

AutoFlow: an interactive Shiny app for supervised and unsupervised flow cytometry analysis

Frey E.R. Woods^{1,2,*}, Emilyanne Leonard³, Timothy Ebbels⁴, Jonathan Cairns², Rhiannon David¹

¹Safety Sciences, Clinical Pharmacology & Safety Sciences, R&D, AstraZeneca, Trumpington, Cambridge, CB2 0AA, United Kingdom

²Data Sciences & Quantitative Biology, Discovery Sciences, R&D, AstraZeneca, Trumpington, Cambridge, CB2 0AA, United Kingdom

³Integrated Bioanalysis, Clinical Pharmacology & Safety Sciences, R&D, AstraZeneca, Trumpington, Cambridge, CB2 0AA, United Kingdom

⁴Section of Bioinformatics, Department of Metabolism, Digestion & Reproduction, Imperial College London, London, W12 0NN, United Kingdom

*Corresponding author. Safety Sciences, Clinical Pharmacology & Safety Sciences, R&D, AstraZeneca, Biomedical Campus, 1 Francis Crick Ave, Trumpington, Cambridge, CB2 0AA, United Kingdom. E-mail: freywoods94@gmail.com.

Associate Editor: Anthony Mathelier

Abstract

Motivation: Flow cytometry (FC) is a widely used technique for analysing cells or particles based on the fluorescence of specific markers. Thresholds for fluorescence are typically set manually, a laborious, subjective process that scales poorly as FC technology advances. Machine learning (ML) methods can address these issues but often require technical expertise many bench scientists do not possess. Thus, accessible, open-source, and cross-domain ML-based FC tools are needed.

Results: We present AutoFlow, an easy-to-use, adaptable R Shiny application for automated flow cytometry (FC) analysis. AutoFlow supports two workflows: supervised and unsupervised learning. The application automates key preprocessing steps including fluorescence compensation, debris exclusion, single-cell identification, viability marker gating, and downstream classification or clustering. Across three datasets, two publicly available (Mosmann and Nilsson Rare) and a novel bone marrow microphysiological system (BM-MPS) dataset, AutoFlow demonstrated robust performance. In the supervised workflow, multiclass classification on BM-MPS achieved 97.2% accuracy under a single-timepoint training and multi-timepoint testing scheme, with high sensitivity and specificity across major lineages. For rare populations, performance was strong: Mosmann Rare (0.03% prevalence) achieved 87.5% sensitivity, and 100% specificity, while Nilsson Rare (0.08% prevalence) achieved 87.9% sensitivity, and 99.9% specificity. The unsupervised workflow accurately grouped cells into biologically meaningful clusters, recovering known populations and identifying additional candidate populations with marker profiles consistent with true biology. AutoFlow offers a fast, reproducible, and scalable solution for FC analysis, enabling high-throughput studies and improving the discovery of rare or unexpected cell types.

Availability and implementation: The application is available at <https://github.com/FERWoods/AutoFlow> for download using R. An archived version is available at DOI: 10.5281/zenodo.18235796.

1 Introduction

Flow cytometry (FC) is a technology to analyse single cells or particles and is widely used in cell biology research for applications including immunology, oncology, and drug discovery and development (Chattopadhyay and Roederer 2010, Woo *et al.* 2014, Bonilla *et al.* 2020, Ullas and Sinclair 2024). FC works by first isolating single cells using fluidics, then exposing the particles to laser light of a specific wavelength and detecting this light after scattering or absorption and emission of fluorescence. By staining cells using a panel of specific antibodies conjugated to fluorescent markers, FC can identify cells when combined with downstream ‘gating’ analysis (Adan *et al.* 2017, McKinnon 2018). Manual gating is the process by which populations of cells expressing or not

expressing markers are determined in a 2d plot. FC is an attractive assay in many laboratories, due to its ability to analyse thousands of particles in real time inexpensively, and its flexibility to many different applications. However, as staining panels contain more biomarkers for better granularity of cell type identification and FC technology advances, manual gating analysis becomes more cumbersome. In contrast, automated methods for cellular identification (gating) scale well and offer many other potential benefits—such as improved accuracy and speed, lower bias, better comparability between labs, more explicit quantification of uncertainty, novel cell type discovery, and better potential for integration with downstream ML tasks.

Many ML-based methods for FC have been developed, such as FlowSOM (Flow self-organizing maps), flowMeans, SWIFT,

Received: 31 December 2024. Revised: 26 January 2026. Accepted: 29 January 2026

© The Author(s) 2026. Published by Oxford University Press.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

flowClust, and others (Lo *et al.* 2008, Finak *et al.* 2009, Lo *et al.* 2009, Pyne *et al.* 2009, Aghaeepour *et al.* 2011, Van Gassen *et al.* 2015). These broadly rely on a semi-supervised ML approaches—a typical workflow consists of dimensionality reduction and clustering to identify similar groups of cells. Biological domain expertise is subsequently required to link these clusters back to the biological outcome (Pedersen and Olsen 2020). These methods have made great progress; however, they usually require some pre-processing of the data and, due to the need for hands-on coding expertise, are inaccessible to most wet lab scientists. Hauchamps *et al.* (2024) recently published their automatic pre-processing pipeline for FC data, but it does not extend to downstream analysis.

We present here an approach to use expert knowledge to train supervised machine learning algorithms for different specific-purpose classifications. In addition, we include an unsupervised workflow to aid with exploratory analysis and enable a data-driven approach to identifying cells from FC. This approach uses dimensionality reduction, clustering, and differential expression analysis (Pyne *et al.* 2009, Aghaeepour *et al.* 2011, Pedersen and Olsen 2020, Hauchamps *et al.* 2024).

Our methodology is presented in an easy-to-use, and easy-to-adapt R Shiny application available at github.com/ferwoods/AutoFlow. Autoflow supports flow cytometry standard (.FCS) files for end-to-end processing, enabling biologists and bioinformaticians alike to analyse FC data with ease, saving time and improving reproducibility.

2 Methods

2.1 Benchmarking datasets

We demonstrate both Autoflow workflows (supervised and unsupervised) on three datasets.

The first is a novel dataset generated in-house using the BM-MPS. Briefly, BM-MPS chips are cultured with haematopoietic stem cells (HSCs) and mesenchymal stem cells in a ceramic scaffold mimicking the BM. This enables the study of a full cycle of BM injury and recovery in response to therapeutics recapitulating clinical BM toxicities (Chou *et al.* 2020). Table S1, available as [supplementary data](#) at *Bioinformatics* online summarizes the cluster of differentiation (CD) markers used to identify each cell type. Further information on FC data acquisition can be found in the [Supplementary Material](#) for BM-MPS.

The second and third datasets are publicly available data—namely, the Mosmann Rare (Mosmann *et al.* 2014) and Nilsson Rare (Rundberg Nilsson *et al.* 2013) datasets previously published and collated by Weber and Sonesson (2019). Mosmann Rare contains peripheral blood cells with labels provided only for a rare type of cell rare cytokine-producing influenza-specific T cells. Nilsson Rare originates from haematopoietic cells, with labels provided for HSCs.

The approach was trained using cross-validation and tested on a held-out validation set—full details and results are described in the [Supplementary Material](#).

2.2 Optional preprocessing

Before either ML workflow, an optional preprocessing pipeline can be applied. This includes margin removal, automated debris

detection, QC [margin events and PeacoQC algorithm (Emmaneel *et al.* 2022)], and viability identification. Debris is removed using a two-component Gaussian mixture model [GMM, via Mclust (Scrucca *et al.* 2023)] on forward-scatter measurements (FSC-A/H; using a radial summary when both are present). At this stage, raw cytometer values are used (no compensation or transformation), mirroring manual debris screening. The data are then compensated (using the spillover matrix when available) and transformed with a logicle transform. For viable-cell identification, users select a ‘viability’ channel and a two-component GMM proposes a viability threshold, which can be adjusted interactively. Each preprocessing step is fault-tolerant (failures are skipped rather than aborting the run), if all preprocessing fails, the analysis proceeds on raw data. At export, an additional doublet check (FSC.A versus FSC.H) is applied to write singlet flags into the processed FCS files.

2.3 Supervised machine learning

The supervised arm of AutoFlow requires the user to upload a pre-trained model, i.e. one trained on manually gated data. In this paper, all results were generated using a random forest classifier. RF is an ensemble machine-learning method that combines predictions from many decision trees (Breiman 2001), excelling in capturing complex non-linear relationships, handling high-dimensional feature spaces, and mitigating issues such as overfitting.

The app itself is agnostic to model type, provided it is compatible with R’s `predict()` function, including any caret-based models. These models are uploaded in a ‘bundle’ that contains the fitted model and scaling information. When new data is uploaded, events are z-score normalized based on the mean and standard deviation of the training dataset. At runtime, the app automatically maps features in the uploaded dataset to the expected model features; users can review or adjust these mappings within the interface. If features are missing or mismatched, the app flags them to prevent generating predictions until alignment is complete.

For benchmarking, the Nilsson and Mosmann datasets were split into 70% training and 30% testing sets. For the BM-MPS dataset, the model was trained on Day 14 (selected for maximal lineage diversity) and tested on all other time points. Hyperparameters (mtry and ntree) were tuned via grid search.

2.4 Unsupervised machine learning

The unsupervised version of the application relies on dimensionality reduction using principal component analysis (PCA) and uniform manifold approximation and projection (UMAP, with default parameters), clustering with Leiden by default, (if available, else Louvain), and differential expression analysis (DEA) to identify defining cluster characteristics (Jolliffe 1986, Blondel *et al.* 2008, McInnes *et al.* 2018). In this application we utilize the Seurat workflow, which is developed for use with RNA-sequencing datasets (Satija *et al.* 2015, Stuart *et al.* 2019). Users can optionally alter clustering resolution and fold change threshold for DEA.

The unsupervised side of Fig. 1 shows a graphical representation of the workflow used in this section of the application.

for identifying major cell types and the need to model haematopoiesis as a continuum when analysing transitional populations.

4 Discussion

The AutoFlow application uses supervised and unsupervised ML to enable cell identification. This app provides a bespoke, fast-running interface to make FC data analysis more accessible. In this study we automatically gated cells of known origin using manually gated data as a training set. In the case of the unsupervised version, cells are identified based on expression of surface markers and compared with manually gated study results.

The supervised version of the application provides a structured and guided approach to data analysis with high performance of >90% accuracy in the three-benchmarking dataset. Researchers can leverage their expertise to define the gating criteria, offering flexibility and the ability to incorporate domain knowledge into the process. This method is particularly useful when dealing with well-understood, standard datasets for which specific populations of interest need to be identified. Hence, the supervised approach streamlines the process, reduces subjectivity, and ensures reproducibility that otherwise remains a challenge in manual gating.

Conversely, the unsupervised gating version caters to the increasingly prevalent challenge of managing complex, high-dimensional data. With this method, the application automatically identifies clusters and patterns within the data, revealing potentially novel and unexpected cell populations. By being data-driven, this method is ideal for exploratory analysis and can unveil hidden insights within heterogeneous datasets. The unsupervised approach enhances the ability to discover rare or unanticipated cell populations, offering a more comprehensive understanding of the biological system under investigation.

The combined use of both supervised and unsupervised versions within the application fosters a comprehensive data analysis pipeline. Researchers can use either workflow or use both by, e.g. using supervised gating at an early stage to refine known populations and subsequently apply the unsupervised approach to uncover novel and more subtle cell types, or outliers. Used together, these workflows ensure that both well-established and hidden insights are fully explored, ultimately enhancing the depth and quality of the analysis.

Nonetheless, a key limitation lies in the underlying data. Both supervised and unsupervised methods rely on the discriminatory power of the chosen markers; when marker expression is noisy or overlapping, algorithms cannot fully resolve certain populations. This reflects constraints of experimental design rather than AutoFlow itself, underscoring the importance of thoughtful panel design in future applications.

In summary, the application's integration of supervised and unsupervised gating methods brings a holistic approach to FC data analysis. It equips researchers with a versatile toolset to tackle various research questions, from precise quantification of known populations to the discovery of unexpected and novel cell subsets. This capability holds great promise for advancing the understanding of complex biological systems and it has the potential to drive transformation across research domains.

Author contributions

Freya E.R. Woods (Conceptualization [equal], Data curation [equal], Formal analysis [lead], Investigation [equal], Methodology [equal], Project administration [equal], Software [lead], Validation [lead], Visualization [lead], Writing—original draft [lead], Writing—review & editing [equal]), Emilyanne Leonard (Data curation [equal], Investigation [equal], Methodology [equal], Resources [equal], Writing—original draft [equal], Writing—review & editing [equal]), Timothy Ebbels (Conceptualization [supporting], Methodology [supporting], Supervision [supporting], Writing—review & editing [supporting]), Jonathan Cairns (Conceptualization [equal], Formal analysis [supporting], Funding acquisition [lead], Investigation [equal], Methodology [equal], Resources [equal], Supervision [lead], Writing—original draft [equal], Writing—review & editing [equal]), and Rhiannon David (Conceptualization [equal], Funding acquisition [lead], Investigation [equal], Methodology [equal], Supervision [lead], Writing—original draft [equal], Writing—review & editing [equal])

Supplementary material

[Supplementary material](#) is available at *Bioinformatics* online.

Conflicts of interest

None declared.

Funding

This work was funded through a postdoctoral fellowship with AstraZeneca.

Data availability

The data underlying this article are available in publicly accessible repositories. The Mosmann rare and Nilsson rare datasets were obtained from the Flow Repository (FlowRepository ID: FR-FCM-ZZPH), available at <http://flowrepository.org/id/FR-FCM-ZZPH>. The BM-MPS dataset generated and analysed during the current study is available in Zenodo at <https://doi.org/10.5281/zenodo.18235796>.

References

- Adan A, Alizada G, Kiraz Y *et al.* Flow cytometry: basic principles and applications. *Crit Rev Biotechnol* 2017;**37**:163–76. <https://doi.org/10.3109/07388551.2015.1128876>
- Aghaeepour N, Nikolic R, Hoos HH *et al.* Rapid cell population identification in flow cytometry data. *Cytometry A* 2011;**79**: 6–13. <https://doi.org/10.1002/cyto.a.21007>
- Blondel VD, Guillaume J-L, Lambiotte R *et al.* Fast unfolding of communities in large networks. *J Stat Mech* 2008;**2008**: P10008. <https://doi.org/10.1088/1742-5468/2008/10/P10008>
- Bonilla DL, Reinin G, Chua E. Full spectrum flow cytometry as a powerful technology for cancer immunotherapy research.

- Front Mol Biosci 2020;**7**:612801. <https://doi.org/10.3389/fmolb.2020.612801>
- Breiman L. Random forests. *Mach Learn* 2001;**45**:5–32. <https://doi.org/10.1023/A:1010933404324>
- Chattopadhyay PK, Roederer M. Good cell, bad cell: flow cytometry reveals T-cell subsets important in HIV disease. *Cytometry A* 2010;**77**:614–22. <https://doi.org/10.1002/cyto.a.20905>
- Chou DB, Frismantas V, Milton Y *et al*. On-chip recapitulation of clinical bone marrow toxicities and patient-specific pathophysiology. *Nat Biomed Eng* 2020;**4**:394–406. <https://doi.org/10.1038/s41551-019-0495-z>
- Emmaneel A, Quintelier K, Sichien D *et al*. PeacoQC: peak-based selection of high quality cytometry data. *Cytometry A* 2022;**101**:325–38. <https://doi.org/10.1002/cyto.a.24501>
- Finak G, Bashashati A, Brinkman R *et al*. Merging mixture components for cell population identification in flow cytometry. *Adv Bioinformatics* 2009;**2009**:247646–12. <https://doi.org/10.1155/2009/247646>
- Hauchamps P, Bayat B, Delandre S *et al*. CytoPipeline and CytoPipelineGUI: a bioconductor R package suite for building and visualizing automated pre-processing pipelines for flow cytometry data. *BMC Bioinformatics* 2024;**25**:80. <https://doi.org/10.1186/s12859-024-05691-z>
- Jolliffe IT. *Principal Component Analysis*. New York: Springer; 1986. <https://doi.org/10.1007/978-1-4757-1904-8>
- Lo K, Brinkman RR, Gottardo R. Automated gating of flow cytometry data via robust model-based clustering. *Cytometry A* 2008;**73**:321–32. <https://doi.org/10.1002/cyto.a.20531>
- Lo K, Hahne F, Brinkman RR *et al*. flowClust: a Bioconductor package for automated gating of flow cytometry data. *BMC Bioinformatics* 2009;**10**:145. <https://doi.org/10.1186/1471-2105-10-145>
- McInnes L, Healy J, Saul N *et al*. UMAP: uniform manifold approximation and projection for dimension reduction. *J Open Source Softw* 2018;**3**:861. <https://doi.org/10.21105/joss.00861>
- McKinnon KM. Flow cytometry: an overview. *Curr Protoc Immunol* 2018;**120**:5.1.1. <https://doi.org/10.1002/cpim.40>
- Mosmann TR, Naim I, Rebhahn J *et al*. SWIFT—scalable clustering for automated identification of rare cell populations in large, high-dimensional flow cytometry datasets, part 2: biological evaluation. *Cytometry A* 2014;**85**:422–33. <https://doi.org/10.1002/cyto.a.22445>
- Pedersen CB, Olsen LR. Algorithmic clustering of single-cell cytometry data—how unsupervised are these analyses really? *Cytometry A* 2020;**97**:219–21. <https://doi.org/10.1002/cyto.a.23917>
- Pyne S, Hu X, Wang K *et al*. Automated high-dimensional flow cytometric data analysis. *Proc Natl Acad Sci USA* 2009;**106**:8519–24. <https://doi.org/10.1073/pnas.0903028106>
- Rundberg Nilsson A, Bryder D, Pronk CJH. Frequency determination of rare populations by flow cytometry: a hematopoietic stem cell perspective. *Cytometry A* 2013;**83**:721–7. <https://doi.org/10.1002/cyto.a.22324>
- Satija R, Farrell JA, Gennert D *et al*. Spatial reconstruction of single-cell gene expression data. *Nat. Biotechnol* 2015;**33**:495–502. <https://doi.org/10.1038/nbt.3192>
- Scrucca L, Fraley C, Murphy T *et al*. *Model-Based Clustering, Classification, and Density Estimation Using mclust in R*. New York: Chapman and Hall/CRC, 2023. <https://doi.org/10.1201/9781003277965>
- Stuart T, Butler A, Hoffman P *et al*. Comprehensive integration of single-cell data. *Cell* 2019;**177**:1888–902.e21. <https://doi.org/10.1016/j.cell.2019.05.031>
- Ullas S, Sinclair C. Applications of flow cytometry in drug discovery and translational research. *Int J Mol Sci* 2024;**25**:3851. <https://doi.org/10.3390/ijms25073851>
- Van Gassen S, Callebaut B, Van Helden MJ *et al*. FlowSOM: using self-organizing maps for visualization and interpretation of cytometry data. *Cytometry A* 2015;**87**:636–45. <https://doi.org/10.1002/cyto.a.22625>
- Weber LM, Soneson C. HDCytoData: collection of high-dimensional cytometry benchmark datasets in Bioconductor object formats. *F1000Res* 2019;**8**:1459. <https://doi.org/10.12688/f1000research.20210.1>
- Woo J, Baumann A, Arguello V. Recent advancements of flow cytometry: new applications in hematology and oncology. *Expert Rev Mol. Diagn* 2014;**14**:67–81. <https://doi.org/10.1586/14737159.2014.862153>