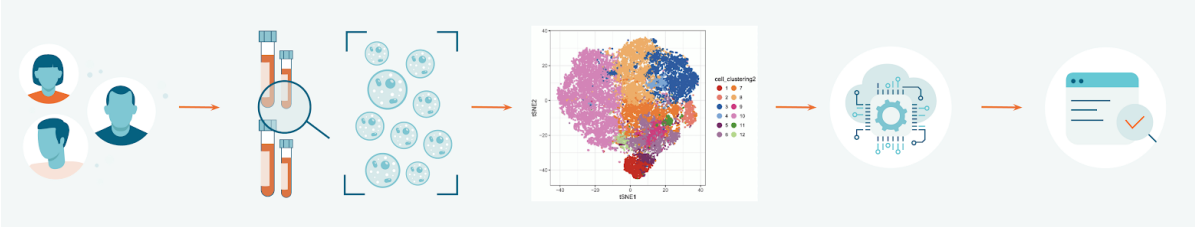
Scailyte's ScaiVision performs best-in-class at sample class prediction



472 patients' blood samples

9 cohorts of single cell samples

Mass Cytometric (CyTOF) analysis

ScaiVision data integration & analysis

Clinically - relevant outcome prediction

Introduction

Named Nature "Method of the year" in 2013, single-cell transcriptome and genome sequencing have paved the way for novel fundamental discoveries as well as the characterization of previously unknown cell types and subtypes in normal and diseased tissues[1]. Since then, a rapid development of technologies has improved our ability to profile different aspects of cells, including the epigenome and the proteome at a single-cell resolution[2,3].

While huge strides have been made in identifying and describing cell types across human tissues, the next major frontier in single-cell analysis is the prediction of disease state in individual patients through a precise identification of the molecular changes that cause cells to deviate from normal physiological trajectories[4]. Such an ability is crucial for the development of true precision medicine. Nonetheless, the analysis of such large and complex data presents major challenges. Single-cell datasets can be extremely rich, containing measurements of thousands of parameters from millions of individual cells. Moreover, samples from different patients can be highly variable, which may confound the detection of true disease-level patterns. Finally, the overwhelming scale of data may result in only a fraction of this insightful resource being utilized and translated into clinically-relevant knowledge and applications.

To solve this problem, Scailyte has developed ScaiVision, an Artificial Intelligence (AI) based software that offers a best-in-class performance for automated pattern-recognition and interpretation of single-cell data. Originally created by Prof. Manfred Claassen and colleagues at the ETH Zürich as CellCnn[5], ScaiVision has been further developed by Scailyte to integrate high-parameter single-cell data together with clinical data for the identification of disease-specific biomarkers and, therefore, prediction of disease outcomes. ScaiVision unravels the full potential of single-cell datasets, such as mass cytometry (CyTOF) and single-cell transcriptomics data. Based on a convolutional neural network, ScaiVision automatically learns molecular profiles corresponding to individual cells or cell populations whose presence or frequency is associated with a phenotype of interest.[5]

In order to train such a network, ScaiVision requires input datasets consisting of single-cell samples from multiple patients together with information about the clinical endpoint or outcome of each patient. Through this training process, ScaiVision identifies signals that are common across individuals but distinct to a certain disease state or condition. The results of a ScaiVision analysis consist of both predictions of the clinical endpoint for new, unknown patient samples, as well as the molecular characterization of the most-relevant cells associated with that endpoint. Possible clinical applications include disease diagnosis and staging, prognosis, patient stratification by treatment response or prediction of adverse events, as well as in-depth discovery of cellular mechanisms underlying disease states or drug activity.

Comparison of ScaiVision against current methods

ScaiVision presents a number of advantages over currently available methods and algorithms for single-cell data analysis. In comparison to methods relying on a clustering algorithm as a first step, such as FlowSOM[6], ScaiVision is entirely agnostic to cell clusters or pre-determined cell types. This allows the analysis to truly operate on the single-cell level, considering each cell's contribution independently, which leads to a greater potential for discovery of novel results as well as a higher resolution and sensitivity for detecting rare cells. In addition, the need for the user to pre-specify parameters such as the number of clusters to detect, which has been shown to have a significant impact on clustering results[7], is removed.

Two of the methods benchmarked in the current study, CytoDx[8] and deepCNN[9], also circumvent the need for a clustering algorithm to be applied as a first step. However, both of these methods rely on sub-sampling cells to speed up runtime, rather than utilizing the entire dataset. In contrast, ScaiVision operates in a scalable manner to analyze datasets of up to hundreds of millions of cells without any sub-sampling. ScaiVision also retains this single-cell resolution all the way through the interpretation stage, as it calculates a score for every cell in the original dataset indicating how strongly associated it is with the given clinical endpoint. This is in contrast to other methods that rely on clustering of results or decision trees to identify cellular populations of interest.

Given these advantages, we set out to determine whether ScaiVision could also outperform other methods in terms of predictive accuracy in a sample classification task. To do so, we used a massive, publicly available CyTOF dataset containing more than 121 million cells, with the goal of predicting latent CMV infection status in donors.

Methods

Following a recent study, FCS files corresponding to 472 PBMC samples split across nine different immunology cohorts were downloaded, along with metadata indicating whether each sample was taken from a patient positive or negative for latent CMV infection[9]. CMV status was used as a label for training and prediction. One cohort was randomly chosen to be used for validation and one cohort was randomly chosen to be used for testing, with the seven remaining cohorts used for training. This process was repeated 10 times to generate 10 independent cross-validation splits. ScaiVision v1.0 was benchmarked against the following competitor algorithms: CytoDx[8], the "deep CNN" model[9] (herein referred to as "deepCNN"), and a combination of clustering using FlowSOM[6] followed by a random forests model implemented in Python using scikit-learn[10]. CytoDx was tested with either rank-transformed or non- rank-transformed data. DeepCNN was tested with either scaled or non-scaled data. For ScaiVision and FlowSOM, data was scaled only. The Area Under the Receiver-Operator Characteristic (AUROC) was calculated as a performance metric.

Results

ScaiVision outperformed all competitor algorithms at the task of predicting CMV status in the held-out test samples, as measured by the AUC (Figure 1). ScaiVision attained a mean AUC of 0.96 across all 10 cross-validation splits, higher than any other method (Table 1). With a standard deviation of the AUC of 0.031 (Table 1), ScaiVision shows the lowest variability in performance across independently-evaluated test cohorts (Figure 2).

We conclude that ScaiVision is both highly accurate and robust to potential batch effects introduced during the sample collection and processing steps, which may vary across studies.

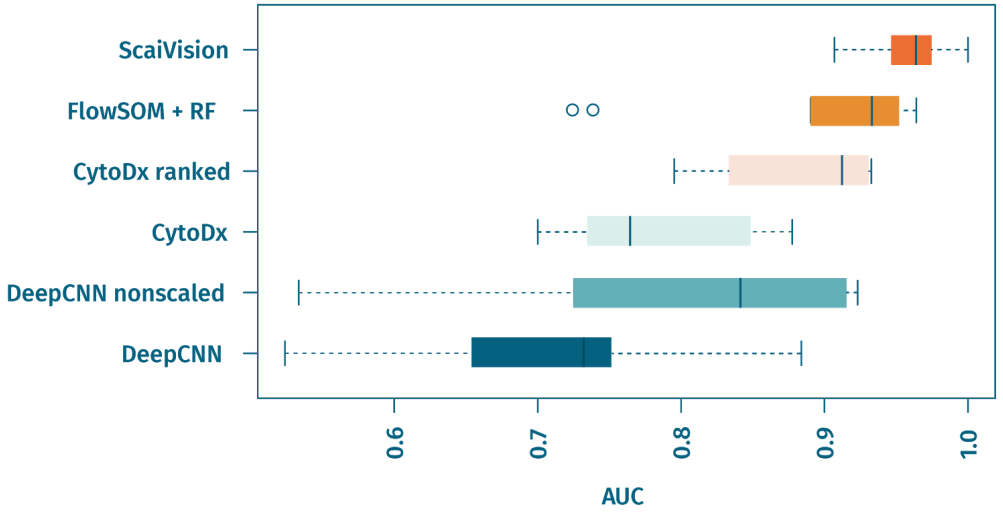


Figure 1. ScaiVision outperforms five competitor algorithms.
The plot shows the performance of ScaiVision and five competitor algorithms in predicting CMV status. Each boxplot represents the AUC from 10 independent cross-validation runs.
The central line shows the median value, the outer box edges show the first and third quartiles, and the whiskers show the most extreme values within 1.5 times the length of the box. Outliers are shown as individual circles.

Algorithm	Mean AUC	Standard deviation AUC
ScaiVision	0.96	0.031
FlowSOM + RF	0.90	0.088
CytoDX ranked	0.89	0.055
CytoDx	0.78	0.065
DeepCNN nonscaled	0.80	0.13
DeepCNN	0.67	0.17

Table 1. Table showing the performance of ScaiVision and five competitor algorithms.
See also Figure 1.

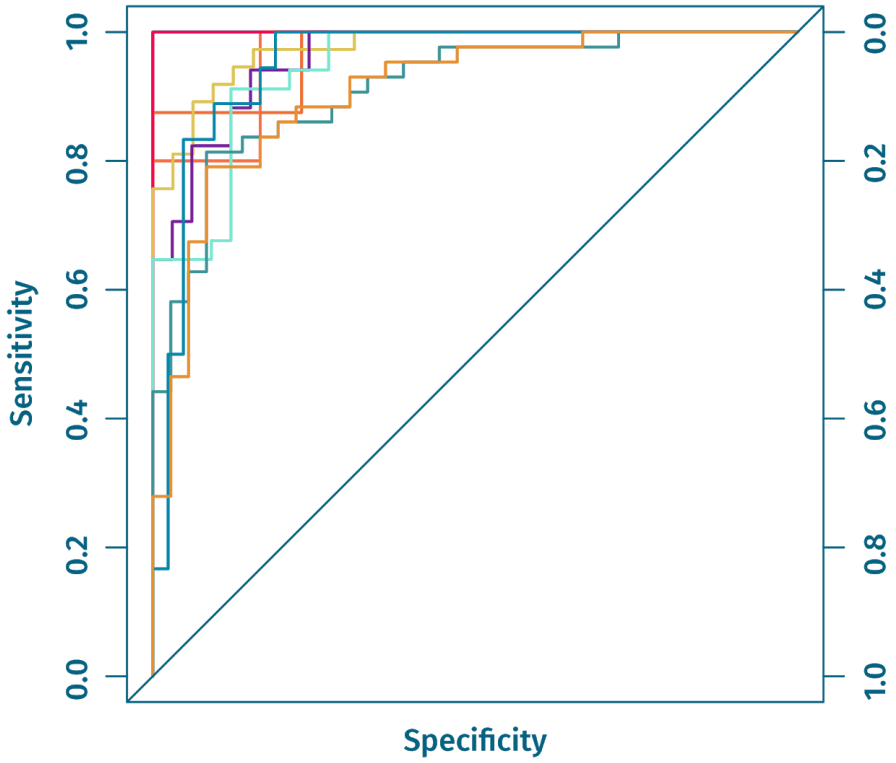


Figure 2. ScaiVision shows high sensitivity and specificity across independent test cohorts.
The Receiver-Operator Characteristic (ROC) curves show the tradeoff between sensitivity and specificity at various possible cutoffs for predicting positive versus negative CMV status. Each colored curve corresponds to predictions made on an independent test cohort, using a network trained and validated on the remaining cohorts.

Conclusions

Our results show that ScaiVision performs best-in-class compared to competitor algorithms at correctly predicting the clinical labels of the samples. The stability of predictions across multiple independent cross-validation runs, together with the ability of ScaiVision to process and interpret massive datasets on a single-cell level without sub-sampling or up-front clustering, shows that analysis with ScaiVision can unlock an unparalleled level of high-resolution and clinically-relevant discoveries in single-cell datasets. Together, these advantages support Scailyte's goal of enabling true precision medicine through single-cell science.

References

1. Method of the Year 2013. *Nat. Methods* **11**, 1–1 (2014).
2. Labib, M. & Kelley, S. O. Single-cell analysis targeting the proteome. *Nat. Rev. Chem.* **4**, 143–158 (2020).
3. Stuart, T. & Satija, R. Integrative single-cell analysis. *Nat. Rev. Genet.* **20**, 257–272 (2019).
4. LifeTime Community Working Groups *et al.* LifeTime and improving European healthcare through cell-based interceptive medicine. *Nature* **587**, 377–386 (2020).
5. Arvaniti, E. & Claassen, M. Sensitive detection of rare disease-associated cell subsets via representation learning. *Nat. Commun.* **8**, 14825 (2017).
6. Van Gassen, S. *et al.* FlowSOM: Using self-organizing maps for visualization and interpretation of cytometry data: FlowSOM. *Cytometry A* **87**, 636–645 (2015).
7. Weber, L. M. & Robinson, M. D. Comparison of clustering methods for high-dimensional single-cell flow and mass cytometry data: Comparison of High-Dim. Cytometry Clustering Methods. *Cytometry A* **89**, 1084–1096 (2016).
8. Hu, Z., Glicksberg, B. S. & Butte, A. J. Robust prediction of clinical outcomes using cytometry data. *Bioinformatics* **35**, 1197–1203 (2019).
9. Hu, Z., Tang, A., Singh, J., Bhattacharya, S. & Butte, A. J. A robust and interpretable end-to-end deep learning model for cytometry data. *Proc. Natl. Acad. Sci.* 202003026 (2020) doi:10.1073/pnas.2003026117.
10. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *Mach. Learn. PYTHON* 6.