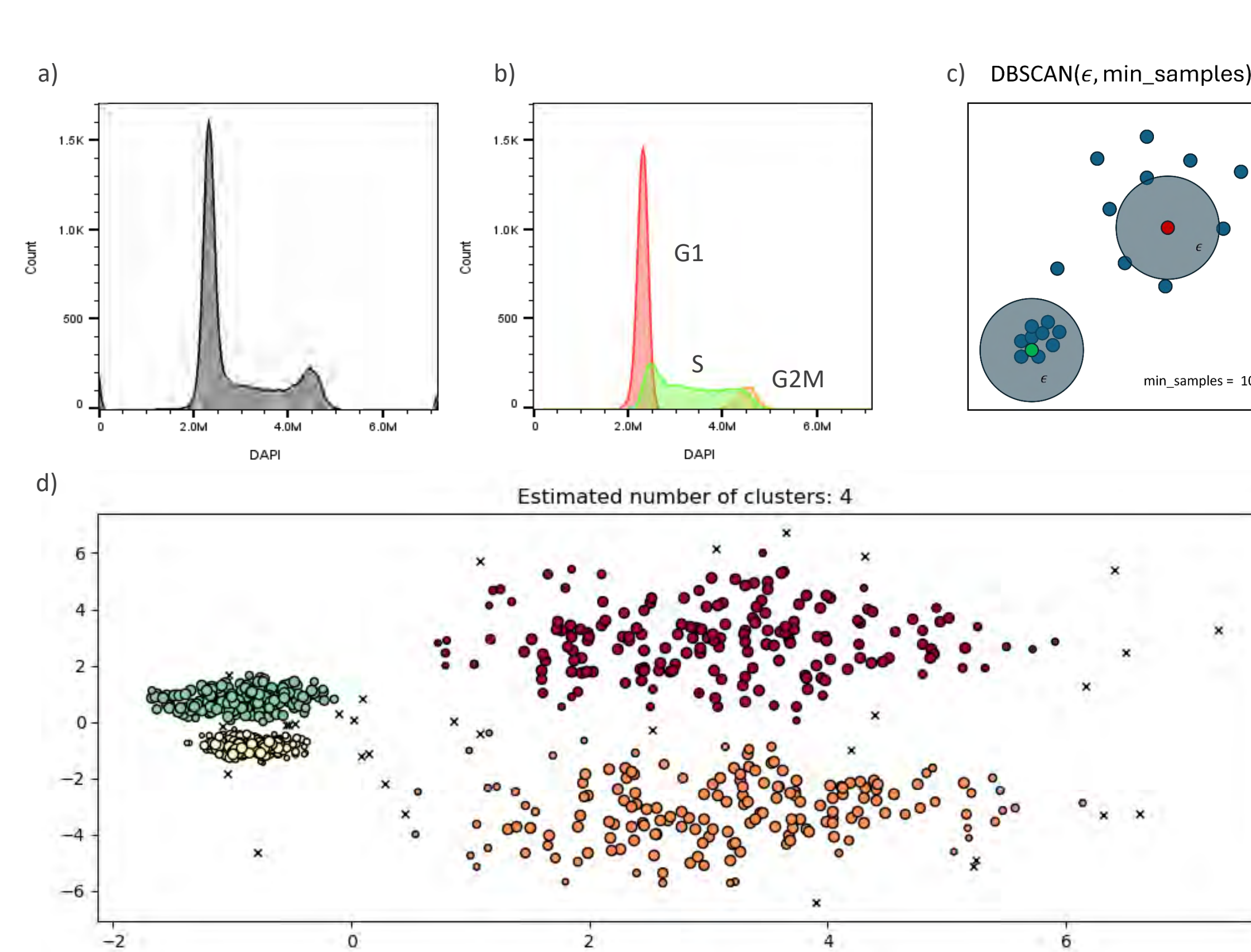
Results: Development and 2D/3D Data

Figure 1 – 1D Cell Cycle Analysis and Potential for HDBSCAN
a) Histogram of DNA content, typically used for 1D cell cycle analysis with DAPI on the x-axis. **b)** Histogram of DNA content overlaid with actual G1, S, and G2M phases, demonstrating overlapping regions and indeterminant nature of single-parameter analysis. **c)** Demonstration of DBSCAN, the underlying clustering algorithm used in HDBSCAN. Clustering is performed by specifying a minimum number samples, min_samples, within a certain distance, eps. **d)** Demonstration of HDBSCAN, which is able to identify outliers while otherwise clustering multiscale data structures.
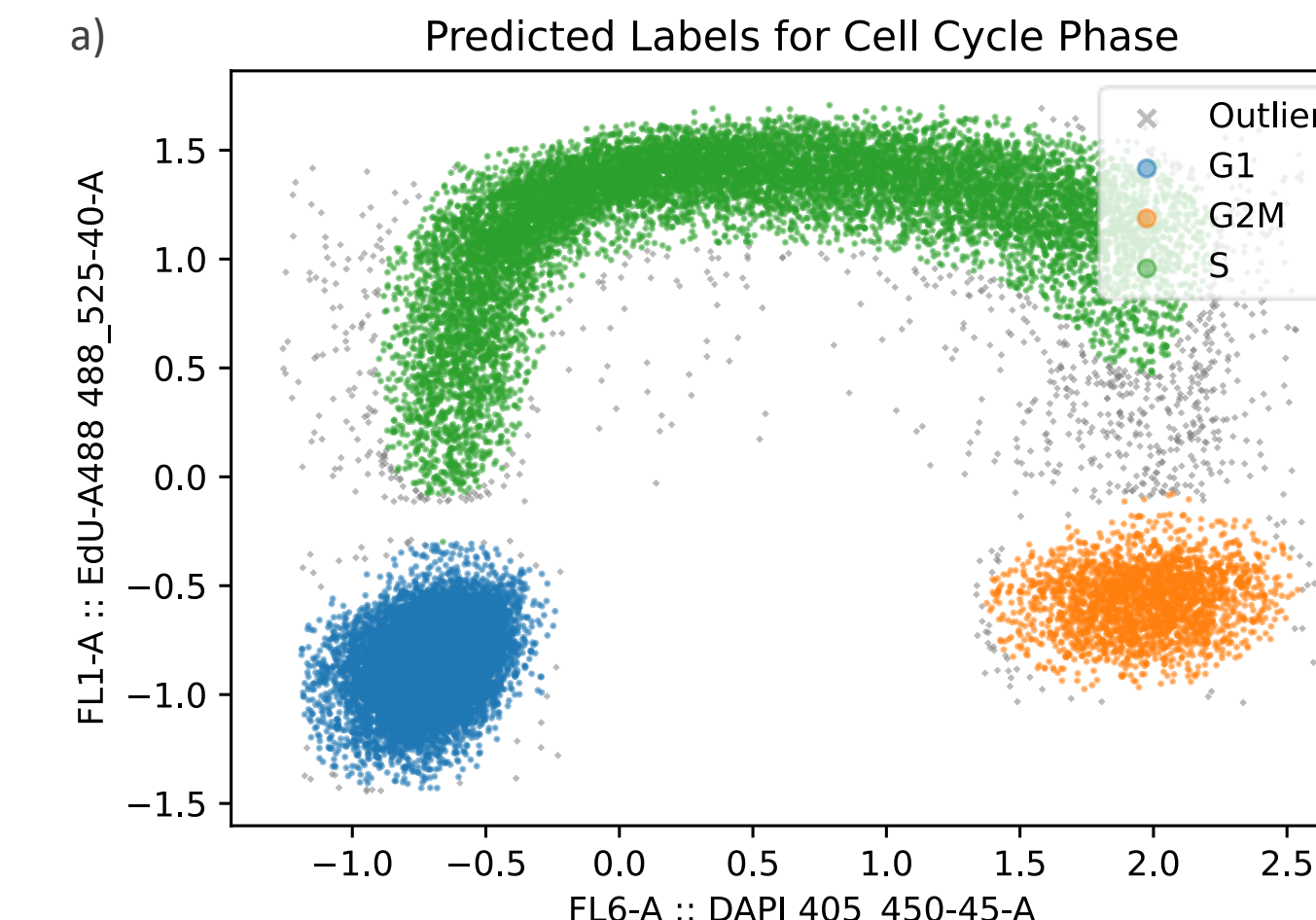
Figure 2 – Demonstration of 2D and 3D Classification using HDBSCAN
a) 2D classification using DAPI and EdU which allow for detection of proliferation. **b)**3D classification using DAPI, EdU, and PH3, which allow for detection of proliferation and mitosis.
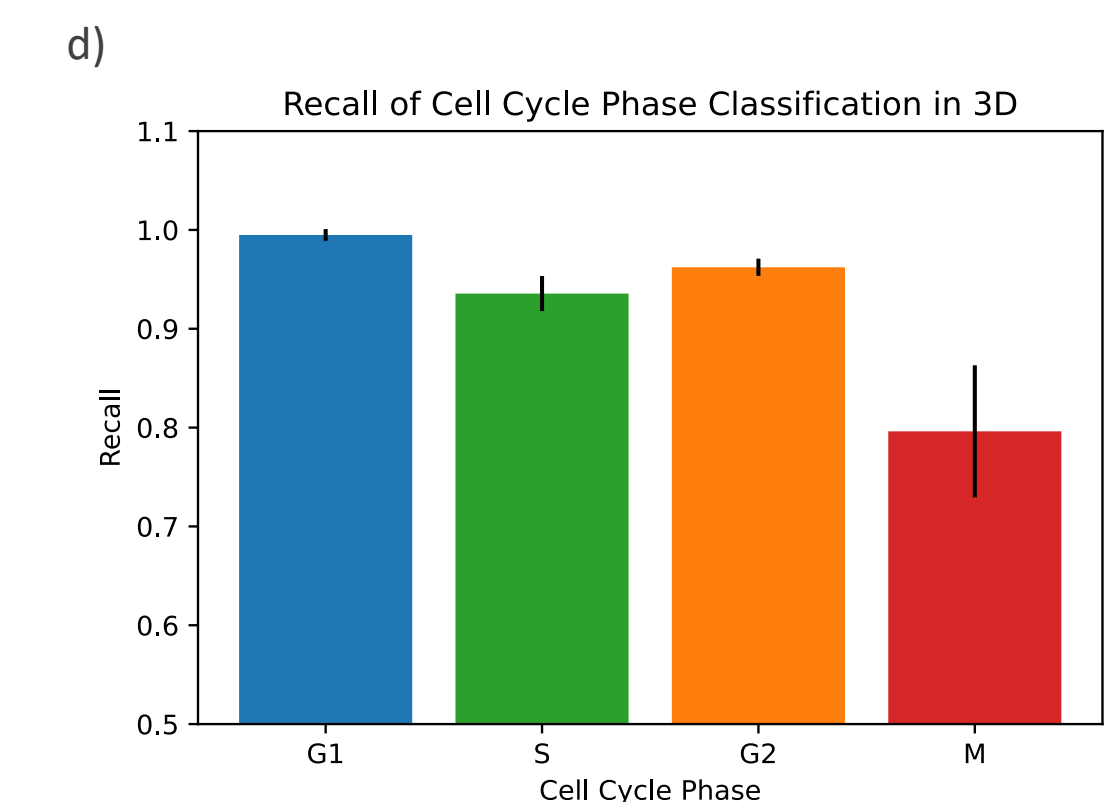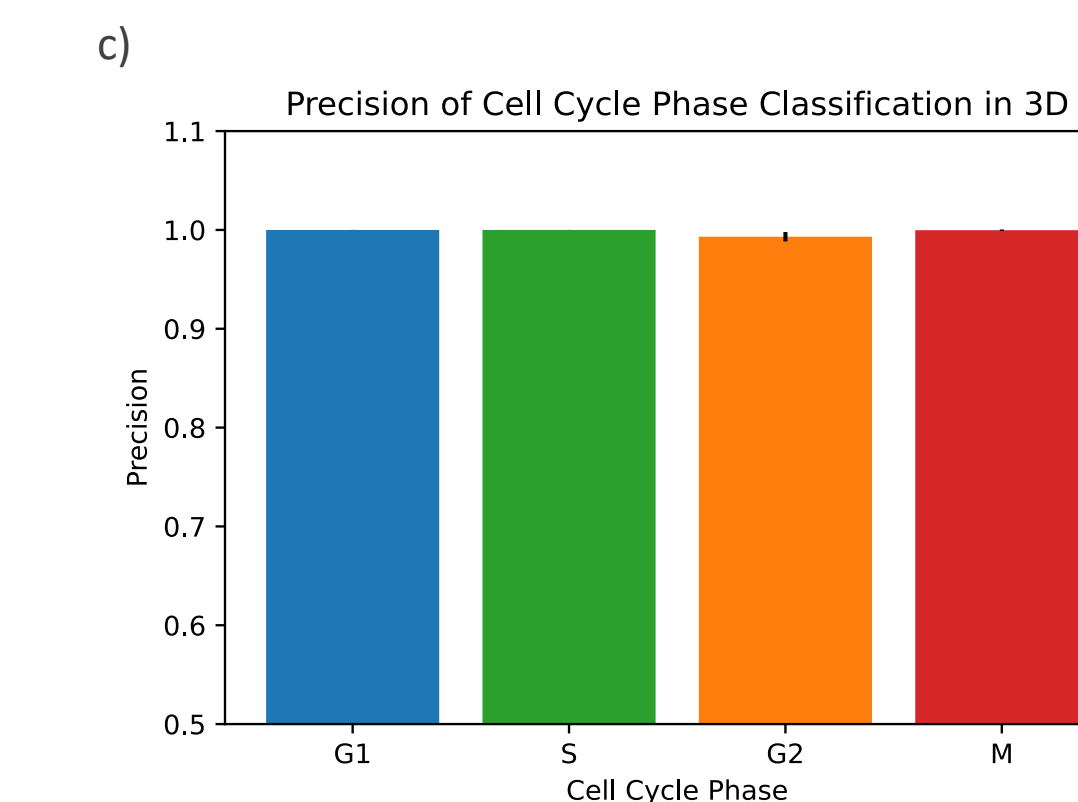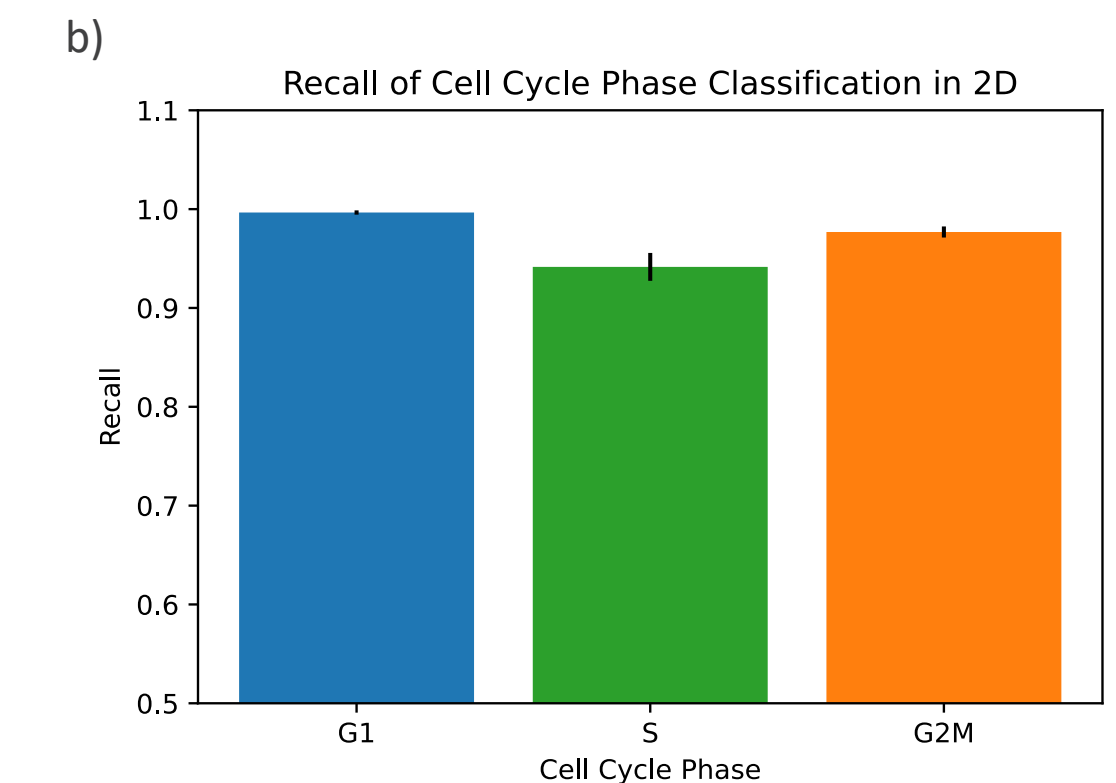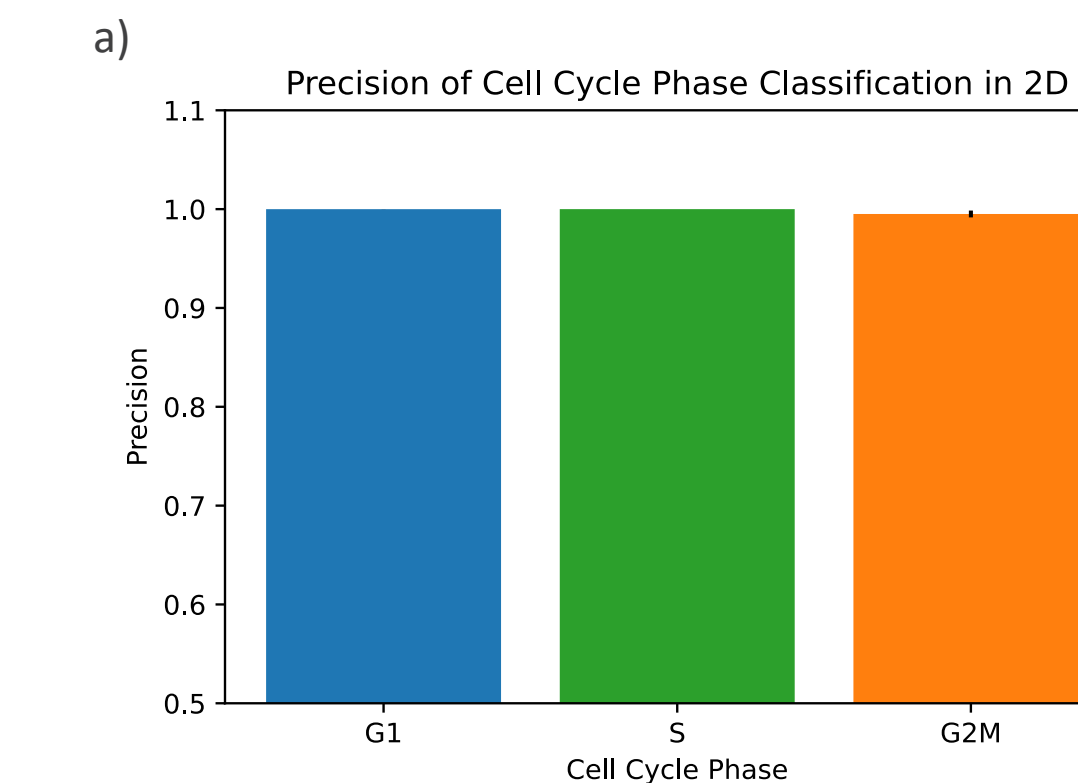
Figure 1 – 1D Cell Cycle Analysis and Potential for HDBSCAN
a) Histogram of DNA content, typically used for 1D cell cycle analysis with DAPI on the x-axis. **b)** Histogram of DNA content overlaid with actual G1, S, and G2M phases, demonstrating overlapping regions and indeterminant nature of single-parameter analysis. **c)** Demonstration of DBSCAN, the underlying clustering algorithm used in HDBSCAN. Clustering is performed by specifying a minimum number samples, min_samples, within a certain distance, eps. **d)** Demonstration of HDBSCAN, which is able to identify outliers while otherwise clustering multiscale data structures.
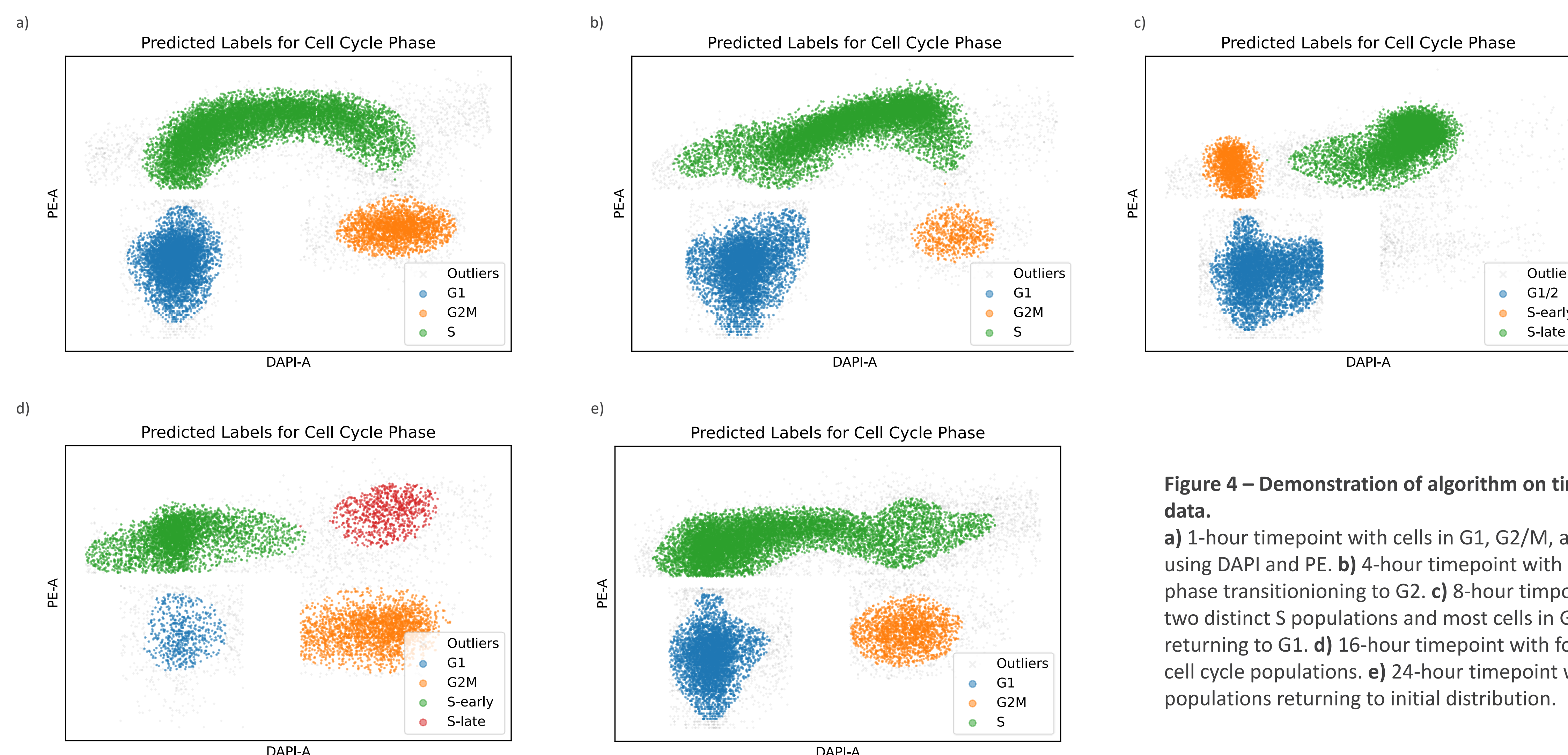
Results: Time-series Data

Figure 4 – Demonstration of algorithm on time-series data.
a) 1-hour timepoint with cells in G1, G2/M, and S-phase, using DAPI and PE. **b)** 4-hour timepoint with cells in S-phase transitionioning to G2. **c)** 8-hour timepoint with two distinct S populations and most cells in G2/M returning to G1. **d)** 16-hour timepoint with four distinct cell cycle populations. **e)** 24-hour timepoint with populations returning to initial distribution.

Conclusions

- We have developed an algorithm for the classification of multi-dimensional cell cycle data into various cell cycle phases.
- Algorithm provides >99% precision and >80% recall according to test data.
- Demonstrated applicability for 2D, 3D, and time-series data.
- May improve speed and reproducibility when analyzing data, especially for new users.
- The algorithm may lay the groundwork for new cell cycle analysis tools within FlowJo™.