

Clustering-Based Rare-Event Detection Beyond Singlets in Flow Cytometry

Yejin Kwak^{1,2,3,*}, Roberto Würth^{4,5,8}, Angelo Jovin Yamachui
Sitcheu^{1,3}, Daniel Hübschmann^{2,6,7,8,9}, Ralf Mikut¹

¹Institute for Automation and Applied Informatics (IAI), Karlsruhe Institute of
Technology (KIT)

²Innovation and Service Unit for Bioinformatics and Precision Medicine, German
Cancer Research Center (DKFZ)

³Helmholtz Information and Data Science School for Health (HIDSS4Health)

⁴Heidelberg Institute for Stem Cell Technology and Experimental Medicine
(HI-STEM gGmbH)

⁵Division of Stem Cells and Cancer
German Cancer Research Center (DKFZ) and DKFZ-ZMBH Alliance

⁶National Center for Tumor Diseases (NCT), NCT Heidelberg, a partnership
between DKFZ and Heidelberg University Hospital

⁷Pattern Recognition and Digital Medicine Group, Heidelberg Institute for Stem
Cell Technology and Experimental Medicine (HI-STEM gGmbH)

⁸German Cancer Consortium (DKTK), Core Center Heidelberg

⁹Division of Translational Precision Medicine, Institute of Human Genetics,
Heidelberg University

*E-mail: yejin.kwak@kit.edu

Abstract

Physical cell interactions are biologically relevant, but they are routinely discarded as doublets or multiplets in standard flow cytometry workflows. A recent clustering-based cytometry framework has instead retained multiplets while

using non-uniform sampling, or sketching, to preserve rare cell types during large-scale analysis. However, it remains unclear how well such sampling strategies support downstream recovery of extremely rare events and how reliably singlets can be distinguished from multiplets. To address this, we compare leverage-score and geometric sketching for recovering circulating tumor cells (CTCs) among singlet events, and evaluate whether adding side scatter (SSC) ratio to forward scatter (FSC) ratio improves singlet/multiplet discrimination across all events. The analysis comprises 4 blood samples from 3 patients, ranging from 116,507 to 1,361,065 events each. Geometric sketching yielded higher CTC recovery on average, particularly in recall, though performance remained sensitive to sample identity and clustering resolution. FSC ratio alone provided strong singlet/multiplet separation, while adding SSC ratio yielded no consistent improvement. These findings highlight the importance of both sampling strategy and feature selection for extending clustering-based cytometry to extremely rare events while retaining information beyond singlets.

1 Introduction

Cellular interactions are important for understanding development, homeostatic function, and pathophysiology of multicellular organisms, shaping processes from immune activation to tumor progression. Characterizing the frequency, composition, and dynamics of specific interaction types can reveal disease-associated alterations, inform patient stratification, and support the monitoring of therapeutic response. Representative examples include circulating tumor cell-neutrophil clusters, whose physical association enhances metastatic competence and has been proposed as a therapeutic target [1], and T cell-tumor cell conjugates, which are therapeutically induced by bispecific T cell engagers to promote tumor cell killing [2].

Despite the widespread adoption of single-cell genomics and spatial transcriptomics for dissecting cellular heterogeneity, direct analysis of physically interacting cells remains limited. Computational approaches such as CellPhoneDB [3] and CellChat [4] infer cell-cell communication indirectly, from the expression of ligands, receptors, and associated signaling components

across cell populations, without confirming physical contact between interacting cells. Spatial transcriptomics methods [5, 6] preserve spatial context and can identify neighboring cells, but spatial proximity alone does not establish direct physical interaction. Moreover, transcriptome-based inference is constrained by prior knowledge of signaling relationships, whereas high-resolution spatial approaches involve trade-offs among gene coverage, cell throughput, acquisition time, and cost.

Flow cytometry [7], by contrast, offers the potential to capture and characterize physical cell interactions through light scatter properties and surface marker co-expression at high throughput and comparatively low cost. Standard flow cytometry workflows, however, are generally centered on the analysis of individual singlet cells. This is traditionally achieved through hierarchical gating, which evaluates one or two parameters at each step [8], or, as panel sizes have grown, through clustering on the joint feature space [9]. In either case, doublets and multiplets are routinely excluded as technical artifacts prior to analysis, discarding potential interaction signal before it is examined. Historically, flow-cytometric analysis of cell-cell conjugates has instead been applied in targeted assays designed to quantify specific interaction types [10].

More recently, interact-omics [11] has begun to recover the discarded multiplet signal. We extend this line of work specifically to the rare-event regime, where events of interest constitute only a small fraction of all measured events. Because cytometry datasets can contain large numbers of events, computational pipelines often perform clustering on a downsampled subset, using sampling strategies designed to retain rare events, with cluster assignments subsequently extended to the full event pool [11, 12]. However, it remains unclear how effectively different downsampling strategies preserve extremely rare events in clinical flow cytometry.

To examine an extreme rare-event setting, we focus on circulating tumor cells (CTCs), which can occur at frequencies as low as approximately one cell per 10^9 blood cells, yet are implicated in metastatic spread and therapy resistance [13]. We therefore compare downsampling strategies for rare CTC recovery among singlets, and assess whether augmenting clustering with an additional scatter-derived feature can improve singlet/multiplet discrimination.

2 Background and Related Work

Hierarchical gating has long been the standard approach for cell population identification in flow cytometry [14]. However, as the number of measured parameters increases, hierarchical gating becomes increasingly time-consuming and difficult to scale [8]. The number of possible bi-axial marker combinations grows quadratically with panel size. Spectral flow cytometry has enabled panels of 50 or more fluorescent parameters with instruments recording fluorescence across dozens to more than one hundred detection channels [15]. At this scale, comprehensive analysis through sequential bi-axial gating becomes increasingly impractical. Consequently, high-dimensional cytometry analyses have increasingly adopted clustering-based analytical tools that consider multiple markers simultaneously [9].

Among the widely used clustering algorithms for this purpose are FlowSOM [16] and PhenoGraph [17], both of which have been implemented in commonly used cytometry analysis software. FlowSOM applies a self-organizing map followed by meta-clustering, whereas PhenoGraph constructs a nearest-neighbor graph and applies Louvain community detection. Both approaches, however, have been shown to vary substantially in their ability to resolve rare cell populations [18].

Cytometry pipelines have long incorporated downsampling strategies to better preserve rare populations during analysis. SPADE [12], for example, uses density-dependent downsampling to reduce differences in local cell density, thereby increasing the relative representation of rare populations prior to clustering. More recently, sketching approaches adopted for large-scale single-cell genomics have begun to be applied to cytometry data. Interact-omics [11], for instance, uses leverage-score sketching [19].

Leverage-score sketching [19] and geometric sketching [20] have both been applied in single-cell genomics to preserve rare populations during downsampling, but rely on contrasting criteria for determining which observations to retain. Leverage-score sketching, as implemented in Seurat v5 [19], assigns sampling probabilities according to statistical leverage, prioritizing high-leverage observations and thereby enriching rare populations. Geometric sketching [20],

in contrast, covers the feature space with equal-volume regions and samples approximately evenly across them, reducing the dominance of densely populated regions and increasing the relative representation of sparse populations. Whether these contrasting sketching strategies reliably preserve extremely rare events in clinical flow cytometry has not been systematically evaluated.

Beyond this uncertainty in sketching performance, the clustering algorithms discussed above can also show reduced performance in resolving rare populations [18]. Moreover, these algorithms are typically applied within cytometry workflows that exclude doublets and multiplets through upstream singlet gating. Interact-omics [11] is a closely related prior approach that recovers cell-interaction events from cytometry data through clustering, addressing this multiplet-exclusion gap. The method jointly clusters events using surface marker expression, light scatter properties, and the forward scatter (FSC) ratio, defined as FSC-A (area) divided by FSC-H (height). Multiplet events tend to exhibit higher FSC ratios than singlets, and clusters are subsequently classified based on the proportion of events above an FSC-ratio threshold.

This framework was primarily validated in immune-cell interaction settings in which interacting events were considerably more abundant than the CTC events examined here. Because downstream analysis depends on the events retained during sketching, it remains unclear whether extremely rare CTC events are retained sufficiently during sketching to enable reliable downstream detection after clustering and classification. It has also not been established how FSC ratio-based classification performs in CTC-containing clinical samples and whether it can be further refined using additional scatter-ratio features. To address these gaps, we compare leverage-score and geometric sketching for rare CTC recovery at both the sketch and full-pool levels. Separately, we test whether FSC ratio-based singlet/multiplet discrimination is effective in CTC-containing samples and whether incorporating an additional side scatter (SSC) ratio improves this discrimination.

3 Methodology

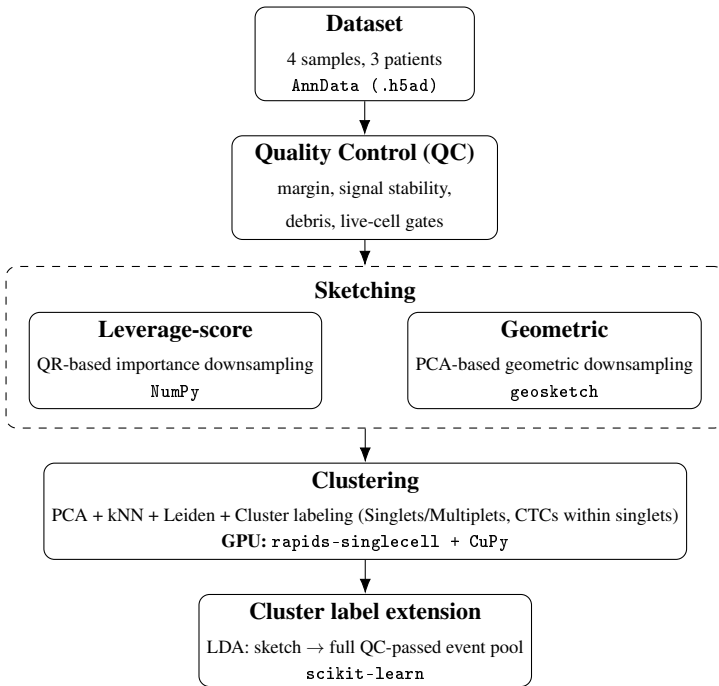


Figure 1: Overview of the analysis pipeline implemented in Python 3.11. After QC, samples are downsampled by leverage-score or geometric sketching, projected using principal component analysis (PCA), represented by a k-nearest-neighbor (kNN) graph, and clustered using Leiden clustering. Clusters are labeled as singlets or multiplets based on FSC ratio, with CTCs assigned only within singlet clusters, and the resulting labels are extended to the full QC-passed pool using linear discriminant analysis (LDA).

3.1 Dataset

The dataset comprises 4 conventional flow cytometry blood samples from 3 patients. These are non-spiked clinical samples, with CTCs present at naturally occurring frequencies rather than experimentally introduced or enriched. Up to two longitudinal timepoints per patient were available, spanning several months to years of clinical follow-up. Samples were acquired on a BD FACSaria

III Cell Sorter using four lasers (405, 488, 561, and 633 nm). Acquisition parameters included FSC and SSC area, height, and width measurements (FSC-A/H/W and SSC-A/H/W), together with three fluorescence channels (CD45-APC (allophycocyanin), EpCAM-FITC (fluorescein isothiocyanate), and DAPI). The subsequent analysis workflow, including QC, sketching, clustering, and cluster label extension, is summarized in Fig. 1.

3.2 Quality Control (QC)

Quality Control (QC) comprised margin dynamic-range event QC, signal stability QC, a debris gate, and a live cell gate. Margin dynamic-range event QC and signal stability QC are custom Python procedures inspired by PeacoQC [21], enabling integration into our analysis workflow.

Margin dynamic-range event QC is applied to untransformed data. Events within the upper 0.1% of each channel's range, as specified by the FCS \$PnR, are removed as upper-range events potentially affected by saturation or clipping. For scatter channels, negative values are also removed, as they are not expected for raw scatter measurements. For fluorescence channels, negative values are retained, as they can arise from compensation spread and are required to estimate the Logicle [22, 23] width parameter (W).

Signal stability QC is applied to transformed data across all channels ($\log(1+x)$ for scatter parameters, Logicle for fluorescence markers). Events are binned by equal event count to maintain comparable statistical support despite variation in acquisition rate. Bins whose median differs from the global median by more than a threshold defined using the median absolute deviation (MAD) are removed. This captures signal drift and abrupt perturbations associated with acquisition anomalies such as clogs.

Debris gating excludes events with low forward-scatter signals, which are more likely to represent cellular fragments than intact cells. In this dataset, debris exclusion was implemented through an instrument-level FSC threshold ($FSC > 5000$) applied during acquisition rather than through an additional post-acquisition gate.

Live cell gating retains events with low viability-marker signal. In this dataset, DAPI was used as the viability marker, with DAPI⁻ events considered live. The cutoff was determined per sample by one-dimensional kernel density estimation on the Logicle-transformed DAPI channel and set at the density valley between the dominant peak and the next sufficiently separated peak.

No further morphological cell gate is applied to avoid removing rare or morphologically atypical events, including CTCs and multiplets, before clustering.

3.3 Sketching

Sketching downsamples the QC-passed pool for each sample to 10,000 events. This sketch size was chosen to retain sufficient rare events while maintaining a manageable dataset size for downstream clustering, given QC-passed pool sizes of up to 1,361,065 events in this dataset. Baseline features comprised **raw scatter parameters** (FSC-A/H/W and SSC-A/H/W) and **Logicle-transformed fluorescence markers** (CD45-APC and EpCAM-FITC). All baseline features were standardized (z-scored and clipped to ± 10) prior to sketching.

Leverage-score sketching assigns sampling weights based on each event's contribution to the linear structure of the data. For the standardized event-by-feature matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$, where n and d denote the numbers of events and features, respectively, a reduced QR decomposition is computed as $\mathbf{X} = \mathbf{Q}\mathbf{R}$, where $\mathbf{Q}$ has orthonormal columns and $\mathbf{R}$ is upper triangular.

The leverage score h_i of event i is the squared Euclidean norm of the corresponding row of $\mathbf{Q}$, and the normalized sampling weight p_i is defined as

$$h_i = \|\mathbf{Q}_{i,:}\|_2^2, \quad p_i = \frac{h_i}{\sum_{j=1}^n h_j}, \quad (1)$$

where j indexes all n events. Events are sampled without replacement according to the weights p_i . Leverage scores were computed directly from the reduced QR decomposition using a custom NumPy-based implementation.

Geometric sketching aims to provide a more balanced representation of sparsely and densely populated regions of the data. For the standardized feature matrix $\mathbf{X}$,

PCA is performed using full singular value decomposition (full SVD), retaining all 8 available components (baseline features) in this study. The resulting embedding is covered by non-empty, equal-sized axis-aligned boxes,

$$\mathcal{C} = \{B_1, B_2, \dots, B_K\}, \quad (2)$$

where B_k denotes a covering box and K is the number of non-empty boxes. For the target sketch size of 10,000 events, events are sampled approximately evenly across the covering boxes, with events selected uniformly within each box. This reduces the dominance of densely populated regions in the resulting sketch. Geometric sketching is implemented using the Geosketch package.

3.4 Clustering

3.4.1 Clustering algorithm

Based on the clustering benchmark in Interact-omics [11], **Leiden clustering** was selected and implemented in `rapids-singlecell` with a fixed random seed for reproducibility.

Features of the sketched events (baseline features + FSC ratio (*raw FSC-A / raw FSC-H*)) are standardized (z-scored, clipped to ± 10) and a PCA embedding is computed using full SVD. Given the low-dimensional feature space, all 9 PCs are retained to avoid discarding low-variance structure that may be relevant to rare populations. PCA therefore acts only as an orthogonal transformation, preserving Euclidean distances for kNN graph construction.

A k -nearest-neighbor (kNN) graph is then constructed ($k = 15$, brute-force search) on this embedding, followed by Leiden clustering with a fixed random seed. The Leiden resolution parameter controls clustering granularity, with higher values generally producing a larger number of smaller clusters. For each sample and sketching method, the same 10,000-event sketch was reused across three Leiden resolutions ($r = 0.5, 1.0, 1.2$) to assess how downstream results depend on clustering granularity.

3.4.2 Cluster labeling

Singlets and multiplets are distinguished following the FSC ratio-based strategy of Interact-omics [11]. Otsu thresholding is applied to the event-level FSC ratio distribution across the sketch. In this implementation, a cluster is classified as a multiplet cluster when at least 50% of its member events exceed this threshold, and as a singlet cluster otherwise.

CTCs within singlets are identified based on their relative CD45 and EpCAM expression. Rather than relying on conventional manual inspection of marker expression [11, 16, 17], each singlet cluster is assigned a reproducible, rule-based score derived from its median CD45 and EpCAM expression. Because the two markers are measured on different fluorescence intensity scales, cluster medians for each marker are z-scored across singlet clusters within each sample, and a score is computed as $\text{score} = z_{\text{EpCAM}} - z_{\text{CD45}}$.

The highest-scoring singlet cluster, representing the strongest relative high-EpCAM/low-CD45 profile, is designated as the CTC cluster. Using relative comparisons across singlet clusters within the same sample instead of an absolute threshold makes the approach less sensitive to sample-to-sample variation in fluorescence intensity while remaining consistent with the $\text{CD45}^-/\text{EpCAM}^+$ CTC phenotype used in this study.

3.5 Cluster label extension

Leiden cluster labels are extended from the sketched pool ($n = 10,000$) to the full QC-passed pool ($n = N$) using **linear discriminant analysis (LDA)** as in Interact-omics [11]. The clustering features (scatter channels, marker channels, and FSC ratio) are standardized using the sketch mean and standard deviation (z-score, clipped to ± 10), and a multiclass LDA classifier is trained to predict the Leiden cluster labels and is then applied to the non-sketched events. Together with the original sketch labels, these predictions assign cluster labels across the full QC-passed pool. Each event inherits the singlet/multiplet label of its assigned cluster and, within singlet clusters, its CTC label where applicable.

3.6 Evaluation

3.6.1 Pseudo ground truth

Because no orthogonal confirmation (e.g., imaging-based validation) is available, hierarchical manual gating in FlowJo (v11) is used as a pseudo ground truth. The gating hierarchy used to define CTCs and multiplets is shown in Figure 2.

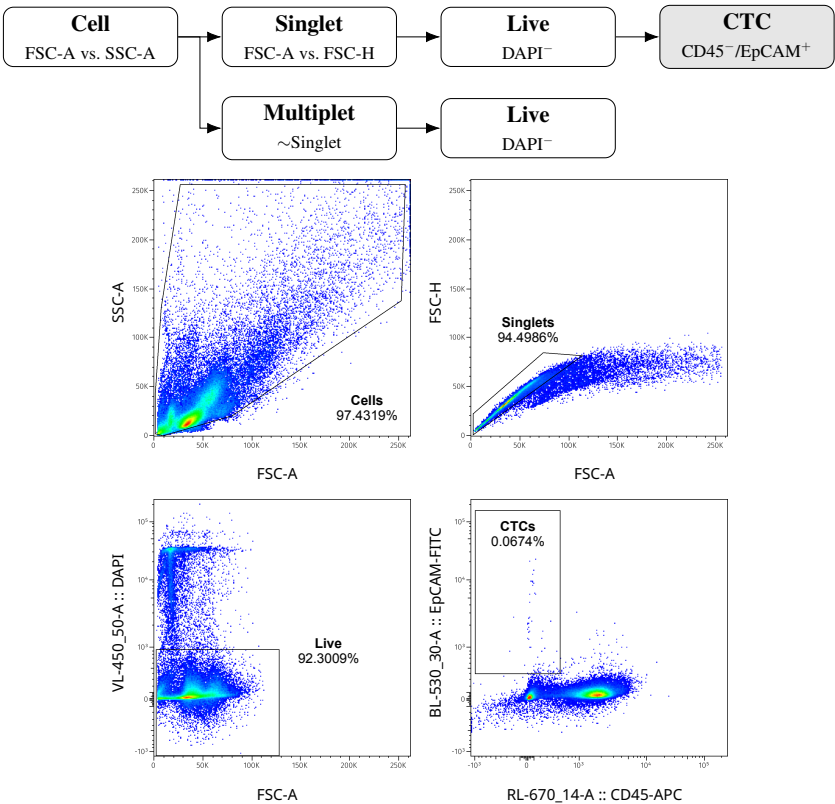


Figure 2: Manual gating hierarchy for P3 (T1), with representative gating plots for each step: an initial scatter-based cell gate, followed by singlet gating based on FSC-A vs. FSC-H, with events outside the singlet gate classified as multiplets. Live cells were subsequently selected using DAPI⁻ within both the singlet and multiplet branches, and CD45⁻/EpCAM⁺ singlet events were used to define the CTC population.

4 Results and Discussion

4.1 CTC recovery under QC

The QC pipeline (Section 3.2) is designed to remove technical noise while avoiding the exclusion of rare or heterotypic events. To assess this, we examine whether manually gated CTCs are preserved after QC (Table 1). Overall event retention ranged from 79.8% to 83.4% across the 4 samples, whereas manually gated CTCs were retained at a higher rate, from 86.9% (73 out of 84) to 100% (553 out of 553). This indicates that the QC procedure preserved the majority of manually gated CTCs. Manually gated CTC prevalence ranged from 0.054% to 0.238% of the QC-passed pool across samples, showing that CTCs constitute a rare population even after QC.

Table 1: Per-sample event and manually gated CTC counts before and after QC, with retention rates and post-QC CTC prevalence. (P: patient; T: timepoint; prev.: prevalence)

Sample	Events		CTCs		Prev. %
	Raw	Post-QC (%)	Manual	Post-QC (%)	
P1 (T1)	659,308	526,348 (79.8)	553	553 (100.0)	0.105
P1 (T2)	864,009	712,824 (82.5)	410	387 (94.4)	0.054
P2 (T1)	1,703,663	1,361,065 (79.9)	3,556	3,244 (91.2)	0.238
P3 (T1)	139,716	116,507 (83.4)	84	73 (86.9)	0.063

4.2 CTC recovery through sketching

We compare leverage-score sketching (lev-sketch) and geometric sketching (geo-sketch) across all 4 samples and 3 Leiden resolutions ($r = 0.5, 1.0, 1.2$), yielding 24 scenarios. For each scenario, CTC recovery within singlet clusters is evaluated after extending sketch-derived cluster labels to the full QC-passed pool (Section 3.5).

Cluster-based CTC predictions exceeded manual CTC counts in most scenarios (20 of 24), although underprediction was observed in four scenarios. The frequent overprediction may reflect both the conservative nature of hierarchical manual gating (Figure 2), which requires events to pass a sequence of strict gating boundaries, and possible overestimation by the clustering approach when CTC and non-CTC events are grouped within the same cluster. Given this uncertainty and the rare-event setting, recall and F_2 are emphasized to prioritize successful CTC recovery, while precision and F_1 are also reported to capture potential overprediction. F_1 and F_2 are defined as

$$F_1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}, \quad F_2 = 5 \cdot \frac{\text{precision} \cdot \text{recall}}{4 \cdot \text{precision} + \text{recall}}, \quad (3)$$

where F_1 balances precision and recall equally, whereas F_2 places greater emphasis on recall and is therefore more sensitive to successful CTC recovery.

Geo-sketch achieved higher full-pool mean and median values than lev-sketch across every summary metric (Table 2): mean F_1 (**0.445** vs. 0.370), mean precision (**0.410** vs. 0.338), and mean recall (**0.719** vs. 0.465). The gap was most pronounced for recall, where the median for geo-sketch exceeded the median for lev-sketch (**0.771** vs. 0.363) by more than a factor of two. This recall-driven advantage was reflected in the larger between-method difference in median F_2 score (**0.631** vs. 0.280) than in median F_1 score (**0.488** vs. 0.227).

Table 2: Mean and median CTC recovery metrics across 12 sample-resolution combinations (4 samples $\times$ 3 resolutions), for the sketch and full pool.

		Precision	Recall	F_1	F_2
Full pool	lev-sketch (mean)	0.338	0.465	0.370	0.409
	geo-sketch (mean)	0.410	0.719	0.445	0.541
	lev-sketch (median)	0.210	0.363	0.227	0.280
	geo-sketch (median)	0.366	0.771	0.488	0.631
Sketch	lev-sketch (mean)	0.322	0.516	0.375	0.433
	geo-sketch (mean)	0.347	0.708	0.430	0.536
	lev-sketch (median)	0.200	0.488	0.236	0.314
	geo-sketch (median)	0.391	0.717	0.483	0.618

Despite geo-sketch’s higher full-pool averages, **lev-sketch** showed more extreme performance, with higher maxima for F_1 (**0.882** vs. 0.836) and precision (**0.857** vs. 0.789), but lower minima for both (F_1 : **0.000** vs. 0.078, precision: **0.000** vs. 0.041) across all scenarios (Table 3). For recall, **geo-sketch** had both a higher maximum (**0.953** vs. 0.914) and a higher minimum (**0.297** vs. 0.000), although its range remained substantial. Thus, its higher mean and median recall partly reflected avoidance of very low-recall scenarios rather than uniformly stable performance.

Pronounced dependence of recall on resolution was observed more often for lev-sketch in these samples. In P1, lev-sketch’s recall varied sharply across resolutions (from 0.604 to 0.119 for T1, and from 0.054 to 0.912 for T2, between $r = 0.5$ and $r = 1.0$), while geo-sketch’s recall remained stable in both (swings of 0.024 and 0.015, respectively). In P3, lev-sketch’s recall instead declined monotonically toward zero across resolutions (0.110, 0.068, 0.000), a modest absolute change in magnitude but a complete loss of recall. At $r = 1.2$, this corresponded to complete recovery failure in P3, despite 48 manually gated CTCs being retained in the lev-sketch (Figure 3). This suggests that rare-event retention during sketching alone does not guarantee cluster-based recovery. In P2, however, the pattern was reversed. Geo-sketch recall dropped from 0.906 to 0.297, while lev-sketch remained stable (0.897–0.914).

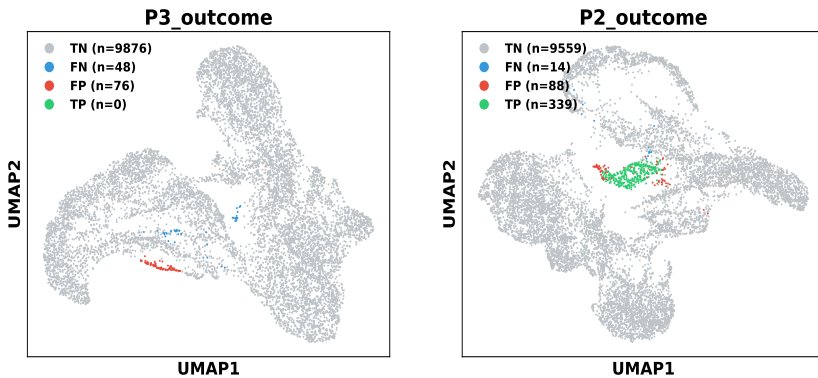


Figure 3: Comparison of sketch-level classification outcomes for P3 (left) and P2 (right) using lev-sketch at $r = 1.2$, highlighting the complete loss of recall observed in P3 compared with P2 under the same analysis settings.

Sketch-only performance is also examined as a secondary comparison to assess whether the relative performance of the two sketching methods is preserved after label extension. For precision and F_1 , the relative ranking was preserved. Whichever method scored higher within the sketch also scored higher after cluster extension, in all 12 paired scenarios. Recall was the exception. In 4 of the 12 scenarios, lev-sketch had the higher sketch-only recall but was overtaken by geo-sketch after extension, with no instances of the reverse.

Table 3: CTC recovery performance of geo-sketch and lev-sketch by sample and resolution (Res.).

Sample	Sketch	Res.	TP	FP	FN	Precision	Recall	F_1	F_2
P1 (T1)	geo	0.5	420	716	133	0.370	0.759	0.497	0.627
		1.0	433	763	120	0.362	0.783	0.495	0.635
		1.2	421	575	132	0.423	0.761	0.544	0.656
	lev	0.5	334	1350	219	0.198	0.604	0.299	0.429
		1.0	66	232	487	0.221	0.119	0.155	0.131
		1.2	67	393	486	0.146	0.121	0.132	0.125
P1 (T2)	geo	0.5	369	781	18	0.321	0.953	0.480	0.684
		1.0	368	655	19	0.360	0.951	0.522	0.716
		1.2	363	489	24	0.426	0.938	0.586	0.756
	lev	0.5	21	2493	366	0.008	0.054	0.014	0.026
		1.0	353	231	34	0.604	0.912	0.727	0.828
		1.2	335	851	52	0.282	0.866	0.426	0.613
P2	geo	0.5	2939	845	305	0.777	0.906	0.836	0.877
		1.0	1021	273	2223	0.789	0.315	0.450	0.358
		1.2	964	260	2280	0.788	0.297	0.432	0.339
	lev	0.5	2910	485	334	0.857	0.897	0.877	0.889
		1.0	2965	517	279	0.852	0.914	0.882	0.901
		1.2	2957	501	287	0.855	0.912	0.882	0.900
P3	geo	0.5	57	1338	16	0.041	0.781	0.078	0.169
		1.0	43	199	30	0.178	0.589	0.273	0.403
		1.2	43	446	30	0.088	0.589	0.153	0.275
	lev	0.5	8	1873	65	0.004	0.110	0.008	0.018
		1.0	5	182	68	0.027	0.068	0.038	0.052
		1.2	0	99	73	0.000	0.000	0.000	0.000

UMAP visualizations qualitatively illustrate differences in the sampled distributions between the two sketching methods (Figure 4). In the separately

computed UMAP embeddings, geo-sketch points appeared more evenly distributed, whereas lev-sketch points formed more compact and visually distinct groupings.

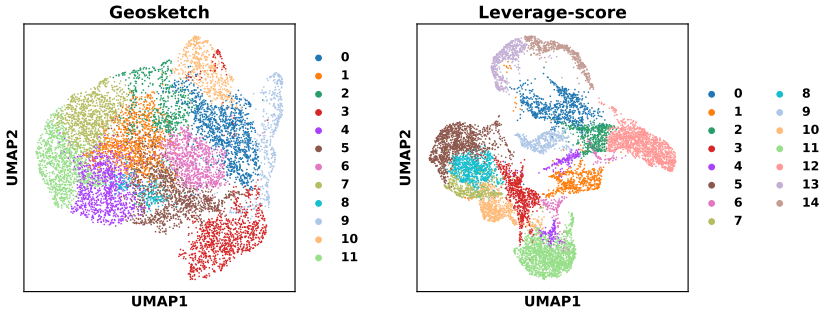


Figure 4: UMAP embeddings for geo-sketch (left) vs. lev-sketch (right), P2 at $r = 0.5$. Numbers indicate Leiden cluster labels.

4.3 Singlet/multiplet discrimination with SSC ratio

Next, we assess whether SSC ratio (raw SSC-A / raw SSC-H) improves singlet/multiplet discrimination within the QC-passed pool. For P2 at $r = 0.5$, we compare four conditions: lev-sketch with FSC ratio alone, lev-sketch with FSC and SSC ratios, geo-sketch with FSC ratio alone, and geo-sketch with both ratios. This setting was chosen because both sketching methods achieved high CTC F_2 scores (Table 3). The same sketch is reused within each sketching method, so within-method differences between the FSC-only and FSC+SSC conditions arise solely from the added clustering feature.

Adding SSC ratio had opposite effects for the two sketching methods (Table 4). For lev-sketch, it substantially improved specificity (0.789 to 0.940) and balanced accuracy (0.894 to 0.960), at a small cost to recall (0.999 to 0.980) and F_1 (0.997 to 0.989) (Figure 5). For geo-sketch, the same addition instead reduced specificity (0.720 to 0.660) and balanced accuracy (0.860 to 0.830), while recall (both near 1.0) and F_1 (0.996 to 0.995) remained comparably high.

Because singlets are treated as the positive class and substantially outnumber multiplets, precision and recall primarily characterize singlet classification

and do not directly quantify multiplet recovery. Specificity and balanced accuracy are therefore also reported to capture performance on the multiplet class. Because the manual singlet/multiplet gate is itself defined on FSC-A versus FSC-H, the strong performance of FSC ratio should not be interpreted as independent validation. The more informative comparison is whether SSC ratio provides additional discriminatory value beyond FSC ratio. Overall, SSC ratio did not provide a consistent improvement across the two sketching strategies. Specificity and balanced accuracy improved for lev-sketch but decreased for geo-sketch. This method-dependent effect suggests that the impact of SSC ratio depends on the sampled event composition and subsequent clustering structure, although these contributions were not separately evaluated here.

Table 4: Singlet/multiplet classification against the manual gate for P2 at $r = 0.5$, using FSC ratio alone or combined with SSC ratio. (Spec.: specificity; Bal. acc.: balanced accuracy)

Sketch	Features	Precision	Recall	F_1	Spec.	Bal. acc.
lev-sketch	FSC ratio	0.9937	0.9994	0.9965	0.7892	0.8943
lev-sketch	FSC + SSC ratio	0.9982	0.9798	0.9889	0.9404	0.9601
geo-sketch	FSC ratio	0.9916	0.9999	0.9958	0.7197	0.8598
geo-sketch	FSC + SSC ratio	0.9899	0.9996	0.9947	0.6601	0.8298

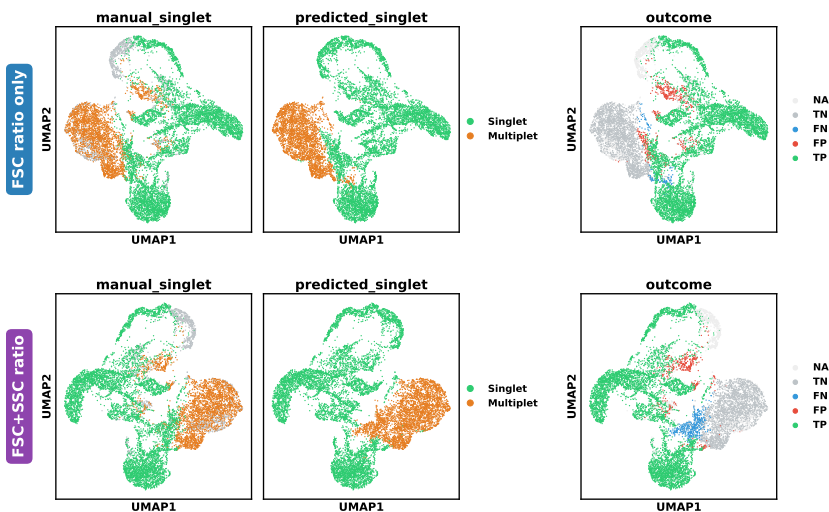


Figure 5: Manual singlet/multiplet (left), predicted singlet/multiplet (center), and classification outcome (right) for lev-sketch, P2 at $r = 0.5$, using FSC ratio only (top) and FSC and SSC ratios combined (bottom). Events without a corresponding manual singlet/multiplet label are shown as NA and were excluded from metric calculation.

5 Conclusion

In this study, we investigated whether clustering-based cytometry analysis can be extended to an extreme rare-event setting while retaining multiplets that are conventionally excluded during singlet gating. Using CTCs in clinical blood samples, we compared geometric and leverage-score sketching prior to Leiden clustering and extended sketch-derived cluster labels to the full QC-passed event pool. Geometric sketching showed higher average CTC recovery from singlet clusters, particularly in recall, and avoided the complete recovery failure observed for leverage-score sketching. However, performance remained sample- and resolution-dependent, showing that sketching strategy can substantially affect rare-event recovery. Adding SSC ratio to FSC ratio did not consistently improve singlet/multiplet discrimination in the evaluated P2 setting at $r = 0.5$, indicating that additional scatter-derived features should be evaluated in the context of the sampling and clustering strategy.

Several limitations constrain the interpretation of these findings. No orthogonal ground truth, such as imaging-based confirmation of CTCs and physical cell interactions, was available. Manual gating was therefore used as a pseudo ground truth, although this means that the clustering results are evaluated against labels derived from the same fluorescence and scatter measurements. The data were originally generated to isolate CTCs for downstream molecular profiling, rather than rare-event clustering or multiplet identification. Consequently, the acquisition panel, available markers, and reference annotations were not optimized for the present task. The dataset comprised only 4 blood samples from 3 patients, limiting the scope for inferential statistical analysis.

The present analysis selected the single most CTC-like singlet cluster based on relative CD45 and EpCAM expression. CTC labeling was restricted to singlets because CD45⁺/EpCAM⁺ multiplets could represent CTC-leukocyte conjugates, but their cellular composition could not be unambiguously determined with the available marker panel. Moreover, phenotypically heterogeneous CTCs may occupy multiple clusters or form a minority subpopulation within larger mixed clusters, which may lead to underestimation of CTC abundance. Conversely, assigning an entire mixed cluster as CTC-positive may overestimate CTC abundance by including non-CTC events. With more reliable ground truth, future studies should therefore evaluate multiple CTC-cluster assignments and downstream subclustering of CTC-enriched or mixed clusters.

The LDA-based extension of sketch-derived labels could also be compared with non-linear classifiers, such as kernel support vector machines [24] or random forests [25], which can capture more complex decision boundaries than linear discriminant analysis. Beyond label extension, alternative clustering strategies may also be useful for rare-event recovery. In particular, noise clustering explicitly introduces a noise class for observations that are poorly represented by the main clusters [26]. Further examination of these events could provide a complementary strategy for identifying rare events, although the noise class may also contain technical noise and debris.

Finally, the analysis was implemented in Python 3.11 using AnnData (.h5ad) as the central data structure, with GPU acceleration using rapids-singlecell and CuPy. In contrast to R-based Interact-omics, the present framework operates

within the Python/AnnData ecosystem, allowing these GPU-enabled tools to support analysis of larger event pools. The use of `.h5ad` as the central data structure also provides a natural route for integrating `.h5`-based imaging flow cytometry features into the same event-level representation. Such morphological information could complement scatter and marker measurements and provide a more direct basis for distinguishing singlets from multiplets.

Acknowledgements

The present contribution is supported by the Helmholtz Association under the joint research school "HIDSS4Health - Helmholtz Information and Data Science School for Health" and the program "Natural, Artificial and Cognitive Information Processing (NACIP)". The authors used the AI language model Claude Sonnet 5 to improve the language and style of this manuscript. The authors reviewed all AI-assisted text and take full responsibility for the content.

References

- [1] B. M. Szczerba, F. Castro-Giner, M. Vetter, I. Krol, S. Gkoutela, J. Landin, M. C. Scheidmann, C. Donato, R. Scherrer, J. Singer, et al. Neutrophils escort circulating tumour cells to enable cell cycle progression. *Nature*, 566(7745):553–557, 2019.
- [2] H. Kantarjian, A. Stein, N. Gökbüget, A. K. Fielding, A. C. Schuh, J.-M. Ribera, A. Wei, H. Dombret, R. Foà, R. Bassan, et al. Blinatumomab versus chemotherapy for advanced acute lymphoblastic leukemia. *New England Journal of Medicine*, 376(9):836–847, 2017.
- [3] M. Efremova, M. Vento-Tormo, S. A. Teichmann, and R. Vento-Tormo. CellPhoneDB: inferring cell–cell communication from combined expression of multi-subunit ligand–receptor complexes. *Nature Protocols*, 15(4):1484–1506, 2020.

- [4] S. Jin, M. V. Plikus, and Q. Nie. CellChat for systematic analysis of cell–cell communication from single-cell transcriptomics. *Nature Protocols*, 20(1):180–219, 2025.
- [5] K. H. Chen, A. N. Boettiger, J. R. Moffitt, S. Wang, and X. Zhuang. Spatially resolved, highly multiplexed RNA profiling in single cells. *Science*, 348(6233):aaa6090, 2015.
- [6] M. F. d. Oliveira, J. P. Romero, M. Chung, S. R. Williams, A. D. Gottscho, A. Gupta, S. E. Pilipauskas, S. Mohabbat, N. Raman, D. J. Sukovich, et al. High-definition spatial transcriptomic profiling of immune cell populations in colorectal cancer. *Nature Genetics*, 57(6):1512–1523, 2025.
- [7] J. P. Robinson and M. Roederer. Flow cytometry strikes gold. *Science*, 350(6262):739–740, 2015.
- [8] E. Lugli, M. Roederer, and A. Cossarizza. Data analysis in flow cytometry: the future just started. *Cytometry Part A*, 77(7):705–713, 2010.
- [9] J. P. Robinson, G. B. Gmyrek, and B. Rajwa. Flow cytometry: advances, challenges and trends. *Bioessays*, 48(1):e70091, 2026.
- [10] G. Iserentant, C. Seguin-Devaux, and J. Zimmer. Flow cytometry conjugate formation assay between natural killer cells and their target cells. In *Methods in Cell Biology*, volume 193, pages 213–228. Elsevier, 2025.
- [11] D. Vonficht, L. Jopp-Saile, S. Yousefian, V. Flore, I. Simó Vesperinas, R. Teuber, B. Avanesyan, Y. Luo, C. Röthemeier, F. Grünschläger, et al. Ultra-high-scale cytometry-based cellular interaction mapping. *Nature Methods*, 22(9):1887–1899, 2025.
- [12] P. Qiu, E. F. Simonds, S. C. Bendall, K. D. Gibbs Jr, R. V. Bruggner, M. D. Linderman, K. Sachs, G. P. Nolan, and S. K. Plevritis. Extracting a cellular hierarchy from high-dimensional cytometry data with SPADE. *Nature Biotechnology*, 29(10):886–891, 2011.
- [13] Z. Myrkhieva, K. Seitkamal, Z. Ashikbayeva, A. Taizhanova, D. Tosi, and A. Bekmurzayeva. Circulating tumor cells: overcoming challenges

of detecting a needle in a haystack. *Exploration of Targeted Anti-tumor Therapy*, 6:1002321, 2025.

- [14] A. Cossarizza, H. D. Chang, A. Radbruch, S. Abrignani, R. Addo, M. Akdis, I. Andrä, F. Andreato, F. Annunziato, E. Arranz, et al. Guidelines for the use of flow cytometry and cell sorting in immunological studies. *European Journal of Immunology*, 51(12):2708–3145, 2021.
- [15] E. Astakhova, A. Gubaeva, D. Naumova, A. Egorova, A. Maznina, I. Rybkina, I. Osmanov, D. Tabakov, O. Mityaeva, and P. Y. Volchkov. Spectral flow cytometry: the current state and future of the technology. *International Journal of Molecular Sciences*, 26(12):5911, 2025.
- [16] S. Van Gassen, B. Callebaut, M. J. Van Helden, B. N. Lambrecht, P. Demeester, T. Dhaene, and Y. Saeys. FlowSOM: using self-organizing maps for visualization and interpretation of cytometry data. *Cytometry Part A*, 87(7):636–645, 2015.
- [17] J. H. Levine, E. F. Simonds, S. C. Bendall, K. L. Davis, E.-a. D. Amir, M. D. Tadmor, O. Litvin, H. G. Fienberg, A. Jager, E. R. Zunder, et al. Data-driven phenotypic dissection of AML reveals progenitor-like cells that correlate with prognosis. *Cell*, 162(1):184–197, 2015.
- [18] K. Märtens, M. Bortolomeazzi, L. Montorsi, J. Spencer, F. Ciccarelli, and C. Yau. Rarity: discovering rare cell populations from single-cell imaging data. *Bioinformatics*, 39(12):btad750, 2023.
- [19] Y. Hao, T. Stuart, M. H. Kowalski, S. Choudhary, P. Hoffman, A. Hartman, A. Srivastava, G. Molla, S. Madad, C. Fernandez-Granda, et al. Dictionary learning for integrative, multimodal and scalable single-cell analysis. *Nature Biotechnology*, 42(2):293–304, 2024.
- [20] B. Hie, H. Cho, B. DeMeo, B. Bryson, and B. Berger. Geometric sketching compactly summarizes the single-cell transcriptomic landscape. *Cell Systems*, 8(6):483–493, 2019.
- [21] A. Emmaneel, K. Quintelier, D. Sichien, P. Rybakowska, C. Marañón, M. E. Alarcón-Riquelme, G. Van Isterdael, S. Van Gassen, and Y. Saeys.

- PeacoQC: peak-based selection of high quality cytometry data. *Cytometry Part A*, 101(4):325–338, 2022.
- [22] D. R. Parks, M. Roederer, and W. A. Moore. A new “logicle” display method avoids deceptive effects of logarithmic scaling for low signals and compensated data. *Cytometry Part A: The Journal of the International Society for Analytical Cytology*, 69(6):541–551, 2006.
- [23] W. A. Moore and D. R. Parks. Update for the logicle data scale including operational code implementations. *Cytometry. Part A: the journal of the International Society for Analytical Cytology*, 81(4):273, 2012.
- [24] C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20:273–297, 1995.
- [25] L. Breiman. Random forests. *Machine Learning*, 45:5–32, 2001.
- [26] R. N. Dave. Robust fuzzy clustering algorithms. In *IEEE International Conference on Fuzzy Systems*, pages 1281–1286. IEEE, 1993.