Accurate automated diagnosis of B-acute lymphoblastic leukemia using deep learning and flow cytometry

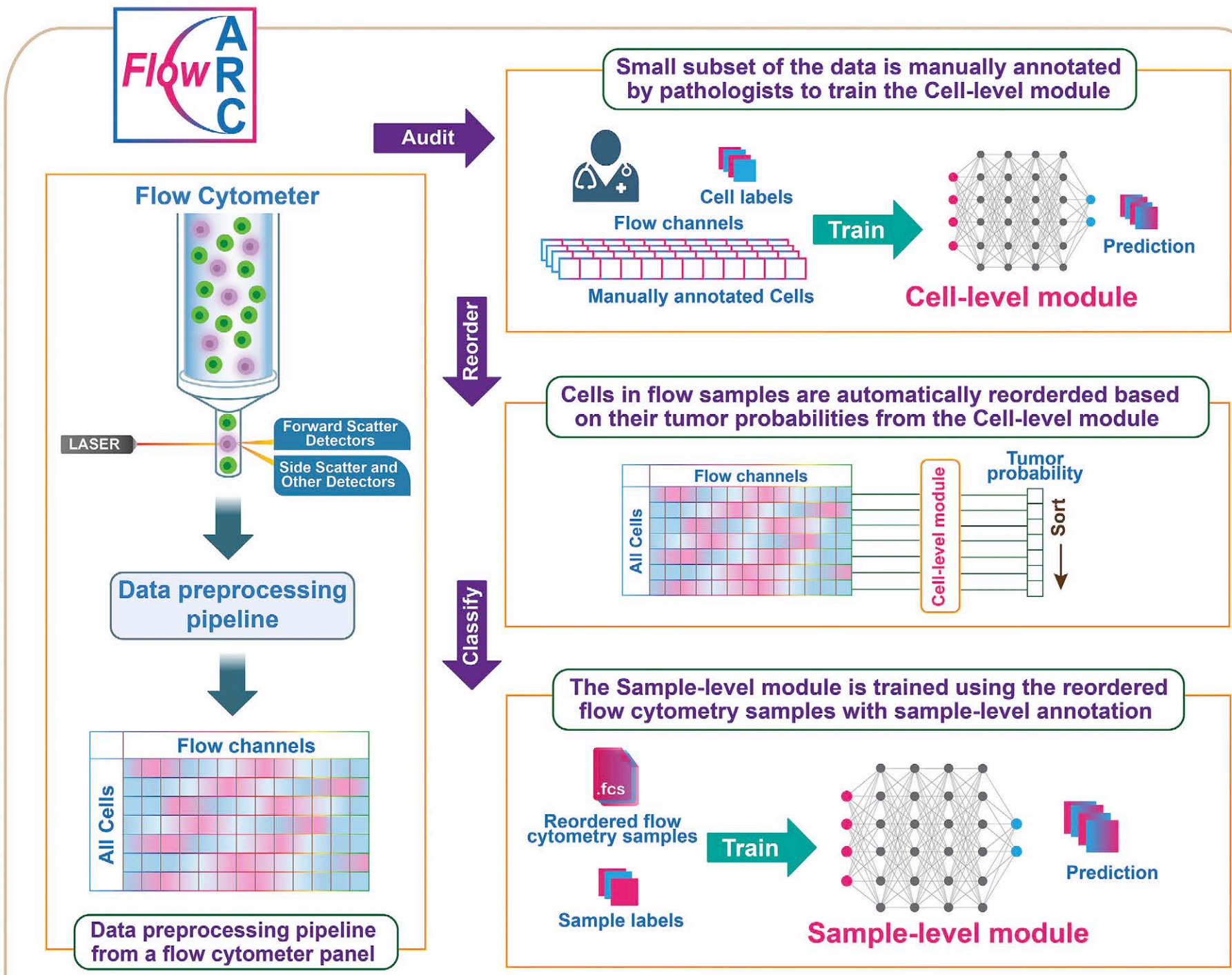
Multi-parameter flow cytometry (MFC) is an essential ancillary technique used in B-lymphoblastic leukemia / lymphoma (B-ALL) diagnosis and monitoring to identify abnormal B-precursor cell populations in patient specimens. Minimal (measurable) residual disease (MRD) analysis is essential for clinical management of B-ALL.[1] In experts' hands, MFC is a well-established, relatively inexpensive, and highly accurate technique for MRD with sensitivities as high as 2 in 10^6 cells.[2] Currently, accurate interpretation of MFC results primarily depends on the expertise of highly trained personnel. Despite panel and instrument standardization, interpretive challenges remain a barrier to broader utilization. For example, a recent study involving expert, high-volume laboratories participating in Children's Oncology B-ALL trials revealed substantial variability in distinguishing normal immature B cells from leukemic counterparts.[3] Automated analysis approaches may, in theory, improve the situation, but have yet to show sufficient accuracy to be impactful in routine clinical practice.

In recent years, automated MFC analysis for cell population identification and characterization has been demonstrated in research studies.[4] CellCNN, a supervised deep-learning model that utilizes annotated data, can identify rare disease-associated cells using a small flow cytometry dataset.[5] Recent machine learning studies on B-ALL flow cytometry analysis, utilizing methods such as Gaussian mixture model-based pipelines, SVM-optimized radar plots, and deep neural networks, have developed high-performing tools that support expert interpretation and show strong concordance with manual gating.[6-8] However, comparisons across studies remain difficult due to differences in analytical focus, MRD inclusion, and evaluation metrics. Despite these advances, current systems have not yet achieved clinical-grade accuracy as standalone tools and are best suited for human-in-the-loop applications rather than fully automated assessment.[9]

To narrow the gap to the real-world clinical practice, we developed FlowARC (Flow analysis using Residual Convolutional network), a novel deep-learning approach for precise and automated B-ALL detection, including in the MRD setting. The FlowARC model employs a cascaded 3-stage architecture (Figure 1), a cell-level module (Audit stage), a cell-ranking step (Reorder stage), and a sample-level module (Classify stage). In the Audit stage, individual cells within a flow cytometry sample are classified as normal or tumorous, with a leukemic probability score assigned to each cell. The Reorder stage ranks all cells by their leukemic probability and retains the top 7,500 most likely leukemic cells for further analysis. In the Classify stage, this subset is used by the sample-level module to determine whether the overall sample is tumorous or normal. Additionally, a separate quantification module estimates tumor burden, an essential feature for clinical reporting, particularly in the context of MRD.

FlowARC was trained and evaluated using a large retrospective clinical cohort of patients tested for B-ALL at our institution between 2015 and 2019, using an 8-color flow cytometry assay that followed standard diagnostic protocols.[10,11] This assay included surface markers (CD20, CD34, CD10, CD33, CD58, CD45, CD19, CD38) along with forward scatter (FSC-H, FSC-A) and side scatter (SSC-H, SSC-A) parameters. The cohort comprised 1,681 flow cytometry samples from 333 patients (184 male, 149 female), including 1,149 B-ALL-positive and 532 B-ALL-negative cases, totaling approximately 85.7 million B cells. Most samples were derived from bone marrow (1,141 positive and 532 negative), with a small number from peripheral blood (6 positive) and tissue (2 positive). The mean age of patients was 32.8 years (Standard Deviation 22.2, range 0-81 years). The Institutional Review Board at Memorial Sloan Kettering Cancer Center approved the study.

All FCS files from the cohort were processed using a standardized preprocessing pipeline including compensation and transformation using flowCore[12] (R package version 2.14.2) as well as automatic B-cell extraction by flowDensity (*Online Supplementary Figure S1*).[13] In negative cases, all extracted B cells were labeled as normal. For a representative subset of 137 B-ALL positive bone marrow samples, abnormal immature B cells were manually annotated by an expert hematopathologist (MR). These annotations were used to train the cell-level module and to generate synthetic datasets for training the sample-level and quantification modules. Synthetic data were generated to enhance the diversity of tumor cell populations across a wide range of tumor burdens, including 25 to 500 B cells (low MRD), 500 to 25,000 B cells (high MRD), 25,000 to 50,000 B cells (low tumor burden), and 50,000 to 500,000 B cells (high tumor burden). In our synthetic sample generation, we set a minimum leukemic population of 25 cells, corresponding to a theoretical sensitivity of up to 10^{-5}, depending on the total number of cells analyzed, reflecting the real-world detection threshold of the underlying flow cytometry assay (*Online Supplementary Figure S2*). To avoid data leakage, patient-level splits were performed prior to synthetic data generation. Among negative cases, 65% were randomly allocated for training. Of the 137 manually annotated positive samples, 75% were used for training, while the remaining 25%, along with all



Aspect	FlowARC *vs*. prior approaches	Key impact
Architecture	Three-stage cascade (cell → ranked subset → sample) instead of single-stage model	Greater accuracy; lower noise
Data focus	Selects top 7,500 high-probability abnormal cells, not all events	Better sensitivity for MRD cases
Training	Synthetic MRD samples from real data	Reduces need for extensive manual labeling
Interpretability	SHAP marker attribution and probability mapping	Transparent, biologically consistent results
Clinical utility	Integrated tumor burden estimate; visual review on standard plots	Enables MRD reporting and pathologist verification

Figure 1. FlowARC architecture. FlowARC, a novel and multi-module deep-learning-based model, follows a cascaded approach for detecting tumor presence in flow cytometry data. (A) Flow cytometry data acquisition and preprocessing pipeline: raw flow cytometry samples undergo preprocessing to extract B-cell populations based on forward/side scatter detectors and fluorescent markers. (B) Cell-level module training (Audit stage): a small subset of cells is manually annotated by pathologists as normal or tumor cells. These annotations train the cell-level neural network to classify individual cells and assign tumor probability scores. (C) Probability-based cell reordering (Reorder stage): within each flow cytometry sample, cells are automatically reordered based on their tumor probability scores from the cell-level module, ranking cells from highest to lowest tumor likelihood. This allows sam-

Continued on following page.

ple truncation from hundreds of thousands of cells to the most relevant thousands for efficient processing. (D) Sample-level classification (Classify stage): the reordered and truncated flow samples, labeled at the sample level (normal *vs*. tumor), train the sample-level module to classify entire flow cytometry samples as normal or tumorous. (E) Summarization of key innovations and clinical impact of the FlowARC model. MRD: minimal (measurable) residual disease.

unannotated positive cases, were reserved for testing. To evaluate model robustness, five independent instances of the cell-level, sample-level, and quantification modules were trained using different random weight initializations. Data preprocessing and model training (*Online Supplementary Table S1*) codes are available at https://github.com/MSK-Computational-HemePath/FlowARC.

FlowARC demonstrated high performance in differentiating abnormal B-cell precursors from normal cells at the cell level, achieving an area under the receiver operating characteristic curve (AUROC) of 0.995 (95%CI: 0.994-0.995). The corresponding confusion matrix included 10,028,906 true normal cells, 2,057,835 true tumor cells, 77,621 false normal cells (0.6%), and 341,424 false tumor cells (2.7%) (Figure 2C). At the sample level, reflecting clinical diagnoses, FlowARC continued to show excellent results. It achieved an AUROC of 0.994 (95%CI: 0.990-0.998) on the synthetic test set and maintained similarly high performance on real-world patient samples with an AUROC of 0.991 (95%CI: 0.987-0.995), accuracy of 97.5%, sensitivity of 96.4%, and specificity of 98.7% (Figure 2A). The sample-level confusion matrix comprised 225 true normal, 245 true tumor, 4 false normal (0.8%), and 6 false tumor cases (1.6%) (Figure 2D). Compared with the CellCNN model trained and tested on identical preprocessed data, FlowARC significantly outperformed CellCNN (AUROC: 0.995 *vs*. 0.869; $P< 10^{-11}$ by DeLong's test) (Figure 2B). Tumor burden quantification model was performed on real-world patient samples, and demonstrated strong agreement with expert-assessed tumor content, achieving an R^2 of 0.879 (95%CI: 0.877-0.881), a slope of 0.963 (95%CI: 0.960-0.966), and an intercept of 0.111 (95%CI: 0.095-0.127), indicating that the model's estimates closely matched the true measured values (Figure 2C).

Beyond accurate B-ALL detection, FlowARC provides explainable outputs and visual tools to aid clinical interpretation. Using the SHAP (Shapley Additive ExPlanations) algorithm, we identified key markers contributing to predictions, with low CD38, high forward scatter (FSC-A), and increased CD58 emerging as the strongest predictors of tumor cells, consistent with known disease phenotypes (Figure 2D). These SHAP values help elucidate the phenotypic profiles underlying the model's abnormal cell predictions.

FlowARC's cell-level predictions can be visualized using standard bi-parametric flow cytometry plots. In true tumor cases, high-probability cells form distinct clusters across multiple markers (Figure 3A, bottom), whereas in true negatives or false positives, misclassified cells appear scattered without clustering (Figure 3A top, B top). In false negatives, abnormal cells may still form clusters resembling true positives (Figure 3B, bottom). These visualization features support pathologist-in-the-loop verification, improving diagnostic accuracy and interpretability. The same approach is effective for low-cell-count samples (<7,500 B cells), where true tumors form compact clusters and true negatives remain diffuse (Figure 3C), extending FlowARC's utility even in challenging cases.

In summary, we developed the FlowARC model for automated detection of B-ALL, achieving over 0.99 AUROC in real-world clinical cases (Figure 1E). Its novel cascaded architecture, combining cell-level and sample-level modules, overcomes the limitations of previous methods and enables clinical-grade detection of lymphoblastic lymphoma in both diagnostic and MRD samples. The model achieved performance comparable to expert pathologists, with superior reproducibility and consistency, whereas human performance is influenced by inter-observer variability with reported concordance of 74-93%.[3,9,14] Misclassifications primarily occurred in samples with very low tumor burden (<0.01%) or overlapping aberrant and regenerative immunophenotypes, which are scenarios that challenge both AI and human reviewers, while cell-level visualization of prediction scores helped identify and correct some of these borderline cases.

Although FlowARC was developed and validated using data from our institution, its architecture and training strategy are broadly adaptable. The model is panel-specific, but the training framework is universally applicable and can be retrained on any clinical flow cytometry panel with limited locally annotated data. Model performance aligns with the analytical capability of the target clinical assay, facilitating cross-institutional consistency in diagnostic quality. Moreover, the same architecture can be adapted to other leukemia types, such as acute myeloid leukemia and T-cell acute lymphoblastic leukemia, by retraining on disease-specific panels and ensuring sufficient positive cases to account for immunophenotypic variability.

FlowARC addresses the extensive data requirements of deep learning by using synthetic samples generated from real clinical data, which closely mirror real-world cellular variability. This approach enables effective training with a limited number of curated cases. An ablation study showed that approximately 200 annotated cases (100 normal and 100 abnormal) are sufficient to achieve clinically acceptable performance, with only a modest decline from the full model. The use of locally representative cases, ideally verified by an expert hematopathologist, further ensures FlowARC's practicality and accessibility for clinical laboratories integrating AI-driven diagnostics into routine workflows.

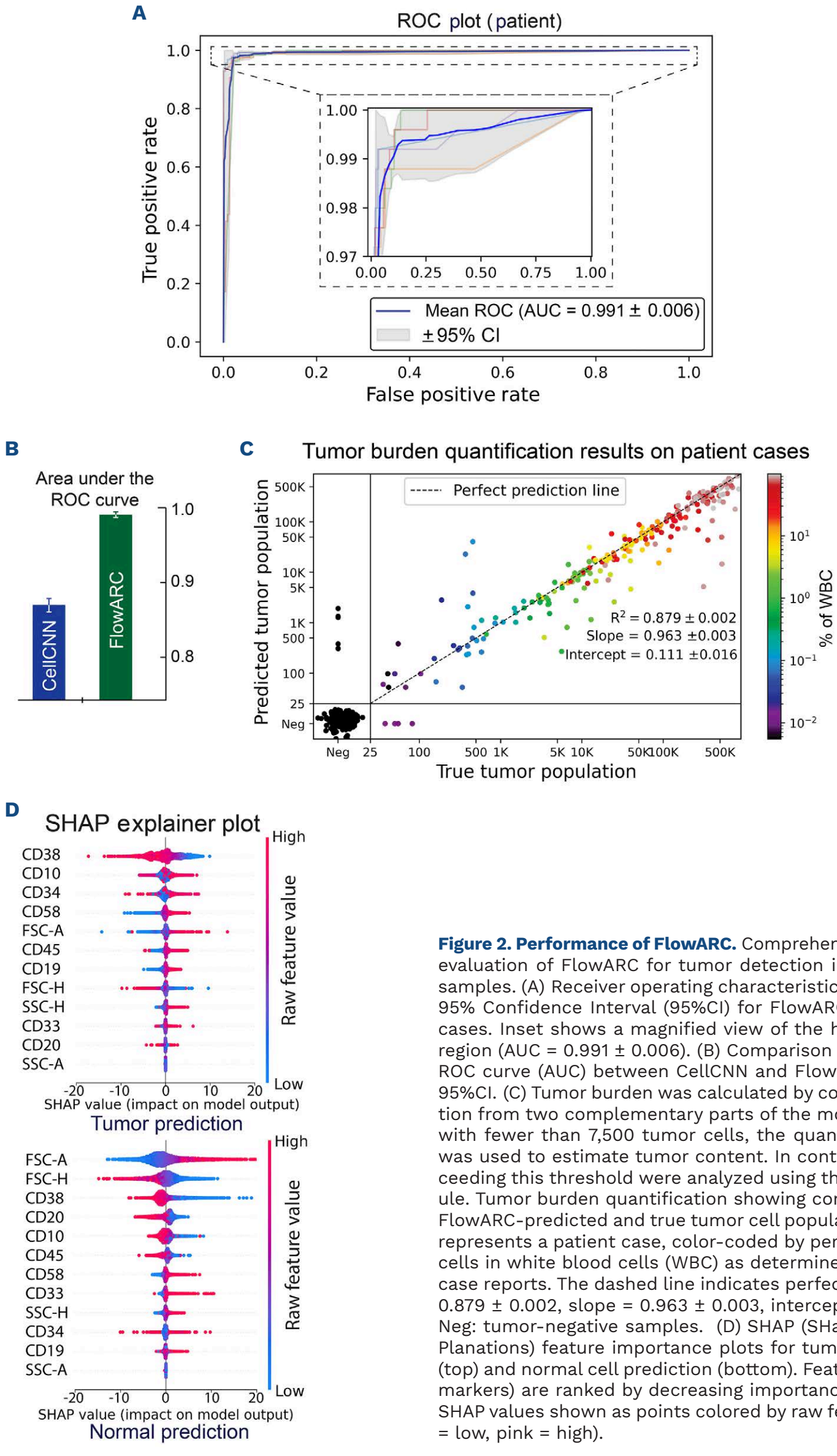


Figure 2. Performance of FlowARC. Comprehensive performance evaluation of FlowARC for tumor detection in flow cytometry samples. (A) Receiver operating characteristic (ROC) curve with 95% Confidence Interval (95%CI) for FlowARC on patient test cases. Inset shows a magnified view of the high-performance region (AUC = 0.991 ± 0.006). (B) Comparison of area under the ROC curve (AUC) between CellCNN and FlowARC models with 95%CI. (C) Tumor burden was calculated by combining information from two complementary parts of the model. For samples with fewer than 7,500 tumor cells, the quantification module was used to estimate tumor content. In contrast, samples exceeding this threshold were analyzed using the cell-level module. Tumor burden quantification showing correlation between FlowARC-predicted and true tumor cell populations. Each point represents a patient case, color-coded by percentage of tumor cells in white blood cells (WBC) as determined by institutional case reports. The dashed line indicates perfect prediction (R^2 = 0.879 ± 0.002, slope = 0.963 ± 0.003, intercept = 0.111 ± 0.016). Neg: tumor-negative samples. (D) SHAP (SHapley Additive exPlanations) feature importance plots for tumor cell prediction (top) and normal cell prediction (bottom). Features (cell surface markers) are ranked by decreasing importance, with individual SHAP values shown as points colored by raw feature value (blue = low, pink = high).

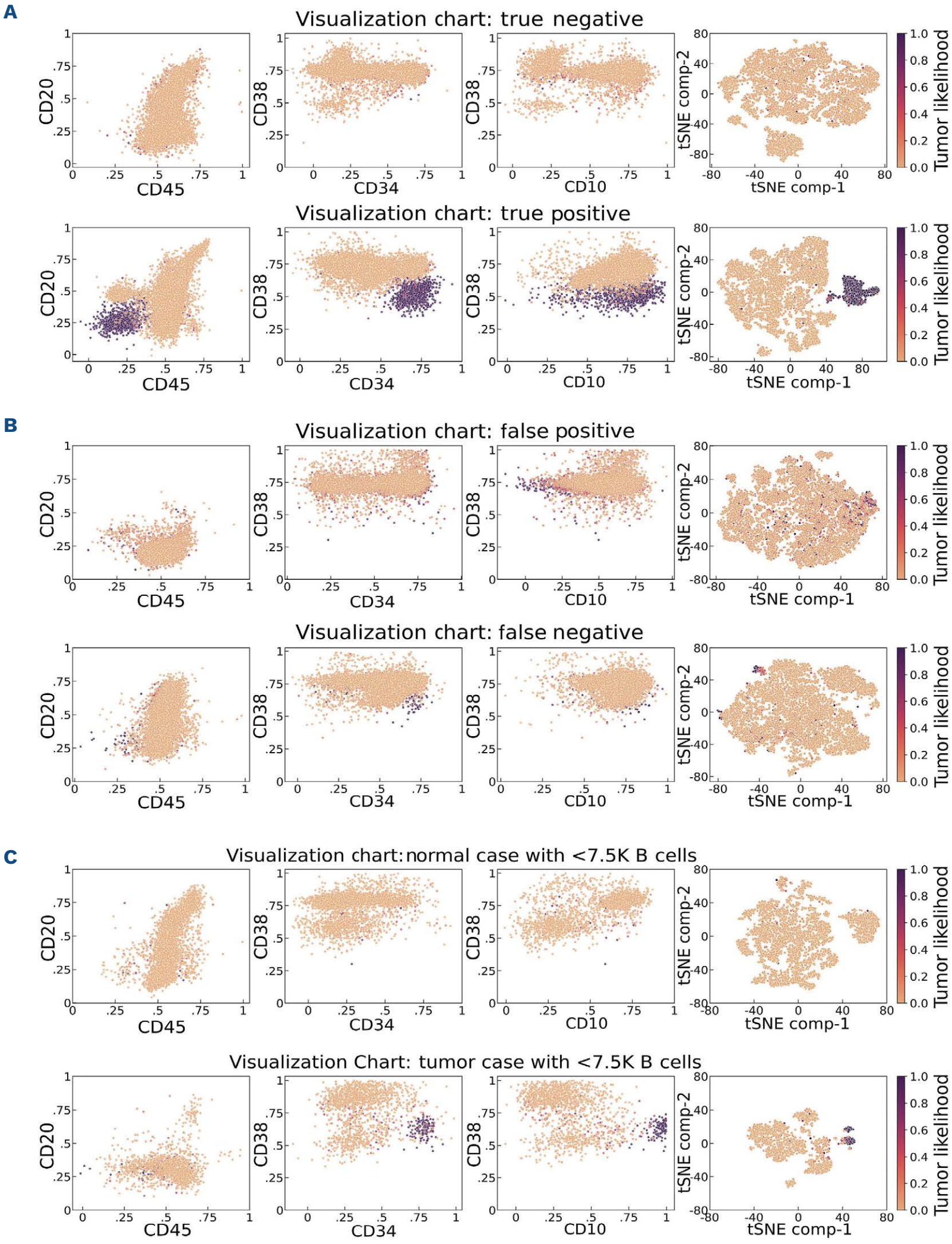


Figure 3. Flow cytometry visualization of B-acute lymphoblastic leukemia detection by FlowARC cell-level module. Direct visualization of FlowARC's cell-by-cell tumor probability assignments across representative patient cases. Each row represents a single patient case displayed across four biaxial plots: CD45 *versus* CD20, CD34 *versus* CD38, CD10 *versus* CD38, and t-SNE dimensionality reduction (components 1 and 2). Individual cells are color-coded by FlowARC-predicted tumor likelihood (purple = high tumor probability, orange = normal cells, as indicated by color bar). (A) Correctly classified cases: true negative (top row) shows uniform orange coloring indicating correct identification of normal B cells; true positive (bottom row) displays distinct

Continued on following page.

purple tumor cell populations clearly separated from normal cells. (B) Misclassified cases: false positive (top row) shows scattered likely-tumor cells that do not form tight clusters in any biaxial plots, indicating misclassification of the case as abnormal; false negative (bottom row) demonstrates likely-tumor cells that, although few, form tight clusters in CD34 *versus* CD38 and t-SNE plots, indicating model misclassification as normal. (C) Cases with insufficient B-cell counts (<7,500 cells) excluded from sample-level analysis: normal case (top row) does not show clusters of likely-tumor cells, while the abnormal case (bottom row) displays a distinct purple tumor cluster as expected. Despite being unable to process these cases through the complete FlowARC pipeline due to low B-cell numbers, the cell-level module visualization remains highly informative for identifying tumor populations even in these low cell count samples. These visualizations demonstrate FlowARC's ability to identify phenotypically distinct tumor populations while highlighting edge cases and current limitations.

The resulting automation enabled by FlowARC holds considerable potential for optimizing diagnostic and MRD evaluation workflows. It streamlines key analytical processes, resulting in notably shorter diagnostic turnaround times, improved workflow efficiency, and reduced variability from manual interpretation. These enhancements in diagnostic accuracy and speed are likely to lead to more informed clinical decisions and improved patient outcomes.

https://doi.org/10.3324/haematol.2025.288277

Received: May 30, 2025.
Accepted: December 24, 2025.
Early view: January 8, 2026.

Disclosures

No conflicts of interest to disclose.

Contributions

SC, MZ, HV, MR and AD designed the study; SC conducted data processing, model training, and performance evaluation; MR performed expert annotation of flow cytometry data; SR, JB, QG and AL assisted in data collection and technical validation; SC, MR, AL, MZ and HV wrote and reviewed the manuscript.

Data-sharing statement

The data that supports the findings of this study are available from the corresponding authors upon reasonable request.

Authors

Sulov Chalise,[1] Mikhail Roshal,[1] Sophia Roshal,[2] Jeeyeon Baik,[1] Qi Gao,[1] Anyi Li,[3] Ahmet Dogan,[1] Harini Veeraraghavan[3] and Meng-Lei Zhu[1]

[1]Department of Pathology and Laboratory Medicine, Memorial Sloan Kettering Cancer Center, New York, NY; [2]Carnegie Mellon University, Pittsburgh, PA and [3]Department of Medical Physics, Memorial Sloan Kettering Cancer Center, New York, NY, USA

Correspondence:
H. VEERARAGHAVAN - veerarah@mskcc.org
M. ZHU- zhum1@mskcc.org

References

1. Borowitz MJ, Wood BL, Devidas M, et al. Prognostic significance of minimal residual disease in high risk B-ALL: a report from Children's Oncology Group study AALL0232. Blood. 2015;126(8):964-971.
2. Tembhare PR, Subramanian Pg PG, Ghogale S, et al. A high-sensitivity 10-color flow cytometric minimal residual disease assay in B-lymphoblastic leukemia/lymphoma can easily achieve the sensitivity of 2-in-10^6 and is superior to standard minimal residual disease assay: a study of 622 patients. Cytometry B Clin Cytom. 2020;98(1):57-67.
3. Keeney M, Wood BL, Hedley BD, et al. A QA program for MRD testing demonstrates that systematic education can reduce discordance among experienced interpreters. Cytometry B Clin Cytom. 2018;94(2):239-249.
4. Hu Z, Bhattacharya S, Butte AJ. Application of machine learning for cytometry data. Front Immunol. 2021;12:787574.
5. Arvaniti E, Claassen M. Sensitive detection of rare disease-associated cell subsets via representation learning. Nat Commun. 2017;8:14825.
6. Reiter M, Diem M, Schumich A, et al. Automated flow cytometric MRD assessment in childhood acute B-lymphoblastic leukemia using supervised machine learning. Cytometry A. 2019;95(9):966-975.
7. Shopsowitz KE, Liu L, Setiadi A, et al. Machine learning optimized multiparameter radar plots for B-cell acute lymphoblastic leukemia minimal residual disease analysis. Cytometry B Clin Cytom. 2022;102(5):342-352.
8. Seheult JN, Otteson GE, Timm MM, et al. Artificial intelligence accelerates the interpretation of measurable residual B lymphoblastic leukemia by flow cytometry. Blood Adv. 2026;10(1):58-69.
9. Verbeek MWC, Rodriguez BS, Sedek L, et al. Minimal residual disease assessment in B-cell precursor acute lymphoblastic leukemia by semi-automated identification of normal hematopoietic cells: a EuroFlow study. Cytometry B Clin Cytom. 2024;106(4):252-263.
10. Gao Q, Liu Y, Aypar U, et al. Highly sensitive single tube B-lymphoblastic leukemia/lymphoma minimal/measurable residual disease test robust to surface antigen directed therapy. Cytometry B Clin Cytom. 2023;104(4):279-293.
11. Geyer MB, Ritchie EK, Rao AV, et al. Pediatric-inspired chemotherapy incorporating pegaspargase is safe and results in high rates of minimal residual disease negativity in adults up to age 60 with Philadelphia chromosome-negative acute

lymphoblastic leukemia. Haematologica. 2021;106(8):2086-2094.
12. Hahne F, LeMeur N, Brinkman RR, et al. flowCore: a Bioconductor package for high throughput flow cytometry. BMC Bioinformatics. 2009;10:106.
13. Malek M, Taghiyar MJ, Chong L, et al. flowDensity: reproducing manual gating of flow cytometry data by automated density-based cell population identification. Bioinformatics. 2015;31(4):606-607.
14. Maurer-Granofszky M, Schumich A, Buldini B, et al. An extensive quality control and quality assurance (QC/QA) program significantly improves inter-laboratory concordance rates of flow-cytometric minimal residual disease assessment in acute lymphoblastic leukemia: an I-BFM-FLOW-Network report. Cancers (Basel). 2021;13(23):6148.