


Automated cytometric gating with human-level performance using bivariate segmentation

Received: 19 June 2024

Accepted: 21 January 2025

Published online: 12 February 2025

 Check for updates

Jiong Chen^{1,2}, Matei Ionita^{3,4}, Yanbo Feng², Yinfeng Lu^{2,5}, Patryk Orzechowski^{2,6}, Sumita Garai², Kenneth Hassinger³, Jingxuan Bao², Junhao Wen⁷, Duy Duong-Tran^{2,8}, Joost Wagenaar³, Michelle L. McKeague^{3,4}, Mark M. Painter^{3,4}, Divij Mathew^{3,4}, Ajinkya Pattekar^{3,4}, Nuala J. Meyer⁹, E. John Wherry^{3,4}, Allison R. Greenplate^{3,4} & Li Shen²✉

Recent advances in cytometry have enabled high-throughput data collection with multiple single-cell protein expression measurements. The significant biological and technical variance in cytometry has posed a formidable challenge during the gating process, especially for the initial pre-gates which deal with unpredictable events, such as debris and technical artifacts. To mitigate the labor-intensive manual gating process, we propose UNITO, a framework to rigorously identify the hierarchical cytometric subpopulations. UNITO transforms a cell-level classification task into an image-based segmentation problem. The framework is validated on three independent cohorts (two mass cytometry and one flow cytometry datasets). We compare its results with previous automated methods using the consensus of at least four experienced immunologists. UNITO outperforms existing methods and deviates from human consensus by no more than any individual does. UNITO can reproduce a similar contour compared to manual gating for post-hoc inspection, and it also allows parallelization of samples for faster processing.

Cytometric analysis has gained tremendous attention in immunological experiments, as a method that produces reliable, high-throughput measurements of single cells. Flow cytometry, introduced in the 1960s to separate and count immune cell subtypes, underwent a critical evolution with increased speed and parameters^{1–4}. Mass cytometry, also known as cytometry by time-of-flight, was developed in 2009 to include more parameters in the analysis and avoid the difficulty of fluorescence compensation^{5–7}.

Flow cytometry labels the antibodies with fluorophores, whose emission spectra overlap to some extent, causing spillover of signal between channels. In contrast, mass cytometry tags antibodies with heavy metal ions, whose discrete atomic masses can be distinguished by mass spectrometry very precisely and with minimal spillover. Although flow cytometry and mass cytometry aim for similar functionality with labeled antibodies, they do not have similar strategies to detect cell aggregates⁸.

¹Department of Bioengineering, University of Pennsylvania School of Engineering and Applied Science, Philadelphia, PA, USA. ²Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA, USA. ³Department of Systems Pharmacology & Translational Therapeutics, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA, USA. ⁴Institute for Immunology and Immune Health, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA, USA. ⁵Department of Mathematics, University of Pennsylvania School of Arts and Sciences, Philadelphia, PA, USA. ⁶Department of Automatics and Robotics, AGH University of Science and Technology, Krakow, Poland. ⁷Laboratory of AI and Biomedical Science (LABS), University of Southern California, Los Angeles, CA, USA. ⁸Department of Mathematics, United States Naval Academy, Annapolis, MD, USA. ⁹Division of Pulmonary and Critical Care Medicine, Department of Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA. ✉e-mail: li.shen@pennmedicine.upenn.edu

The most important and widely used method of subtyping cell populations is referred to as gating, which is usually a boundary or polygon defined on a bivariate density map of the entire cell population (2D density map represented by two selected protein measurements). Gating for different cell types is usually configured hierarchically⁹, to mimic the biological hierarchy of immune cell differentiation. Even though mass cytometry and flow cytometry rely on different technologies, they both employ similar gating procedures for cell type identification. However, manually defining boundaries for each cell subtype and each sample can be time-consuming and labor-intensive. Thus, algorithms have been developed that attempt to automate this task, including unsupervised clustering to identify cell subtypes^{10–13}, supervised deep learning methods^{14–17}, as well as some density-based gating methods^{18,19}. Although unsupervised methods can target certain cell subpopulations effectively, their results often contain some remaining cell clusters that are unrecognizable and/or sometimes can fail to capture rare cell types making interpretation challenging²⁰. Such methods usually lack the ability to reproduce cytometric gating from human experts, and it is hard for those methods to adapt to existing domain knowledge to satisfactorily explain the clustering results.

A typical manual analysis pipeline of this data first separates live, viable single cells from unwanted events, then gates these into different cell populations. A major challenge is that methods, like those described above, typically assume that single cells have been separated from doublets, debris, and other unwanted events, a process that we call “pre-gating”. Efficient open-source methods employing pre-gating using adaptive approaches to define populations and account for variability in signal intensity between samples, however, do not exist. Indeed, current unsupervised clustering methods typically require manual “pre-gating” before applying unsupervised analysis. While manual pre-gating still requires much human labor to annotate and draw polygons for the region of interest, existing bivariate pipelines also need human input to pre-define the relevant parameters. For example, those parameters include where the actual population is located as well as the approximate percentage of the target cell population. Currently, no software can configure a series of bivariate pre-gating (and gating) tasks where gates are set in a hierarchical order. Our goal is to automate the analysis process.

The application of deep learning has been effective in disentangling complex relationships in different cytometric domains including cell type identification and cell sorting in cytometric time series data^{14–17}. Most of the existing methods focus on providing a global classification of different cell types in one step by using a multi-channel dense neural network. Moreover, these methods also usually assume pre-gated data as input, and they have not been validated for complex pre-gating tasks. In particular, one of the major challenges in performing automatic pre-gating is the technical and biological variance across different subjects. Even if the experimental protocol and the panel of measured proteins are held constant, differences in sample preparation and even instrument variation, can cause fluctuations in protein detection and precise population “shape” in the data. In addition, only predicting terminal cell types (such as naïve T cells) will prevent accurate gating of the intermediate cell types (such as T cells or lymphocytes). In the bivariate setting from manual gating, gates are configured in a tree-like structure, and at each level, the parent cell population is split into smaller subpopulations so that gating results for all intermediate cell types are obtained. The hierarchy structure guarantees the interpretability of the target cell populations. It also clearly defines some cells that are not included in the subsequent gating but still belong to the current gating step (out-of-boundary cells that are not included in any subsequent cell types). The same strategy used for manual pre-gating can be easily extended to downstream automatic gating of immune cell types, since cell-type gates have a more stable and fixed bivariate boundary compared to

pre-gates. Therefore, we combined those two types of tasks and refer to all of them, onward, as “gating”.

While existing methods are suboptimal to accommodate biological variance within the protein expression data, deep neural networks with convolutional kernels have the ability to address such data challenges, using the properties of translational invariance and equivariance^{21,22}. The convolutional architecture was originally designed for image classification tasks. It can detect target objects regardless of their positions in the image space and learn the general features of each target object. This versatility led to their adaptation for image segmentation tasks. The reason manual gating is challenging for automated software is that human experts have a global view of the cell density in certain protein measurement spaces so that they can visually inspect and quantify the desired cell type. Therefore, to address this challenge together with the biological and technical variance, we propose and validate UNITO, a method employing image segmentation for automated gating. UNITO converts the cytometrically derived protein expression into an image of bivariate density to enable a global identification of the cell population. Furthermore, with the ability to perform pixel-level prediction, UNITO intuitively defines the region of interest on the bivariate density maps for cytometric gating. By validating UNITO on three independent study cohorts and two cytometric modalities, we hypothesize that the framework can learn any pre-gating and gating tasks from human annotation, and then adaptively draw contours and assign labels to cells from independent data. This ability to perform inference on unseen samples without human supervision will enable applications to large-scale immunology studies.

Results

Flow and mass cytometry rely on indirect reporters for protein expression, which makes the distribution of the data susceptible to noise from unavoidable variations in sample preparation. This source of technical variation is one of the most important challenges for automated analysis. In particular, taking mass cytometry as an example, due to the high variability of channels such as DNA intercalator in the Single-cell gate 1 and Single-cell gate 2 process, existing automated methods usually require the data to be manually cleaned beforehand. Similar situation also happens in the lymphocyte gate in the flow cytometry gating process. Between different subjects, while the coordinate system for the density map is fixed, the dense population (singlets) not only moves dramatically, but the distribution also changes (Fig. 1). The location and shape of the target cell population can differ widely between samples, which makes automatic detection over the tabular data challenging. With the UNITO framework, our goal is to create the density plot and binary mask from the protein expression matrix as input to the model and predict with an image-based segmentation method. Each gate will be trained separately, and the classifier will then be used for predicting the binary mask on independent data. The final output of the UNITO framework is the cell type label for each cell and a convex contour on the density map, which resembles the manual gates.

UNITO framework overview

Instead of reading the numerical values of cytometric measurement into the model to predict cell types, UNITO is designed to mimic the process of manual gating, where the cell population of interest is decided by drawing polygons on a bivariate density plot. UNITO uses the bivariate density plot of the entire cell population as its training data, while human annotation masks serve as the corresponding training labels (Fig. 2A). In other words, the central goal of UNITO is to draw polygons on the bivariate density map, and the cell annotation is inferred from the output mask. Based on the selected measurement and label, UNITO first contains a preprocessing step that will (1) normalize the protein expression by min and max values to a range of

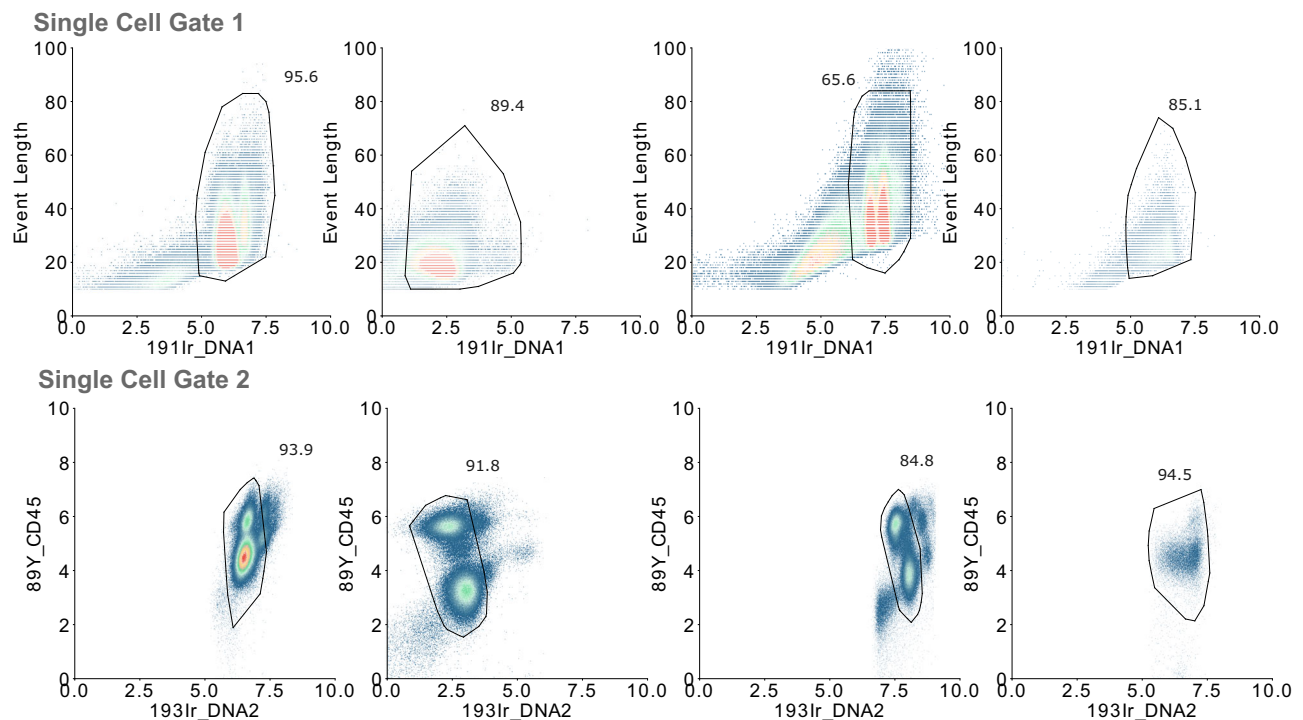


Fig. 1 | Biological and technical heterogeneity in mass cytometry across subjects. Four examples demonstrate the heterogeneity across different subjects in the first two pre-gating manual gating tasks using mass cytometry data. The top rows are single cell gate 1 and the bottom rows are single cell gate 2. The event

length variable, which is the physical size of the ion cloud that results from vaporizing the cell, is an integer data, resulting in a sparser representation in the density plot. All protein markers are Arcsinh-transformed with a cofactor of 5.

[0,100], and (2) convert the normalized protein expression data to the density plot. Subsequently, it leverages binary labels attributed to each cell to generate an overlay mask atop the density plot. A convex hull processing will fill any empty space within the binary mask to improve the gating performance. The bivariate maps and masks are fed into the model for learning the gating pattern. The prediction output from the UNITO model is the binary mask for any independent data for validation, and the mask will undergo an additional post-processing step to interpolate the pixel label back to the single-cell classification results (Fig. 2B). The same procedure is repeated for each gate recursively throughout the training process. When gating on a new cytometry panel, training a new UNITO model may be necessary, which requires a set of 30–40 manually gated cytometry samples and a defined gating hierarchy. Training a new model is also recommended when the panel design or gating requirements are changed dramatically to ensure model performance still aligns with human expectations.

UNITO overall evaluation

Manual gating is considered the gold standard in immunophenotyping, and the accuracy of automated methods is usually defined by comparison to a human annotator. However, this approach is sensitive to subjective choices made by one person. We consider a more robust choice of ground truth, by building consensus gates based on multiple annotators. To validate the performance of UNITO, we constructed the ground truth gating standard by taking the consensus of multiple human annotators and compared the results with five other methods: static gates, FlowDensity¹⁸, FlowSOM¹⁰, logistic regression, and DeepCyTOF¹⁴. In addition, we also compared UNITO's performance on the singlet gate with PeacoQC²³, which is a flow cytometry QC software. While there are other data QC software such as flowAI²⁴ or flowClean²⁵, none of them can handle singlet gating tasks. The static gates method uses the idea of density plot and binary mask construction to perform auto-gating, and every subject will get the exact same mask on the normalized space by averaging the ground truth mask. The static

gating approach serves as a baseline for the proposed method. FlowDensity is a semi-automated tool built in R to gate cytometric data with positional encoding and cell population percentage. FlowSOM uses unsupervised clustering to find each cell population. Finally, logistic regression as well as DeepCyTOF use cell-level prediction by machine learning and neural network methods. Since some methods, like FlowSOM, produce categorical group outputs, we applied a one-hot encoding strategy across all gate types to standardize the prediction results across methods. The final cell type assignment is based on the relative expression on selected markers and conducted from bottom to top in the hierarchical structure.

Overall, the UNITO gating for both mass cytometry data and flow cytometry data showed high correlations with the gold standard manual gating results (Fig. 3A; Fig. S6A). Among automated methods, the UNITO prediction was most highly correlated with the consensus gating, achieving an average correlation of 0.98 for mass cytometry and 0.97 for flow cytometry (Tables S1, 2). Moreover, UNITO had comparable, and sometimes higher, correlation coefficients than the manual gating done by individual annotators (Fig. 3B; Fig. S6B). To assess the performance of the UNITO framework, the average accuracy score, recall, precision, and F1 score (harmonic mean of precision and recall) were calculated across all subjects (Tables 1, 2). We also compared UNITO's performance with a specific tool PeacoQC, designed for flow cytometry data quality control (Table 3). Since PeacoQC can only handle the singlet gate in flow cytometry, we used ground truth gating for lymphocytes to filter out the non-lymphocyte cells for singlet gating. The accuracy score measures the number of correct predictions over all the data, the recall measures the number of true positive predictions over all of the ground true positive data, and the precision measures the number of true positive predictions over all predicted positives. Additionally, the F1 score can further disclose a more comprehensive evaluation especially when the data is imbalanced. The gating results from UNITO outperform other methods in F1 measurements for all gating tasks, and its consistency over all gating tasks

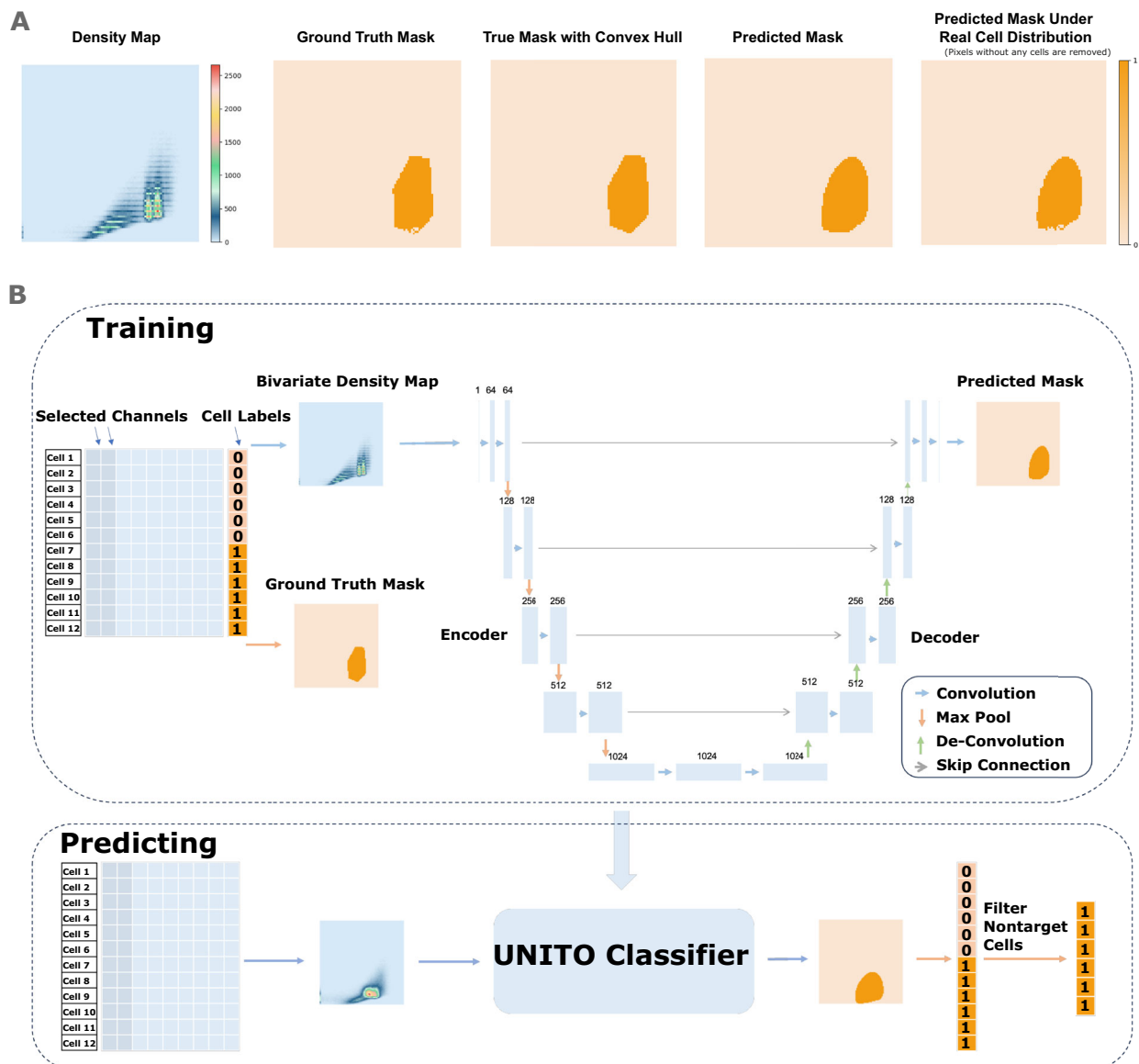


Fig. 2 | UNITO framework illustration. A Example subject input and output from single cell gate 1. The five figures from left to right are: density plot as training input, binary mask, binary mask after convex hull as training label, predicted binary mask, and reconstructed binary mask after interpolation. **B** Two protein measurements are selected from the expression matrix (blue matrix) and used to construct the density map, and the corresponding cell label (orange column) is used to map the

overlap mask. The density maps and binary masks for the desired cell cluster are taken as training input. Prediction is performed on new density maps without annotation, and the output is the binary mask and the mapping results for all cells. The demonstration is only used for a single gate, multiple gates need to be cascaded to achieve the hierarchical gating structure.

guarantees its robustness for applications in flow and mass cytometric gating scenarios. In addition, UNITO uniquely performed lower-level gating tasks with high accuracy. Overall results showcased that the UNITO framework can accurately identify singlets in sequential pre-gating settings and downstream cellular populations. For its extended application in flow cytometry data, we also observed close-to-human performance that outperforms other existing methods (Fig. S6B). To evaluate the generalization ability of UNITO, we compared its performance on in-batch and out-of-batch gating (Table S7). Specifically, we randomly selected 5 samples each from the mass cytometry vaccine dataset and the acute dataset for testing (see Fig. S11 for distribution difference between two batches), while training the UNITO model on the remaining samples from the vaccine dataset. The results show that UNITO maintains comparable performance when predicting out-of-batch samples, with some evaluation metrics even higher than in-batch results. This consistency across two independent datasets

demonstrates UNITO's robustness and its strong generalization capability, as a reliable automation for gating on multiple data cohorts.

Moreover, we also performed ablation studies to assess the impact of each component in UNITO on the cell annotation. First of all, we tested a range of image sizes between 26 and 701 pixels (Fig. S10). We found that no other image sizes improved performance, and 101 pixels provided the most intuitive visualization. We also experimented with constructing a 3D volume instead of a bivariate plot by adding an additional feature (Table S5). This significantly decreased performance and greatly increased computational resource demands, as the larger image volume caused the cell density representation to become more sparse. Furthermore, we examined the performance of applying dimensionality reduction before UNITO gating, as well as replacing semantic segmentation with unsupervised clustering (Tables S4 and S6). However, none of these configurations outperformed our default setup.

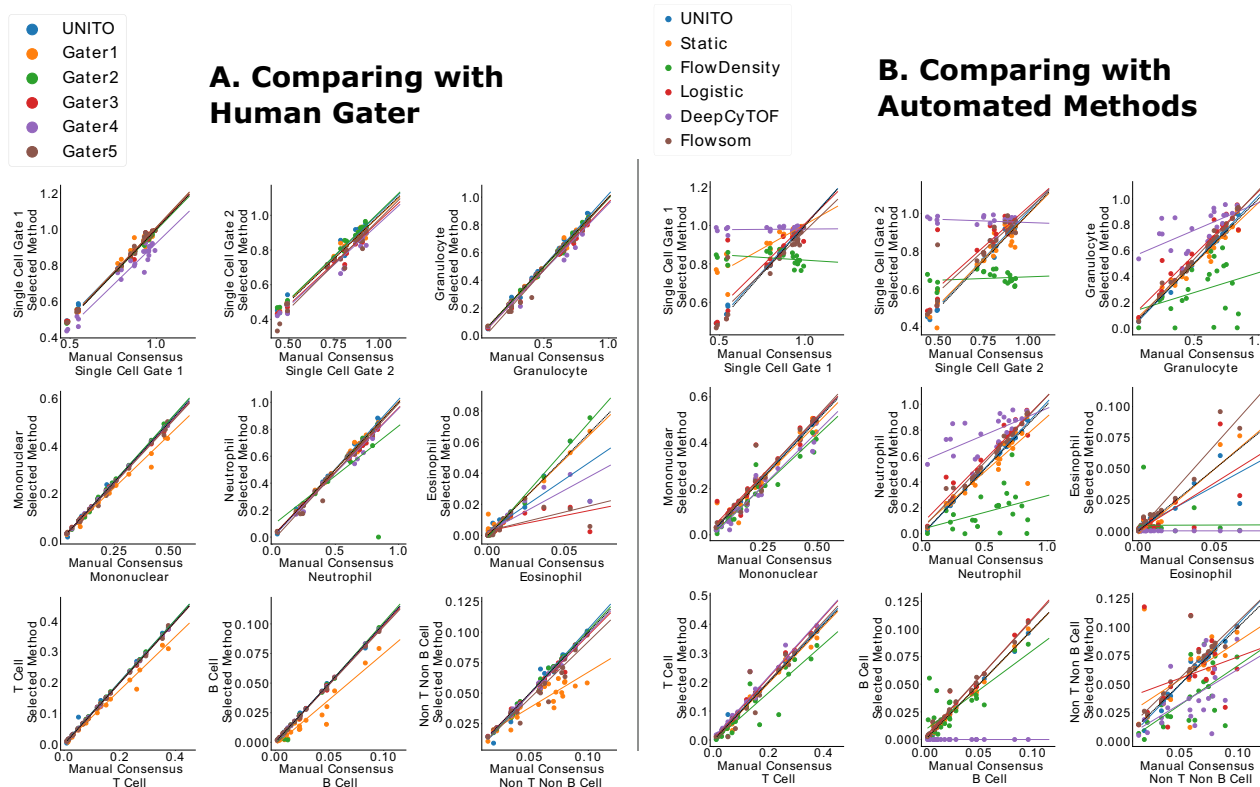


Fig. 3 | Proportion of cells over the entire cell population captured by manual gating versus human/automated gating in mass cytometry. A Comparison between UNITO and human gaters. **B** Comparison between UNITO and other automated gating methods. Each dot in the plot represents a single subject (one FCS file). The plot is separated by every single gate (including single cell gate 1, single cell gate 2, granulocyte, mononuclear, neutrophil, eosinophil, T cell, B cell, and non T non B cell), and within the same coordinate results from each method were visualized by different colors. The dashed line represents a perfect correlation with manual consensus gating, and the Pearson correlation coefficients are reported in the supplementary Table S1. Source data are provided as a Source Data file.

UNITO gating on mass cytometry data

We next asked whether UNITO's density map with decision boundary for the sequential gating task visually aligned with the contour from human experts' consensus (Fig. 4A; Fig. S2). Since different methods may provide different in-gate cell populations, the whole cell density for the visualization of the next gate may also vary across methods, resulting in different background cell population intensity. The ground truth boundary derived from consensus gating is almost the same as the contour predicted by UNITO in all gates. Since different gates within the same gating hierarchy have the same coordinate space, those gates are plotted together. The first two pre-gating steps in the sequential prediction task are more complex than cell subpopulation gating tasks because the presence of debris cells and doublet cells for both gates usually differ in its distribution and position in the density plot across samples. This challenge, if not addressed by manual or automated data cleaning, may affect all subsequent gating results. The gating output from UNITO also showed a high correlation between the proportion of predicted cells with manually gated cells as a nearly straight line in combination with all gating results from UNITO (Fig. 4B), indicating its high consistency with the consensus gating results and ability to perform gating tasks at a level similar to human. In addition, UNITO also shows the lowest disparity with human gaters compared with other automated gating methods (Fig. 4D) when looking at the number of cells included in each gate, where UNITO (red dot) is always closest to the range of values defined by human gaters.

Another key feature of manual annotation is the ability of the expert to identify cells not only by whether the cells that fall into a gate, but also if they are excluded by a gate. UNITO framework mimics this feature by excluding out-of-gate cells in gate 1 as input for the next

prediction gate. This pre-filtering step can also be visualized as the difference in the entire cell density plot between manual gating and automatic gating. The decision boundary from UNITO provides a convex hull that is similar to the human annotation process, which can offer post-prediction adjustment based on the vertex of the convex hull (Fig. 5A). If we visualize using cell-level prediction results, it will not make sense in the bivariate visualization. For image and density-based estimation, all the cells within the boundaries are classified as the target cell type, but when we visualize the results from previously published cell-level prediction, not all cells within the gates are classified by the methods (e.g., FlowDensity), or a simple convex hull over all labeled data may also include cells that are not classified as the target population (e.g., DeepCyTOF). This becomes more apparent when we show the binary distribution of all the cells classified as the target cell type in the same coordinate space as the density map (Fig. 5B). We can see that for both convex hull visualization and binary mask, the UNITO's prediction is the closest estimation approaching the manual consensus gating.

prediction gate. This pre-filtering step can also be visualized as the difference in the entire cell density plot between manual gating and automatic gating. The decision boundary from UNITO provides a convex hull that is similar to the human annotation process, which can offer post-prediction adjustment based on the vertex of the convex hull (Fig. 5A). If we visualize using cell-level prediction results, it will not make sense in the bivariate visualization. For image and density-based estimation, all the cells within the boundaries are classified as the target cell type, but when we visualize the results from previously published cell-level prediction, not all cells within the gates are classified by the methods (e.g., FlowDensity), or a simple convex hull over all labeled data may also include cells that are not classified as the target population (e.g., DeepCyTOF). This becomes more apparent when we show the binary distribution of all the cells classified as the target cell type in the same coordinate space as the density map (Fig. 5B). We can see that for both convex hull visualization and binary mask, the UNITO's prediction is the closest estimation approaching the manual consensus gating.

UNITO gating on flow cytometry data

The same visualization procedure was also repeated for the flow cytometry gating (Fig. 6A; Fig. S3). For automatic gating on flow cytometry, the UNITO framework still performs well in sequential gating tasks with both quantitative evaluation and visual inspection. The UNITO framework can accurately identify gates even when the biaxial map contains multiple, highly dense cell clusters, like in the side scatter-area (SSC-A) by forward scatter-area (FSC-A) gate for lymphocytes (Fig. 6A, left). Because of the presence of a much denser region of debris, the semi-supervised algorithms always require some human-inspected hints to manually eliminate a certain region and give

Table 1 | Evaluation matrices for cell-wise binary label classification using mass cytometry data

	UNITO				Static Gate			
	Accuracy	Recall	Precision	F1	Accuracy	Recall	Precision	F1
Single cell gate 1	0.987	0.997	0.987	0.992	0.829	0.892	0.918	0.859
Single cell gate 2	0.975	0.993	0.976	0.984	0.813	0.823	0.875	0.827
Granulocyte gate	0.986	0.985	0.980	0.982	0.885	0.824	0.894	0.833
Mononuclear gate	0.990	0.982	0.975	0.977	0.932	0.788	0.840	0.789
Neutrophil gate	0.987	0.982	0.983	0.981	0.883	0.813	0.911	0.827
Eosinophil gate	0.998	0.863	0.772	0.790	0.995	0.632	0.602	0.516
T cell gate	0.994	0.984	0.990	0.973	0.957	0.785	0.832	0.792
B cell gate	0.998	0.972	0.959	0.963	0.993	0.779	0.826	0.780
Non-T-non-B cell gate	0.996	0.967	0.971	0.966	0.980	0.792	0.844	0.789
	FlowDensity				FlowSOM			
	Accuracy	Recall	Precision	F1	Accuracy	Recall	Precision	F1
Single cell gate 1	0.845	0.901	0.928	0.903	0.935	0.946	0.976	0.960
Single cell gate 2	0.778	0.789	0.927	0.843	0.922	0.982	0.925	0.951
Granulocyte gate	0.800	0.604	0.923	0.705	0.950	0.967	0.925	0.944
Mononuclear gate	0.945	0.804	0.969	0.868	0.968	0.975	0.906	0.936
Neutrophil gate	0.641	0.251	0.918	0.354	0.956	0.993	0.919	0.953
Eosinophil gate	0.991	0.201	0.402	0.225	0.994	0.930	0.484	0.585
T cell gate	0.961	0.773	0.967	0.837	0.983	0.961	0.930	0.936
B cell gate	0.988	0.695	0.824	0.722	0.994	0.995	0.819	0.894
Non-T-non-B cell gate	0.978	0.684	0.949	0.767	0.989	0.979	0.874	0.921
	Logistic Regression				DeepCyTOF			
	Accuracy	Recall	Precision	F1	Accuracy	Recall	Precision	F1
Single cell gate 1	0.951	0.973	0.969	0.970	0.839	0.974	0.859	0.904
Single cell gate 2	0.910	0.977	0.915	0.944	0.776	0.968	0.793	0.863
Granulocyte gate	0.937	0.985	0.899	0.938	0.768	0.971	0.692	0.789
Mononuclear gate	0.962	0.964	0.896	0.922	0.875	0.702	0.776	0.727
Neutrophil gate	0.938	0.991	0.893	0.937	0.758	0.972	0.677	0.776
Eosinophil gate	0.994	0.503	0.514	0.424	0.992	0	0	0
T cell gate	0.980	0.991	0.903	0.943	0.975	0.992	0.870	0.924
B cell gate	0.995	0.975	0.861	0.910	0.972	0	0	0
Non-T-non-B cell gate	0.983	0.878	0.888	0.863	0.915	0.279	0.363	0.289

an approximate position of where the actual lymphocyte cluster may present. With limited data available, the UNITO framework can easily overcome this limitation and automatically eliminate the debris, disregard other clusters around, and only keep the lymphocytes cluster. The consistency of close-to-manual performance on subsequent gating proves UNITO’s ability to learn gating patterns for different cytometric data modalities. UNITO also maintains a good performance when comparing the proportion of target cells in the UNITO gating and manual gating (Fig. 6B), in which all Pearson correlation coefficients are larger than 0.9. Similarly, UNITO also displayed the closest-to-human gating performance compared with other methods based on the number of gated cells (Fig. 6D).

UNITO is capable of handling strong batch effects

As an image segmentation method, UNITO identifies cell populations based on their location relative to other populations, regardless of changes in absolute location or shape of distribution. This makes UNITO inherently robust to batch effects. For example, despite the large technical variability in DNA intercalator channels for mass cytometry data (Fig. 1), UNITO can gate these channels with significantly higher accuracy than any other method (Table 1).

To test robustness to technical variability more formally, we simulated a batch effect by randomly assigning mass cytometry files to two equal-sized batches, then adding noise to all files in one batch.

Gamma-distributed noise was used because it is non-negative, so adding it preserves the non-negative values measured in CyTOF. The noise was added prior to arcsinh-transformation, causing the negative peak of the transformed marginal distributions to shift right, while the positive peak was almost unaffected (Fig. 7A). By design, the noise affected cell types differently, so that it cannot be removed by normalizing the data. The batch effect was large enough to make affected cells stand out from their original selves in protein space, but not so large to make distinct cell types blend into each other (Fig. 7B, C).

All methods were then evaluated on the dataset with simulated batch effect. The supervised methods were trained on the same cross-validation paradigm and evaluated on held-out files from both batches. All methods had decreased performance on this noisy dataset compared to the original one. But UNITO decreased the least, maintaining an F1 score of around 0.9 or above for all cell types except the comparatively rare Eosinophils, and the highest F1 score overall (Fig. 7D). The gates predicted by UNITO for files in the noisy batch automatically shifted right to account for the change in data distribution, without any need for manual adjustments or data alignment procedures (Fig. 7D).

UNITO identifies immune health signatures by subtyping cell populations

Identifying similarities or differences among subjects in certain cell subpopulations is a common goal with the task of manual gating. To

Table 2 | Evaluation matrices for cell-wise binary label classification using flow cytometry data

	UNITO				Static Gate			
	Accuracy	Recall	Precision	F1	Accuracy	Recall	Precision	F1
Lymphocyte gate	0.981	0.996	0.966	0.981	0.977	0.999	0.957	0.977
Single cell gate	0.975	0.997	0.952	0.974	0.955	0.967	0.937	0.946
CD3 gate	0.988	0.999	0.966	0.982	0.963	0.954	0.930	0.933
CD 4 gate	0.994	0.989	0.975	0.977	0.980	0.955	0.946	0.939
CD 4 naïve gate	0.991	0.986	0.924	0.945	0.982	0.950	0.877	0.897
CD 8 gate	0.994	0.998	0.944	0.968	0.989	0.944	0.912	0.919
CD 8 naïve gate	0.997	0.992	0.986	0.935	0.993	0.933	0.759	0.803
	FlowDensity				FlowSOM			
	Accuracy	Recall	Precision	F1	Accuracy	Recall	Precision	F1
Lymphocyte gate	0.695	0.900	0.650	0.743	0.927	0.900	0.950	0.924
Single cell gate	0.800	0.799	0.810	0.791	0.937	0.905	0.962	0.932
CD3 gate	0.914	0.783	0.981	0.860	0.980	0.978	0.961	0.969
CD 4 gate	0.946	0.796	0.981	0.870	0.989	0.992	0.956	0.973
CD 4 naïve gate	0.908	0.237	0.964	0.269	0.975	0.981	0.812	0.884
CD 8 gate	0.974	0.715	0.983	0.817	0.984	0.983	0.847	0.908
CD 8 naïve gate	0.990	0.702	0.851	0.719	0.996	0.913	0.870	0.887
	Logistic Regression				DeepCyTOF			
	Accuracy	Recall	Precision	F1	Accuracy	Recall	Precision	F1
Lymphocyte gate	0.897	0.868	0.917	0.890	0.173	0.280	0.228	0.240
Single cell gate	0.903	0.879	0.915	0.895	0.264	0.211	0.237	0.205
CD3 gate	0.959	0.936	0.941	0.938	0.656	0.095	0.080	0.086
CD 4 gate	0.981	0.976	0.937	0.956	0.767	0.100	0.071	0.082
CD 4 naïve gate	0.984	0.953	0.899	0.923	0.880	0.097	0.068	0.078
CD 8 gate	0.986	0.928	0.909	0.917	0.909	0.032	0.038	0.033
CD 8 naïve gate	0.996	0.925	0.859	0.886	0.977	0.006	0.023	0.008

Table 3 | Evaluation matrices comparing UNITO with PeacoQC on Lymphocyte gate in flow cytometry

	UNITO				PeacoQC			
	Accuracy	Recall	Precision	F1	Accuracy	Recall	Precision	F1
Singlet gate	0.981	0.996	0.966	0.981	0.977	0.999	0.957	0.977

test whether the UNITO framework could support a workflow in which cell populations were compared between two groups, we analyzed their statistical difference between peripheral blood from healthy donors and patients with acute COVID-19 using the auto-gated cell data as well as the consensus-gated labels. An interesting property of UNITO is its ability to gate cells in a sequential manner, so the statistical tests can be applied either to the end nodes of the gating hierarchy tree or the middle nodes. Here we selected 6 gates after the pre-gating tasks to perform a t-test over the COVID population and healthy population (Fig. 4C, Fig. S7). We found that there are significantly different cell proportions between the two populations such as neutrophil, granulocyte, as well as lymphocyte, which agrees with previous literature^{26–29}. The results show that the UNITO output after the pre-gating tasks can yield the same results among different cell sub-population tests compared to manual gating. In addition, the UNITO pipeline provided enough statistical power to distinguish group disparity between COVID and healthy subjects in many gates that are well-known to be drastically increased when facing immune diseases.

In addition, with the flow cytometry data, we also have the subject phenotype of young and old subjects (Fig. 6C, Fig. S8). While it is more difficult to observe the disparity among cell populations from young and old people than between COVID-19 patients and healthy donors, both UNITO (p-value 3.91e-6) and manual gating (p-value 9.98e-7) observe a significantly higher proportion of CD8 naïve cells in young

donors as previously reported^{30,31}. UNITO again provides enough statistical power as well as an equivalent gating p-value, even for the other cell types that were not significantly different between young and old donors.

Discussion

The rapid development of single-cell technologies has enabled large-scale and high throughput data collection, but at the same time raised challenges for analyzing such large amounts of data. The size of data not only increased within a single sample (number of cells), but we are also accumulating more donors over time. In this study, we first described a common challenge in cytometric pre-gating tasks due to the high technical and biological variability across subjects. Most state-of-the-art unsupervised auto-gating methods typically assume the data is already cleaned from the pre-gating stage and do not explicitly handle events such as debris and doublets, which are unpredictable and difficult to model. Other auto-gating methods such as FlowDensity also require human prior knowledge as hints for the model to target the position or percentage of the desired cell subpopulation. There are also supervised methods that predict cell types using deep neural networks. However, treating every single cell as one data point will require large computational resources to both train the model and predict incoming data. The proposed framework UNITO transformed the numerical classification of protein expression data into an

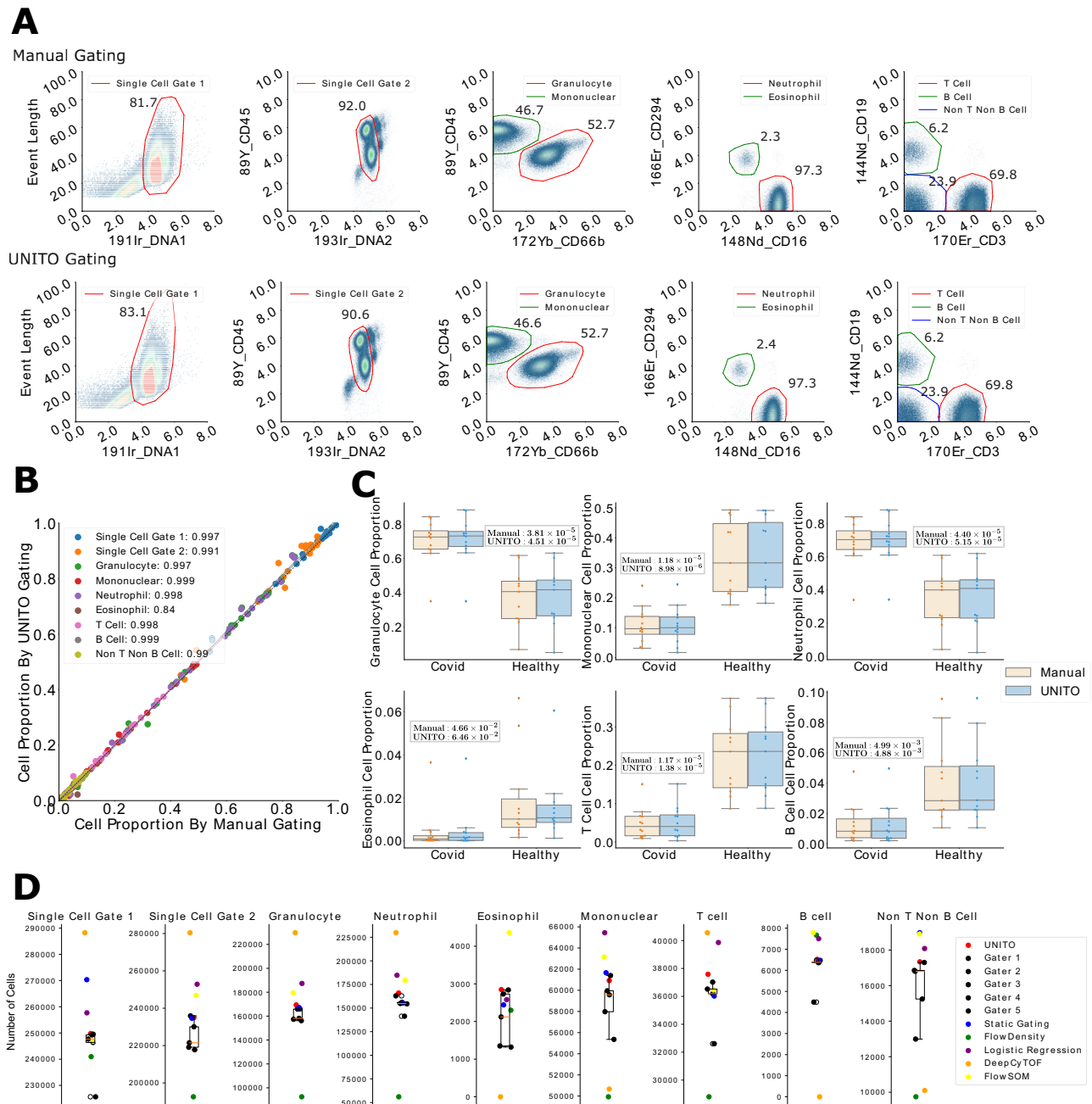


Fig. 4 | Comparison between UNITO automated gating and manual consensus gating (mass cytometry). **A** Density map and target cell annotation for mass cytometry data. The first row shows the ground truth annotation for one validation subject in the sequential prediction task, and the second row shows the predicted decision boundary from the reconstruction of the UNITO output. Each column represents a single level in the gating hierarchy, and gates within the same level are plotted together, such as granulocyte and mononuclear. **B** Visualization of the proportion of cells from UNITO and manual consensus gating with all gating tasks together. **C** Statistical comparison validated on the COVID-19 vs Healthy mass cytometry data (Top row: Granulocyte, Lymphocyte, Neutrophil. Second row: Eosinophil, T cell, B cell). The two-sided t-test between COVID-19 (12 samples) and Healthy (11 samples) group was applied on selected gates and P-values were reported for both manual gating and UNITO gating. The box plot in both panel C

and D display data distribution through its minima and maxima (extreme values excluding outliers), center (median), bounds of the box (interquartile range from the 25th to the 75th percentile), and whiskers (extending to the most extreme data points within 1.5 times the interquartile range). For each gating, the plot is separated by subjects' health condition while the results from manual gating and UNITO for the same population are put next to each other for easier observation.

D Disparities between automated methods and human gaters. Each dot represents the number of cells predicted to belong to a certain population by a given method, averaged across all samples, and the box plots summarize the distribution of values coming from manual gating (sample size of 5 manual gaters). UNITO gating (red) is consistent with the range of values provided by human gaters (black). Source data for panels B, C, and D are provided as a Source Data file.

image-based semantic segmentation task. In this case, by using the convolutional kernel to target the densest region in the cell density plot, the UNITO framework guaranteed the translational invariance property of cell clusters and a global view of the bivariate cell density

distribution. With such properties, the proposed method can be applied and extended to any cytometric data including mass cytometry and flow cytometry to automatically gate for multiple purposes, such as removing debris, doublets, and gating other cell subtypes. By

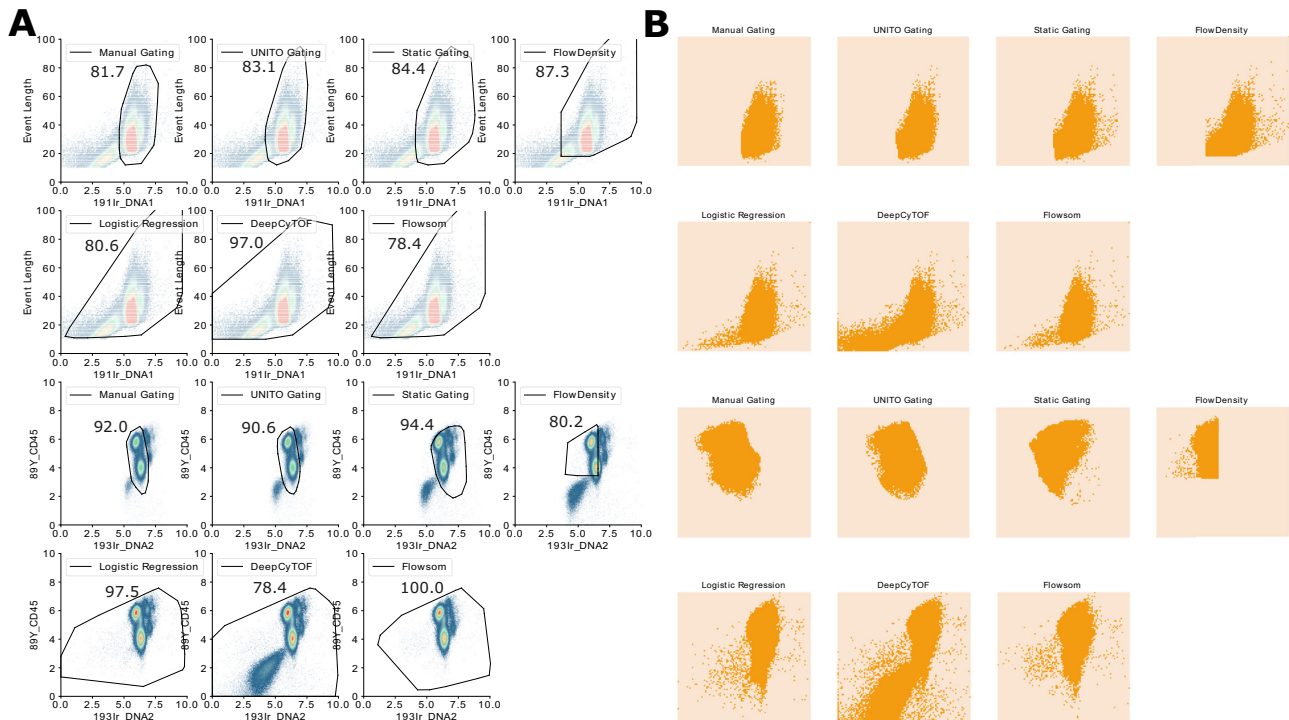


Fig. 5 | Gating results from different automated methods. **A** shows the convex hull based on the cell-level cell type assignment, and **B** shows the cells classified at a raw binary mask level on the density plot. For each panel, the top two rows present

gating results for single cell gate 1, and the bottom two rows show the results for single cell gate 2.

converting the protein expression matrix to the density map, UNITO further allows the incorporation of more training data and is capable of learning the protein expression behavior in a systematic pattern across subjects. One advantage of UNITO is that since the prediction is performed in each single sample file by file, it can query a huge number of samples parallelly with very low computational cost, even computing with CPU is sufficient for prediction. In addition, the computing speed of UNITO will also not be affected by the number of cells in each sample, which guarantees its efficacy even if the throughput of the cytometric experiments continues to grow in the future. Using UNITO for immunophenotyping in different protein panels would require separate training. However, cleanup gates such as Single cell gate 1 and 2 in CyTOF, and the Lymphocyte and Singlet gates in flow cytometry, are the same in most experiments. Therefore, researchers can use the models we already trained to perform cleanup on their own data.

We believe that unsupervised methods like clustering are extremely useful for exploratory analysis, such as understanding how different cell types appear when profiled with different cytometry panels, or uncovering new cell states or unknown sources of variability in the data. In contrast, a supervised method like UNITO is a better fit for what could be called a production environment, rather than an exploratory one. UNITO is intended for users who already understand their panel well, and would like a supervised, automated computational method that can label cell types in a consistent way across many experiments done with the same panel. For example, UNITO may be a good fit for multi-site studies or clinical cytometry settings. Among supervised methods, some of which are included in this manuscript for comparison, UNITO is distinguished by closely following the hierarchy of cell types determined by manual gating, which facilitates interpretation by immunologists. In summary, the strengths of UNITO are robustness and interpretability, rather than discovery of new phenotypes in an exploratory setting.

The UNITO performance on the sequential pre-gating and gating tasks for immune cells proved the similar-to-human boundaries of the

UNITO framework by visually comparing the boundaries of human annotation and UNITO prediction on the density map. One of the biggest advantages of the UNITO framework is that it also gives the boundary of the target cell population, which enables the interpretability of the gating results and allows inspection of intermediate gating steps. While the cell-level prediction using logistic regression or deep neural networks and unsupervised clustering also gives good evaluation scores, the reliability of the gating results with biological knowledge is still in debate. We visualized the gating results with an example subject for the first two single-cell gates in the mass cytometry data (Fig. 5), and we can clearly see that only UNITO has the closest contour compared to the manual gating. The static gating has a similar shape, but the position is off due to the biological variability of the subject and the gating position is fixed. The idea of FlowDensity is to segment the entire cell population into four quadrants and based on the position parameter and percentage to draw the boundaries. However, in reality the percentage can vary a lot across subjects (Figs. S2, 3, 7, 8). Another alternative called flowLearn¹⁹ uses an algorithm instead of a hyper-parameter to find the best cutoff and perform the density-based gating. Nevertheless, the implementation of only relying on peak separation and using one protein measurement at a time makes automatic pre-gating difficult. While UNITO has the highest F1 score for all gates, FlowSOM and logistic regression also have good performance. However, FlowSOM and logistic regression directly label cells with their final cell types, whereas UNITO emphasizes the hierarchical nature of cell phenotypes. For example, UNITO allows cells to be in the T cell gate, but not in the downstream CD4 or CD8 gates. In contrast, FlowSOM may or may not capture separate CD4-CD8- or CD4 + CD8+ clusters, and logistic regression requires these populations to be explicitly modeled. Although not shown in this article, UNITO's hierarchical approach also allows cells to be in multiple terminal gates: for example, a CD4 T cell can have memory subtype "Effector memory" and polarization "Th1", capturing multiple facets of cell identity. The visualization of gating results on the density map for

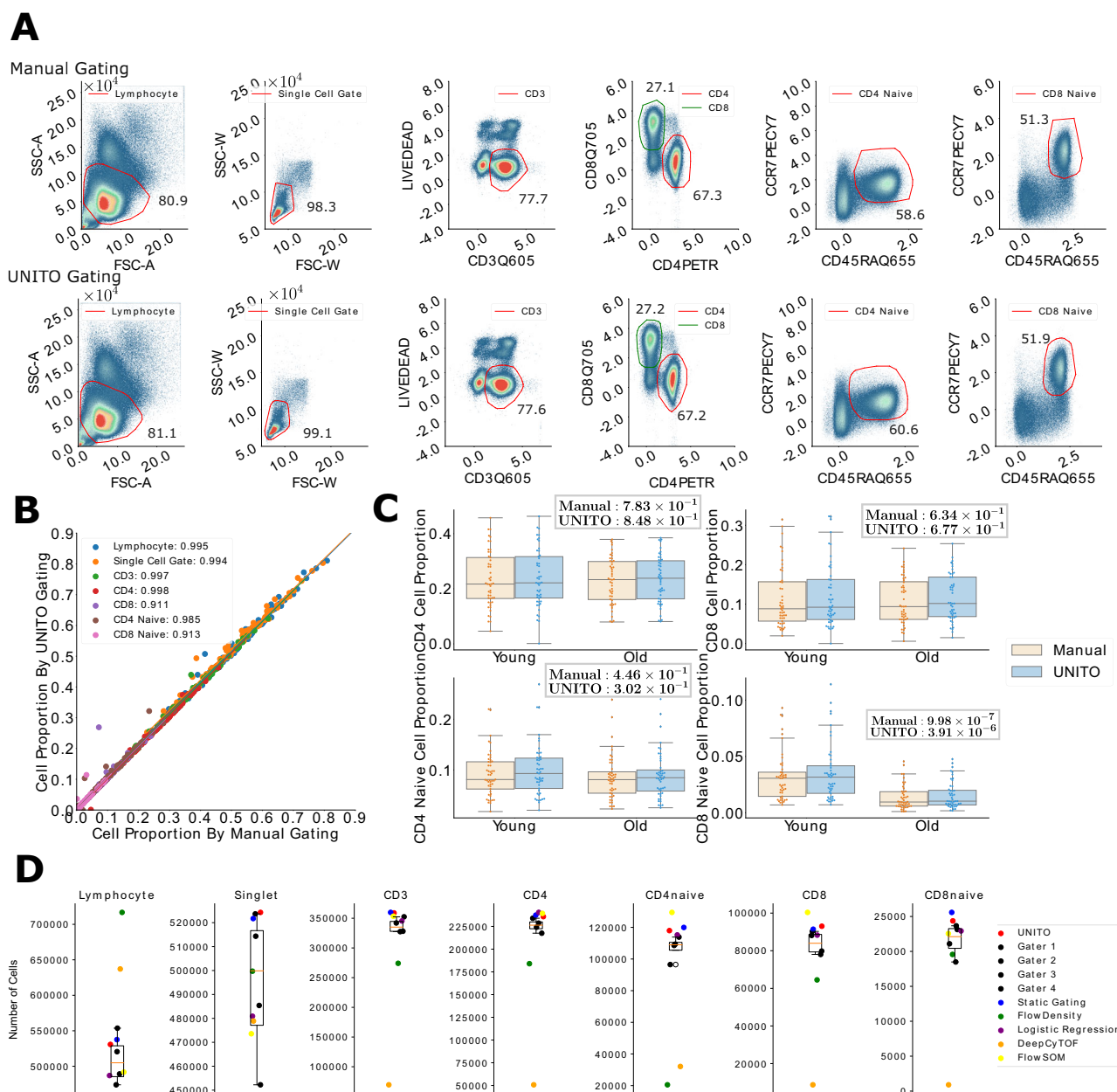


Fig. 6 | Comparison between UNITO automated gating and manual consensus gating (flow cytometry). **A** Density map and target cell annotation for flow cytometry data. The first row shows the ground truth annotation for one validation subject in the sequential prediction task, and the second row shows the predicted decision boundary from the reconstruction of the UNITO output. Each column represents a single level in the gating hierarchy, and gates within the same level are plotted together, such as CD4 cells and CD8 cells. **B** Visualization of the proportion of cells from UNITO and manual gating with all gating tasks together. **C** Statistical comparison validated on the Young vs Old flow cytometry data (Top row: CD4, CD8. Second row: CD4 naive, CD8 naive). The two-sided t-test between Young (49 samples) and Old (48 samples) group was applied on selected gates and P-values were reported for both manual gating and UNITO gating. The box plot in

both (C, D) display data distribution through its minima and maxima (extreme values excluding outliers), center (median), bounds of the box (interquartile range from the 25th to the 75th percentile), and whiskers (extending to the most extreme data points within 1.5 times the interquartile range). For each gating, the plot is separated by subjects' age condition while the results from manual gating and UNITO for the same population are put next to each other for easier observation. **D** Disparities between automated methods and human gaters. Each dot represents the number of cells predicted to belong to a certain population by a given method, averaged across all samples, and the box plots summarize the distribution of values coming from manual gating (sample size of 4 manual gaters). UNITO gating (red) is consistent with the range of values provided by human gaters (black). Source data for (B–D) are provided as a Source Data file.

logistic regression and DeepCyTOF almost included the entire coordinate space, indicating that cell-level predictions based on high dimensional protein measurements are difficult to validate in the bivariate setting familiar to immunologists. A similar situation also happens in the FlowSOM clustering, where the convex hull included a large region with a noisy pattern in the binary mask visualization (Fig. 5). PeacoQC is a tool specifically designed for flow cytometry data

quality control and performs well on singlet gating. However, compared to PeacoQC, UNITO not only achieves a higher F1 score but is also more versatile across various gating tasks. A simple observation we can draw from this comparison is that prediction over bivariate image outperforms prediction on single-cell tabular data in both accuracy and interpretability, and even static gating showed its capacity to provide a good gating prediction. If there are large batch effects

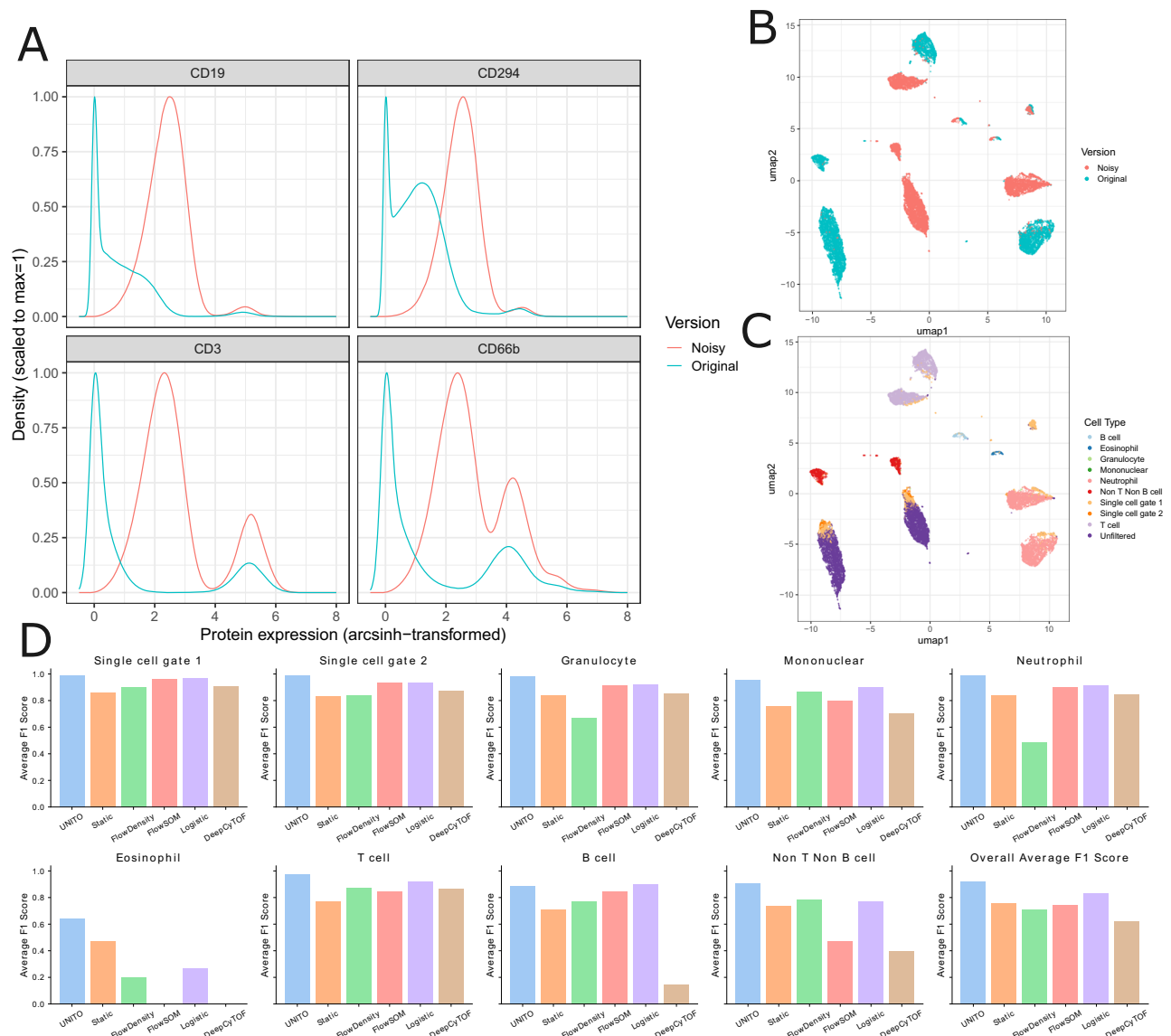


Fig. 7 | Batch effects simulation and evaluation. **A** Distribution shifts on selected protein channels, indicating the simulated batch effects caused significant shifts in protein expression. **B** UMAP visualization of protein expression from original and noisy data combined. **C** Same UMAP visualization showing cell-type annotation

defined by consensus gating. **D** F1 evaluation of 9 gates from mass cytometry for all comparison methods, with an overall F1 score comparison at the end. Source data for (A, D) are provided as a Source Data file.

in the data, the performance of static gating will decrease, but UNITO is still robust to such variability.

While methods like FlowSOM and logistic regression are effective approaches for identifying final cell types and have proven useful in many analytical scenarios, they are not inherently designed to handle batch effects. Their performance may be improved by applying batch effect correction techniques such as CytoNorm³² or CyCombine³³. In contrast, UNITO's robustness to variations and batch effects makes it especially beneficial for large-scale cytometric data processing. Such capacity to natively address batch variability without additional pre-processing or alignment makes UNITO an efficient and robust choice for high-throughput, non-exploratory, as well as production environment applications.

The key goal for all of the pre-gating and gating tasks is to discover the hidden information embedded in the cell subpopulations. The downstream analysis usually reveals statistical differences either among cell subtypes or subject groups. With prior knowledge of the biological difference between COVID-19 patients and healthy donors,

as well as between young and old people, we can further use the statistical power of the gating results to validate the method's performance. The proportion of cell types selected by prior knowledge in UNITO gating confirmed a significant increase in granulocytes and neutrophils with a significant decrease in lymphocytes, T cells, and B cells, therefore validating the efficacy of UNITO. The same experiment on the flow cytometry data in CD8 naïve T cells also agreed with biological prior knowledge. This downstream analysis is not only used for confirming the gating results from UNITO, but also serves as a functionality to explore the group-level difference for scientific discovery purposes.

There are certain limitations to this study. The model may not get very good performance in certain cases where cells are extremely rare, such as eosinophils in certain samples. Even though the UNITO framework still has the highest performance for eosinophil prediction, we found one file in which there were almost no eosinophils. In this case, even in the manual gating, eosinophils can barely be seen. For such situation, one mitigation is to not only pre-filter the cells by the

previous gating level, but also pre-filter the other cell types in the current gating level. For instance, after we filtered out the non-granulocyte cells, we can add additional steps to filter predicted neutrophil cells, leaving only eosinophil and non-neutrophil-non-eosinophil cells in the data to segment the eosinophils. Moreover, we only tuned hyperparameters for the learning rate and batch size when designing UNITO; more advanced deep learning techniques such as adding self-attention blocks or using different image segmentation backbones may improve the performance.

To summarize, we present a framework that can automatically perform pre-gating and gating tasks for both mass cytometry and flow cytometry with close-to-manual performance. We validated that autogating by bivariate images outperforms gating on cell-level protein expression data, and UNITO can further provide the convex boundary on the density map for biological interpretation. In addition, UNITO predictions are easy to use in downstream statistical analysis for cell type-phenotype exploration, and yield results that are as statistically significant as those from manual gating. We then showed that the framework can be adapted to any subsequent gating tasks in a sequential manner while still maintaining high performance. In order to make the UNITO framework more accessible to a broader range of users, there is no requirement to tune the model or choose any format of parameters in addition to the gate settings. The only input needed for UNITO automatic gating is the measurement of two selected cytometric channels. If the user wants to train a model by themselves, they also do not have to worry about preparing the training mask, since all processes are automatic and within the UNITO framework. To enhance the accessibility and user-friendliness of UNITO, our future work includes the development of a web-based or software-based iteration taking raw FCS files as input, which is in progress with the Pennsieve Data Management Platform from the University of Pennsylvania.

Methods

UNITO architecture

In this study, our UNITO framework was adapted from the architecture of the UNet model. UNet is a convolutional neural network (CNN) architecture that was first introduced by Ronneberger et al.³⁴. The purpose of its original design was to perform semantic segmentation of medical images, which is the task of assigning a label to each pixel and thus defining the region of interest. The UNet architecture consists of two parts: an encoder (the contracting path) and a decoder (the expansive path). Both the pooling and up-sampling parts have a large number of feature channels, which allow the network to propagate context information to higher resolution layers which results in the expanding part being symmetric to the contracting path and yields a u-shaped architecture. The architecture of the encoder resembles that of a traditional CNN, which consists of a series of convolutional and pooling layers. The convolutional layers extract high-level features from the input image and the pooling layers reduce its spatial resolution. The decoder consists of a series of convolutional and transposed convolutional layers. The transposed convolutional layers upsample the feature maps to a higher resolution, which is achieved by interpolating the in-between pixels. The convolutional layers then process the upsampled feature maps to extract high-level features. At each decoding stage, the upsampled feature maps are concatenated with the corresponding feature maps directly passed from the encoding stage. These skip connections allow the decoder to use the contextual information from the encoder, which helps recover the lost spatial resolution.

We adapted the UNet architecture in our study to allow binary classification of each gate, and the output bivariate mask was used to produce the final segmentation image. Overall, the network consists of convolutional layers with hidden sizes from 1, 64, 128, 256, 512, and 1024, and then the transposed convolution will upsample the image

back to the same size by following the same hidden size order but in the opposite direction. The skip connections provide the decoder with the original contextual information of the image, allowing U-Net to produce highly accurate segmentation results. In addition, U-Net has shown excellent performance in handling small data sets, which makes it particularly useful in medical applications, where the sample size is often limited.

Cytometric gating structure

The main objective of UNITO is to achieve a similar boundary compared to human annotation with the ability to handle batch effects and heterogeneity between different subjects. Since subsequent cell-type gating tasks have the same setting as the pre-gating task, but with less variability, the original application of pre-gating cytometric data can be further extended to automatically gating cell subtypes in mass cytometry, as well as to gating flow cytometry data. Figure S1A shows all 5 levels of pre-gating and gating tasks (9 gates in total) of mass cytometry gates validated using UNITO. Each level from the hierarchy uses different pairs of channels from protein measurements, and each gate will undergo separate training. Within the same hierarchy, the cells are gated on the same space coordinate with the same pair of protein expressions, such as granulocyte and mononuclear. UNITO is validated by sequentially performing gating tasks on single cell gate 1, single cell gate 2, granulocyte, mononuclear, neutrophil, eosinophil, T cell, B cell, and Non T non B cell to achieve the pre-defined gating hierarchy. In addition to mass cytometry gating, Figure S1B shows the gating procedure for flow cytometry. Lymphocyte gate in flow cytometry data is still difficult to automate due to its high density of debris and noises from other cell populations. Thus, it usually requires manual gating or prior knowledge such as input of a hard threshold and approximate position based on the experiments. We validated the performance of UNITO starting from the lymphocyte gate and examined all gates by sequence of lymphocyte, singlet, CD3, CD4, CD8, CD4 naïve, and CD8 naïve cells.

Construction of bivariate density map and mask

Here we define our cytometric data $X_c = \{X_1, X_2, X_3, \dots, X_n\}$ where each X_i represent a single subject. Let $X_i = [x_{uv}] \in \mathbb{R}^{m \times n}$ be the protein expression matrix, where m is the number of cells in the experiments and n is the number of channels collected. Usually, a specific cell distribution mostly varies among two protein measurements, and using two channels for manual gating is easier to visualize and inspect. Therefore, two selected cytometric measurements are normalized and rounded as a constraint on the boundaries for the density map and mask. The normalization process is defined that for each x_{uv} in the cytometric measurement X :

$$\bar{x}_{uv} = \frac{x_{uv} - \min(x_v)}{\max(x_v) - \min(x_v)} \quad (1)$$

Density maps of the normalized cytometric data were created as training images by counting the number of cells with the same pair of normalized measurements and feeding the number into the corresponding coordinate in the matrix space. In this case, the density matrix has corresponding shape of 101 by 101 representing the normalized measurement. This transformation not only made the process of creating density maps easier, but also allowed faster training and prediction time because this step reduced the training input from a data matrix of millions of entries to a 101 by 101 matrix. We also experimented with different image sizes for UNITO input data generation (Fig. S10), and we found that in general larger image sizes would decrease the performance. Even though a larger image will provide a higher resolution of the bivariate density image, it will make the density representation sparser, and thus make it more difficult to recognize the target cell population. In addition, smaller image sizes

tend to have relatively stable performance. With the 101 matrix size having one of the best performance and straightforward concepts of design, we set the default value of UNITO input to 101. The training data also provides gating labels for each cell inferred from manual annotation, and the corresponding binary masks were created from the manual labels as the training labels. If the cells are identified as the target population, the corresponding coordinate will be marked as 1, creating a 101 by 101 binary image. In addition, a convex hull algorithm was further applied on the binary mask to smooth the boundaries using the `ConvexHull` function from the Python package `SciPy`³⁵. To provide flexibility in handling different feature shapes, we implemented a `true/false` parameter for the convex function. When set to “true,” the convex hull will be applied to smooth boundaries; when set to “false,” only blank pixels within the mask will be filled, while the original boundaries are preserved. This option allows users to adapt the filling behavior based on the specific morphology of their data, ensuring more accurate feature representation.

UNITO training and cell membership prediction

The UNITO was trained using the PyTorch framework in Python. We searched hyperparameters learning rate and batch size for each gate, and we also checked convergence during each training process. Since UNITO follows a sequential gating process, each individual sample requires its own prediction to prepare the training data (in-gate cells) for the subsequent gate. To accommodate this, we implemented a cross-validation approach that iteratively predicts across n batches. In this paradigm, each batch is used in the training data $n-1$ times and serves as the validation data once. Further details on the cross-validation setup can be found in the supplementary material. The Adaptive Moment Estimation Optimizer³⁶ or in short, ADAM optimizer is a popular optimization method that allows the learning rate to adjust to the gradient change during the training process adaptively. This along with the popular Binary Cross-Entropy Loss function with Logits (`BCEWithLogitsLoss`) is used in the binary classification of the pixels. The Binary cross-entropy Loss can provide a good gradient calculation for optimization by combining the sigmoid function and binary cross-entropy loss (`BCELoss`) into a single function and is more numerically stable than using a plain sigmoid followed by a `BCELoss`.

$$\text{BCEWithLogitsLoss} = -(y \cdot \log(p) + (1 - y) \cdot \log(1 - p)) \quad (2)$$

Once we have the trained model, we can use the pipeline to predict the binary labels for each cell. Given a target subject, we first use its normalized cytometric data to create its density map corresponding to its gate. This density map also has a resolution of 101 by 101 and is obtained using the above-described procedure. It is then passed into the trained model to generate the predicted mask for the assigned gate. This binary mask is then interpolated with the protein measurement to generate the binary label for each cell. Specifically, each cell corresponds to a pixel on the binary mask; if a pixel is predicted as in-gate, all the cells corresponding to the selected measurements will be designated as in-gate, and vice versa. Next, we filter out all the cells that are not predicted to be in this gate, and we use the remaining data (in-gate cells for the current gate) to create the density map corresponding to the next gate. Again, the pre-processing and training procedure will be repeated for the next gate using the filtered cytometric data. While all gates are trained and all subjects are predicted, we evaluated the performance of each gate and visually inspected the bivariate polygons on the density map. We can also use intermediate or final gate results to perform downstream statistical analysis. The details of the UNITO algorithm are summarized in **Algorithm 1**.

One advantage of UNITO in automatic gating is that it does not require any human input after training. The framework will take

ungated files as input, and convert the cytometric data into bivariate plots. The prediction stage only takes the density image as input and outputs a corresponding mask. Then the post-processing step in UNITO will interpolate the mask back to the cell type outcome as the final output of UNITO. In addition, the training process for UNITO is also straightforward and automatic, the framework handles all the image conversion and interpolation. The only difference between the input of training and prediction is that the training data has to include an additional column of binary labels, indicating the in-gate and out-of-gate status for each cell. These binary labels can be obtained from the output of manual gating platforms, and we also provided an example script to retrieve the output from the Omiq platform. As shown in Supplementary Table S3, even though UNITO does not have the fastest run time per gate, it requires only seconds to compute each gate, which is competitive with the fastest methods. Moreover, a significant advantage of UNITO is that it operates on each sample independently, making it highly amenable to parallelization. This means that, in practice, the total runtime can be substantially reduced by distributing samples across multiple processors or computational nodes, enabling efficient processing even for high-throughput applications.

Statistics and reproducibility

No statistical method was used during the gating process. T-test was applied to evaluate the group difference between Covid (12 samples) and Healthy (11 samples) for mass cytometry data, and between Young (49 samples) and Old (48 samples) for flow cytometry data. Only baseline samples from the dataset were selected to avoid data leakage during the training process. After the baseline samples were selected, no other data were excluded from the analyses.

Mass cytometry data collection and processing

Human subjects were enrolled with approval of the University of Pennsylvania Institutional Review Board (COVID-19 Vaccine study IRB no. 844642; Acute COVID-19 study IRB no. 808542). All participants or their surrogates provided informed consent in accordance with protocols approved by the regional ethical research boards and the Declaration of Helsinki.

We used two independent datasets consisting of single-cell protein expression data profiled by mass cytometry (Table 4). Both datasets were collected by the Institute for Immunology and Immune Health at the Perelman School of Medicine. For the first (“COVID Vaccine”) dataset, whole blood was obtained from 40 healthy subjects at four timepoints during the course of two-dose mRNA vaccination against COVID-19 and cryopreserved. (T1 = baseline, T2 = one week after the first dose, T3 = prior to the second dose, T4 = one week after the second dose). Only the baseline sample was used, in order to avoid information leaks across training. For the second (“Acute”) dataset, whole blood was obtained from 23 subjects, among whom 12 had acute COVID-19 symptoms and 11 were healthy donors. Samples collected for the COVID-19/healthy dataset were used fresh.

Both the fresh samples from the Acute dataset and the frozen samples from the Vaccine dataset were stained with the Maxpar Direct Immunophenotyping Assay (MDIPA) and run on a CyTOF2 instrument. MDIPA is a single-tube cocktail manufactured by Standard BioTools, which uses standardized clones and dilutions performed by the manufacturer. No additional conjugation or titration is necessary. Samples were stained, stored, and prepared for acquisition as previously described³⁷. MDIPA measures a panel of 30 proteins used for broad characterization of immune phenotypes in whole blood, alongside other channels for control and data cleaning. Raw CyTOF data was transformed using an asinh transformation with a cofactor of 5. Further, a standard data cleaning procedure with manual gating was performed using the OMIQ platform, to remove beads, dead cells, platelets, debris, and other anomalous events, followed by manual

Table 4 | Demographic details of three datasets

Mass Cytometry Dataset			Flow Cytometry Dataset	
COVID Vaccine Dataset		COVID Acute Dataset	Young vs Old Dataset	
		COVID subjects	Healthy subjects	
Subjects	40	12	11	49
Cells	321k ± 156k	304k ± 228k	339k ± 215k	1092k ± 372k
				48
				953k ± 291k

The table shows the number of subjects and cells for each small group in three datasets. The total number of cells is reported in the format of (mean ± standard deviation).

gating on the subsequent cell type to generate the cell type labels for this study.

Flow cytometry data collection and processing

The young/old dataset was created by Rochester Human Immunology Center, David H. Smith Center for Vaccine Biology and Immunology, Rochester, NY (USA)³⁸. The purpose was to use SWIFT’s competitive clustering assignment method to measure the differences between PBMC sub-populations in Old/Young subjects. The Young/Old dataset contains 136 data files in total, with 97 unique subjects; similarly to the COVID Vaccine dataset, only one file from each subject was used, to avoid information leaks. Within those subjects, there are 49 samples from young donors (average age of 39) and 48 samples from old donors (average age of 67). The dataset was downloaded from the publicly available repository (<https://flowrepository.org/id/FR-FCM-ZZGS>). The data cleaning and labeling by manual gating were also done using the OMIQ platform.

Manual gating and consensus

The mass cytometry and flow cytometry data were manually gated by 5 and 4 independent annotators, respectively (see Fig. S4 for mass cytometry data and Fig. S5 for flow cytometry data). The final version of the label used for training and validation was generated by the consensus (majority vote) of the gating output from all annotators. Specifically, for each cell, we calculated the highest frequency of the cell type assignment from the different labels and saved the results as our final consensus voting label. Figure 1 and Fig. S6 show the correlation between consensus voting labels with manual gating from each annotator.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

All datasets used in this study are publicly available and have been uploaded to the Pennsieve platform (<https://doi.org/10.26275/864r-dv00>), and all relevant source data are provided with the paper. In addition, the mass cytometry data used in this study are available in the FlowRepository database under accession code [FR-FCM-Z8XU](#). The flow cytometry data used in this study are available in the FlowRepository database under accession code [FR-FCM-Z8P9](#). This includes both mass cytometry and flow cytometry data, along with manual gating results from each individual annotator and the consensus gating outcomes. All data can be accessed via the provided link. Source data are provided with this paper.

Code availability

The code for data preprocessing, UNITO training, and inference is available at (<https://github.com/KyleeCJ/UNITO>, <https://doi.org/10.5281/zenodo.14474197>³⁹). The GitHub repository also includes a Jupyter notebook that guides users through each step of the UNITO workflow, along with a user-friendly interface for those preferring a non-coding approach to interact with UNITO.

References

1. Fulwyler, M. J. Electronic separation of biological cells by volume. *Science* **150**, 910–911 (1965).
2. Hulett, H. R., Bonner, W. A., Barrett, J. & Herzenberg, L. A. Cell sorting: automated separation of mammalian cells as a function of intracellular fluorescence. *Science* **166**, 747–749 (1969).
3. Gray, J. W. et al. Chromosome measurement and sorting by flow systems. *Proc. Natl Acad. Sci. USA* **72**, 1231–1234 (1975).
4. Baumgarth, N. & Roederer, M. A practical approach to multicolor flow cytometry for immunophenotyping. *J. Immunol. Methods* **243**, 77–97 (2000).
5. Bandura, D. R. et al. Mass cytometry: technique for real time single cell multitarget immunoassay based on inductively coupled plasma time-of-flight mass spectrometry. *Anal. Chem.* **81**, 6813–6822 (2009).
6. Tanner, S. D., Baranov, V. I., Ornatsky, O. I., Bandura, D. R. & George, T. C. An introduction to mass cytometry: fundamentals and applications. *Cancer Immunol. Immunother.* **62**, 955–965 (2013).
7. Spitzer, M. H. & Nolan, G. P. Mass cytometry: single cells, many features. *Cell* **165**, 780–791 (2016).
8. McKinnon, K. M. Flow cytometry: an overview. *Curr. Protoc. Immunol.* **120**, 5.1.1–5.1.11 (2018).
9. Adan, A., Alizada, G., Kiraz, Y., Baran, Y. & Nalbant, A. Flow cytometry: basic principles and applications. *Crit. Rev. Biotechnol.* **37**, 163–176 (2017).
10. Van Gassen, S. et al. FlowSOM: using self-organizing maps for visualization and interpretation of cytometry data. *FlowSOM. Cytometry A* **87**, 636–645 (2015).
11. Samusik, N., Good, Z., Spitzer, M. H., Davis, K. L. & Nolan, G. P. Automated mapping of phenotype space with single-cell data. *Nat. Methods* **13**, 493–496 (2016).
12. Qiu, P. et al. Extracting a cellular hierarchy from high-dimensional cytometry data with SPADE. *Nat. Biotechnol.* **29**, 886–891 (2011).
13. Yang, Z.-K., Pan, L., Zhang, Y., Luo, H. & Gao, F. Data-driven identification of SARS-CoV-2 subpopulations using PhenoGraph and binary-coded genomic data. *Brief. Bioinform.* **22**, bbab307 (2021).
14. Li, H. et al. Gating mass cytometry data by deep learning. *Bioinformatics* **33**, 3423–3430 (2017).
15. Cheng, L., Karkhanis, P., Gokbag, B., Liu, Y. & Li, L. DGCyTOF: deep learning with graphic cluster visualization to predict cell types of single cell mass cytometry data. *PLOS Comput. Biol.* **18**, e1008885 (2022).
16. Li, Y. et al. Deep cytometry: deep learning with real-time inference in cell sorting and flow cytometry. *Sci. Rep.* **9**, 11088 (2019).
17. 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. USA* **117**, 21373–21380 (2020).
18. Malek, M. et al. flowDensity: reproducing manual gating of flow cytometry data by automated density-based cell population identification. *Bioinformatics* **31**, 606–607 (2015).
19. Lux, M. et al. flowLearn: fast and precise identification and quality checking of cell populations in flow cytometry. *Bioinforma. Oxf. Engl.* **34**, 2245–2253 (2018).

20. Critical assessment of automated flow cytometry data analysis techniques | *Nature Methods*. <https://www.nature.com/articles/nmeth.2365>.
21. Understanding of a convolutional neural network | IEEE Conference Publication | IEEE Xplore. <https://ieeexplore.ieee.org/document/8308186>.
22. Gu, J. et al. Recent advances in convolutional neural networks. *Pattern Recognit* **77**, 354–377 (2018).
23. Emmaneel, A. et al. PeacoQC: Peak-based selection of high quality cytometry data. *Cytometry A* **101**, 325–338 (2022).
24. Monaco, G. et al. flowAI: automatic and interactive anomaly discerning tools for flow cytometry data. *Bioinformatics* **32**, 2473–2480 (2016).
25. Fletez-Brant, K., Špidlen, J., Brinkman, R. R., Roederer, M. & Chatopadhyay, P. K. flowClean: automated identification and removal of fluorescence anomalies in flow cytometry data. *Cytom. Part J. Int. Soc. Anal. Cytol.* **89**, 461–471 (2016).
26. Reusch, N. et al. Neutrophils in COVID-19. *Front. Immunol.* **12**, 652470 (2021).
27. Vitte, J. et al. A granulocytic signature identifies COVID-19 and its severity. *J. Infect. Dis.* **222**, 1985–1996 (2020).
28. Huang, W. et al. Lymphocyte subset counts in COVID-19 patients: a meta-analysis. *Cytom. Part J. Int. Soc. Anal. Cytol.* **97**, 772–776 (2020).
29. Illg, Z., Muller, G., Mueller, M., Nippert, J. & Allen, B. Analysis of absolute lymphocyte count in patients with COVID-19. *Am. J. Emerg. Med.* **46**, 16–19 (2021).
30. Hong, M. S., Dan, J. M., Choi, J.-Y. & Kang, I. Age-associated changes in the frequency of naïve, memory and effector CD8+ T cells. *Mech. Ageing Dev.* **125**, 615–618 (2004).
31. Pfister, G. et al. Naïve T cells in the elderly: are they still there? *Ann. N. Y. Acad. Sci.* **1067**, 152–157 (2006).
32. Van Gassen, S., Gaudilliere, B., Angst, M. S., Saeys, Y. & Aghaepour, N. CytoNorm: a normalization Algorithm for cytometry Data. *Cytometry A* **97**, 268–278 (2020).
33. Pedersen, C. B. et al. cyCombine allows for robust integration of single-cell cytometry datasets within and across technologies. *Nat. Commun.* **13**, 1698 (2022).
34. Ronneberger, O., Fischer, P. & Brox, T. U-Net: convolutional networks for biomedical image segmentation. in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015* (eds. Navab, N., Hornegger, J., Wells, W. M. & Frangi, A. F.) 234–241 (Springer International Publishing, Cham, 2015). https://doi.org/10.1007/978-3-319-24574-4_28.
35. Virtanen, P. et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* **17**, 261–272 (2020).
36. Kingma, D. P. & Ba, J. Adam: a method for stochastic optimization. *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings* (2014).
37. A streamlined whole blood CyTOF workflow defines a circulating immune cell signature of COVID-19 - Geanon - 2021 - Cytometry Part A - Wiley Online Library. <https://onlinelibrary.wiley.com/doi/full/10.1002/cyto.a.24317>.
38. Rebhahn, J. et al. Competitive SWIFT cluster templates enhance detection of aging changes. *Cytom. Part J. Int. Soc. Anal. Cytol.* **89**, 59–70 (2015).
39. Chen, J. et al. Automate d cytometric gating with human-level performance using bivariate segmentation. Zenodo, doi:10.5281/zenodo.14474198 (2024).

Acknowledgements

This work was supported in part by the following NIH grants: U01 AG068057 (L.S.) and the Glick COVID-19 research award (E.J.W.). We thank Takuya Ohtani and the CyTOF Core at the University of Pennsylvania for data acquisition.

Author contributions

J.C. conceptualized and developed the model, wrote the source code, and wrote the paper. N.M. coordinated the COVID-19 study, recruited subjects and collected clinical data. M.M., M. P., D.M., A.P., and A.G. collected and processed the data. J.C., M.I., Y.F., and Y.L. conducted relevant computational experiments. M.M. and M.I. developed the framework for aggregating and comparing results from multiple annotators. A.P., M.P., M.M., D.M., M.I. manually annotated data. P.O., S.G., K.H., J.B., J.W. (Junhao Wen), D.D., J.W. (Joost Wagenaar), E.J.W., A.G. provided critical revision to the manuscript. All authors reviewed the paper. L.S. supervised the study and reviewed the method and all the submitted materials.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41467-025-56622-2>.

Correspondence and requests for materials should be addressed to Li Shen.

Peer review information *Nature Communications* thanks Xi Peng and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. A peer review file is available.

Reprints and permissions information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025