Evaluation of Clustering Methods and Large Language Models for Spatial Proteomic Cell Type Annotation

Zachary Deutsch[1], Nick Surguladze[2], Yang Miao[1], John W. Hickey[1]*

[1]Department of Biomedical Engineering, Duke University, Durham, NC 27708, USA
[2]Department of Biostatistics and Bioinformatics, Duke University School of Medicine, Durham, NC 27710, USA

*Corresponding author. E-mail: john.hickey@duke.edu

Abstract

Spatial proteomics enables cell-type mapping while preserving tissue architecture, but annotation remains difficult to scale because its accuracy depends on segmentation, clustering, marker selection, and expert interpretation. Here, we benchmarked these components using a healthy human intestine CODEX dataset comprising 220,082 cells across eight tissue regions. Reference labels included 10,495 Noise cells (4.8% of all cells), representing cells without a confident cell-type assignment. We first evaluated spillover compensation across six segmentation methods and found that its effects on clustering and cell-type recovery varied by segmentation algorithm, tissue region, and cell population, with no consistent improvement across metrics. We next compared four clustering methods—Leiden, FlowSOM, SpatialSort, and PIXIE—against reference annotations. Performance varied across clustering methods, cell types, tissue regions, and evaluation metrics. We then evaluated four large language models—GPT-5.6-sol, Claude Opus 5, Gemini 3.1 Pro Preview, and DeepSeek V4 Pro—for automated annotation of cluster-level marker profiles. Marker-summary and model-specific performance were evaluated retrospectively on the same full cohort used for selection; performance remained dependent on clustering method, model, cell type, and tissue region. In an end-to-end Leiden–Claude pipeline, 56% of cells were correctly annotated, 29% were incorrectly clustered, and 15% were incorrectly annotated after clustering. These results show that scalable CODEX annotation is jointly constrained by image processing, cluster composition, marker representation, and model interpretation. Automated workflows can reduce manual annotation burden, but uncertainty-aware outputs and expert validation remain necessary, particularly for rare cell types and mixed clusters.

Keywords

Spatial omics, multiplexed tissue imaging, cell segmentation, unsupervised clustering, cell-type annotation, large language models

1. Introduction

Spatial proteomics technologies enable multiplexed measurement of protein expression while preserving tissue architecture. CO-Detection by indEXing (CODEX), for example, can measure dozens of protein markers in situ [1], [2]. By retaining spatial context, CODEX can identify cell types and reveal tissue architecture and local cellular neighborhoods across healthy organs, disease progression, immune responses, metastatic remodeling, and therapeutic delivery [3–7]. These capabilities have also supported large-scale atlasing efforts such as the Human BioMolecular Atlas Program (HuBMAP), which is generating reference datasets across the human body [8], [9].

However, cell-type annotation in these images requires several computational steps, including segmentation, feature extraction, dimensionality reduction, clustering, and marker-based interpretation [2], [10], [11]. Each step introduces challenges for accurate cell-type annotation. The quality of segmentation affects which pixels are assigned to each cell and therefore affects measured marker intensities [2], [11]. This has motivated the development of many machine-learning models designed to improve whole-cell segmentation across tissue-imaging platforms, and the choice of segmentation model has been shown to affect downstream cell-type classification [11], [12]. Moreover, compensation algorithms have been developed to correct marker-expression spillover between neighboring segmented cells [13]. Although these algorithms have previously improved cell-type classification [13], it is unclear whether they provide additional benefit when combined with newer cell-segmentation models [11], [14]–[18].

At the same time, benchmarking work in CODEX has shown that downstream cell type identification is sensitive not only to segmentation, but also to normalization and clustering methods [10]. Indeed, preprocessing decisions had as much or more influence on clustering quality than the clustering algorithm itself [10]. These findings suggest that building an end-to-end annotation workflow requires attention not only to one algorithm, but to the interaction between multiple steps in the pipeline.

Several clustering methods have been developed for high-dimensional single-cell and multiplexed imaging data, but they are based on different assumptions and have not been compared on CODEX tissues. Graph-based clustering methods such as Leiden are widely used in single-cell analysis because they are efficient and produce well-connected communities in similarity graphs [10], [19]. Another method arising from cytometry analysis is FlowSOM that applies self-organizing maps and metaclustering to identify cell types in protein-marker space [20], [21]. SpatialSort is a Bayesian model developed specifically for spatial-omics data that uses spatial neighborhood structure together with marker information and optional prior knowledge to improve clustering and population annotation [22]. PIXIE is another algorithm developed specifically for spatial data that performs pixel-level clustering to address issues with cell

segmentation [23]. Together, Leiden, FlowSOM, SpatialSort, and PIXIE represent four different modeling approaches: graph-based clustering, self-organizing map clustering, Bayesian spatial modeling, and pixel-based imaging analysis. Comparing them in one study provides a useful view of how methodological assumptions affect cell type separation in CODEX data.

Another major challenge is cluster annotation. Even if a clustering method separates cells well, researchers still need to assign cell type labels to the resulting clusters. Traditionally, this step is often manual, requiring domain experts to inspect marker-expression patterns and decide which cell type best matches each cluster. This process is slow, difficult to standardize, and often hard to reproduce across datasets or analysts. Earlier automated approaches, such as marker-based scoring, probabilistic assignment models, and reference-based transfer methods, have improved this process but still face challenges related to marker ambiguity, dataset shift, and ontology mismatch [24]–[28]. In spatial datasets, methods such as STELLAR have shown that label transfer can benefit from spatial context and learned representations [29]. However, these approaches still depend on specific references, prior label systems, or supervised training assumptions.

More recently, large language models (LLMs) have emerged as a possible tool for automated cell type annotation. Prior work has shown that LLMs such as GPT-4 can infer cell type names from marker gene lists with strong agreement to expert labels in single-cell RNA-sequencing studies [30]. Follow-up work has also highlighted the need for verification, since LLM outputs may vary in reliability and may require checks against observed marker expression [31]. Importantly, recent benchmarking has shown that different LLM families, including GPT, Gemini, Claude, and DeepSeek, can differ in annotation quality and reproducibility depending on prompt design, marker selection, and output constraints [32]. This suggests that model choice itself is an important experimental variable, though this work has largely focused on transcriptomic datasets. Whether the same principles apply and whether similar model performance is observed for spatial proteomic cluster annotation is unclear. This is increasingly important as others begin to rely on multi-agent frameworks to automate their data analysis pipelines.

This study addresses these gaps by evaluating segmentation compensation, clustering methods, and LLM cell type annotation for CODEX data. We evaluate compensation across multiple recent cell segmentation foundation models. We evaluate clustering across four distinct clustering methods and then evaluate GPT, Gemini, Claude, and DeepSeek for their ability to annotate the resulting clusters from marker patterns. By jointly studying clustering quality and downstream label assignment, this work aims to identify which combinations of clustering and annotation methods are most promising for accurate and scalable CODEX cell type annotation.

2. Methods

2.1 Compensation algorithm

For cell spillover compensation, we used the REDSEA algorithm [13]. The required inputs included the CODEX TIF file, marker channel list, and the segmentation mask produced by each method. The default parameters were used as specified in the repository: element-size=2, element-shape=star, markers-of-interest=all. These parameters define how the algorithm finds boundary cells, and which channels are included in compensation.

REDSEA begins by labeling cell pixels with their corresponding cell IDs, while boundaries are labeled with a value of 0. A 3 × 3 pixel grid searches along the boundary and identifies the cells located on either side to calculate their shared boundaries and perimeters. The element size and shape are used to determine marker signal around the cell boundaries.

The boundary signal associated with a cell is retained from each marker, while the weighted contribution from each neighboring cell boundary is subtracted. Corrected marker values below zero are set to zero.

2.2 Downstream processing and clustering

Cell compensation was applied separately to each tissue region and segmentation method. For each segmentation method, the outputs of each region before and after compensation were merged into separate datasets and processed using the same workflow. This included Z-normalization, noise filtering, feature extraction, and Leiden clustering at a resolution of 1.5 [19]. This produced paired before and after compensation H5AD files for each segmentation method and enabled comparisons of tissue-level clustering outputs (Figure 1A).

2.3 Cluster-quality metrics

Three cluster quality metrics were used to measure changes in cluster structures in PCA space.

The silhouette score measures the similarity of each cell to its assigned cluster relative to neighboring clusters [33], [34]. Values range from -1 to +1, with larger values indicating improved cluster assignment.

The Davies–Bouldin index measures the ratio of within-cluster dispersion to between-cluster separation [35]. Values range from 0 to infinity, with lower values indicating more compact and separated clusters.

The Calinski–Harabasz index compares between-cluster dispersion to within-cluster dispersion [36]. Values range from 0 to infinity, with larger values indicating improved separation relative to within-cluster variation.

These metrics were calculated both for the combined dataset associated with each segmentation method and separately for individual tissue regions. Changes were measured as the after compensation value minus the corresponding before compensation value. A positive change represents an improvement in the silhouette and Calinski-Harabasz scores, whereas a negative change represents an improvement in the Davies-Bouldin index.

2.4 Cell-type agreement

Cell-type agreement was evaluated using HuBMAP human-expert annotations as the reference. Preliminary cell types were assigned to clusters using the simple majority cell type among the cells matched via a nearest neighbors approach to the HubMAP reference. Marker-expression dot plots were generated as part of the clustering process, but these were not the basis for labeling cell types (Figure S1).

For each cell type and segmentation method, agreement between cluster-derived labels and the HuBMAP reference was summarized using the F1 score:

$$F1 = \frac{2 \times precision \times recall}{precision + recall}$$

The F1 score ranges from zero to one and represents the harmonic mean of precision and recall. A larger value indicates a greater agreement between the cluster-based labels and the reference annotation. Changes in F1 score were calculated as the after-compensation score minus the before-compensation score.

2.5 Study design and data source

We performed a retrospective benchmark of clustering and LLM annotation using a previously published HuBMAP CODEX dataset of healthy human large and small intestine [3]. Clustering inputs and reference annotations were linked by exact (File_ID, ID) keys. All 220,082 cells were matched, and no duplicate keys were present. The clustering input tables contained 220,082 cells from eight imaged tissue regions from one donor. The H5AD expression matrix contained 45 protein marker columns. The reference annotations were harmonized into 20 evaluation classes.

The clustering inputs differed by method and are summarized in Table S1. The Leiden configuration used 45 markers; FlowSOM, SpatialSort, and PIXIE used method-specific feature

representations. The cell was the primary unit of analysis, and the imaged tissue region was the unit used for regional comparisons. No new specimens were collected for this study.

We used the cell-type annotations distributed with the processed dataset as reference labels. The predominant harmonization map was Epithelial and CD57+ Epithelial to Enterocyte; CD66+ Epithelial to CD66+ Enterocyte; MUC1+ Epithelial to MUC1+ Enterocyte; ITLN+ Epithelial to Goblet; TA to Cycling TA; CD4+ T to CD4+ T cell; Stromal to Stroma; PDPN+ Stromal to Lymphatic; CD36 high Endothelial and CD36 low Endothelial to Endothelial; CD49a+ Smooth muscle to Smooth muscle; M1 Macrophage to M2 Macrophage; and NK to CD7+ Immune. These harmonized reference labels were used for retrospective evaluation only; they were not used to select clustering configurations and were not supplied to Leiden, FlowSOM, SpatialSort, or PIXIE during clustering.

2.6 Image processing and single-cell features

Raw CODEX images were processed with the published intestine workflow [3]. Cells were segmented with MESMER, and cells that failed segmentation or quality-control criteria were removed [11]. For each retained cell, x-y centroid coordinates were available in the processed H5AD, together with mean intensity for each marker within the cell mask.

Leiden and FlowSOM used protein marker expression features, whereas SpatialSort additionally used spatial-neighborhood information and PIXIE used imaging data. For FlowSOM, marker intensities were transformed as arcsinh(x/1), the 30 highest-variance features were retained, and random-forest-derived feature weights were applied before standardization; PCA was not used. SpatialSort used the 45-marker cell-by-marker H5AD matrix together with cell centroid coordinates and a within-region 24-nearest-neighbor graph. No marker-prior matrix or anchor cells were supplied. PIXIE used paired multichannel expression images and integer-labeled cell-mask OME-TIFFs. The 48 image channels were smoothed and normalized, then clustered at the pixel level; pixel assignments were aggregated within each segmentation mask to represent cells by the fractions of their pixels assigned to each pixel metacluster. These resulting cell-level pixel-composition profiles were used for PIXIE cell clustering. Spatial coordinates were used only by SpatialSort.

2.7 Clustering

Leiden used 45 markers transformed as arcsinh(marker / 5) and standardized within each field of view. A K320 local-Leiden hierarchy was generated and Ward-style coarsening produced K = 300 clusters; seed = 42 [19].

FlowSOM used a 32 × 32 MiniSom map, 300 metaclusters, sigma = 1.0, learning rate = 0.3, 10,000 iterations, and seed = 42 [20], [21].

SpatialSort used 300 clusters, 24 neighbors, precision scale = 0.65, eight trace iterations, one double-Metropolis-Hastings iteration, and seed = 42. Of the 300 configured clusters, 246 were occupied in the final assignment [22].

PIXIE used a 10 × 10 pixel MiniSom followed by 20 pixel Ward metaclusters, a 24 × 24 cell MiniSom, and 300 cell metaclusters. Sigma was 2.0, learning rate was 0.3, training used 5,000 iterations, and seed = 42 [23].

2.8 Clustering configurations and evaluation

We used fixed, method-specific clustering configurations for Leiden, FlowSOM, SpatialSort, and PIXIE. Reference labels were not used to select these configurations. After the configurations were fixed, the resulting clusters were evaluated against the harmonized reference labels using weighted cluster purity.

Let K denote the total number of clusters, N the total number of cells in the evaluation set, and $n_{k,c}$ the number of cells with reference class c in cluster k. Weighted cluster purity, which was subsequently used as the cell-level clustering upper bound, was calculated as

$$A_{UB} = \frac{\sum_{k=1}^{K} max_c n_{k,c}}{N}. \quad (1)$$

This score represents the fraction of cells recovered when every cluster is assigned its majority reference label.

The complete hyperparameters used are provided in Table S1.

2.9 Evaluation of clustering performance

After the clustering configurations were fixed, we mapped each cluster to its majority reference label and compared the mapped labels with the cell-level reference annotations. For each cell type, we calculated precision, recall, and F1 score, where F1 was the harmonic mean of precision and recall. In the regional F1 analysis, cluster-to-label mappings were defined separately within each imaged region. Cell types absent from a region were excluded, and overall F1 within a region was the reference-cell-count-weighted mean of the F1 values for cell types present in that region. The regional boxplot displayed the median, interquartile range, default 1.5-interquartile-range whiskers, and all eight regional observations. In contrast, the cell-type heatmaps defined cluster majorities globally across all regions.

2.10 Marker summaries for LLM annotation

We converted each cluster into a short list of defining protein markers using the 45-marker H5AD matrix. For every cluster-marker pair, we calculated mean expression across cells in that cluster and the fraction of cluster cells whose expression was strictly greater than the expression cutoff. A marker was retained when both this cluster-specific positive fraction and the cluster-specific mean strictly exceeded their respective thresholds; no global or between-cluster comparison was used. Retained markers were ranked by descending within-cluster mean expression. We kept at most the specified maximum number of markers. If no marker passed both filters, we used the specified minimum number of markers with the highest cluster mean expression; the code did not pad a nonempty list that contained fewer than the stated minimum. Persisted marker files were JSON objects mapping cluster identifiers to marker lists. In the only executable external-model prompt, these were wrapped in a payload containing the fixed 20-class allowed vocabulary, generic immunology and gut-epithelium task context, and an array of objects containing cluster_id, marker list, and clustering method. The prompt requested exactly one allowed label per cluster and a JSON-object response.

Exact prompts and example inputs and outputs are provided in Supplementary Note 1.

2.11 Marker-threshold optimization

Marker-summary candidate ranges and selected parameters are reported in Table S2. Parameters were selected retrospectively using the same full reference-labeled cohort used for evaluation. These settings were then used to generate the marker summaries supplied to the four LLMs.

2.12 LLM models and prompting conditions

The analysis used four LLMs: GPT-5.6-sol, Claude Opus 5, Gemini 3.1 Pro Preview, and DeepSeek V4 Pro. Each model was applied to each of the four clustering methods, yielding 16 method-model annotation maps.

2.13 Evaluation of LLM annotation

For each cluster, we defined its majority reference label and its model-predicted label. Using the notation introduced in Eq. (1), let $\hat{c}_k$ denote the model-predicted label assigned to cluster k, and let n_k denote the total number of cells in that cluster.

Absolute cell-level annotation accuracy was calculated as the number of cells whose model-assigned cluster label matched their harmonized reference label, divided by 220,082:

$$A_{cell} = \frac{\sum_{k=1}^{K} n_{k,\hat{c}_k}}{N}. \quad (2)$$

For the overall metrics, the denominator included all 220,082 cells, including 10,495 reference cells labeled Noise (4.77%). Noise was included in the allowed 20-class vocabulary, so assigning a non-Noise label to a reference Noise cell counted as an annotation error. The non-Noise restriction described below applied only to cell-type-specific analyses; the overall cell-level accuracy and clustering upper bound were calculated over the full cohort. Mean within-cluster annotation accuracy gave every cluster equal weight regardless of its number of cells and was calculated as

$$A_{cluster} = \frac{1}{K} \sum_{k=1}^{K} \frac{n_{k,\hat{c}_k}}{n_k}. \quad (3)$$

This cluster-balanced quantity was used only for the post-hoc selector analysis in Figure 3G; it was not the binary fraction of clusters whose predicted label equaled the majority reference label.

The clustering upper bound used to normalize LLM performance was the weighted cluster purity A_{UB} defined in Eq. (1). To express annotation performance relative to the recoverable fraction of the clustering solution, we calculated the percentage of the clustering upper bound as

$$P_{UB} = 100 \frac{A_{cell}}{A_{UB}} = 100 \frac{\sum_{k=1}^{K} n_{k,\hat{c}_k}}{\sum_{k=1}^{K} max_c n_{k,c}}. \quad (4)$$

Equivalently, the numerator was the total number of cells whose reference label matched the model label assigned to their cluster, and the denominator was the total number of cells matching their cluster's majority reference label. Thus, accidental matches to a minority reference type contributed to the numerator even when the predicted label differed from the cluster majority. For cell-type-specific values, global cluster majorities were first defined across retained non-Noise cells; both model accuracy and upper-bound accuracy were then calculated after restricting to all reference cells of the specified type, and their ratio was reported. We use cell-level annotation accuracy, mean within-cluster annotation accuracy, and percentage of the clustering upper bound for these quantities, reserving cluster purity for the majority-label fraction.

2.14 End-to-end Leiden-Claude analysis

Leiden clustering with Claude Opus 5 annotation was selected for the end-to-end analysis because it achieved the highest absolute annotation accuracy.

3. Results

3.1 Spillover compensation benefit is segmentation algorithm, region, and cell type dependent

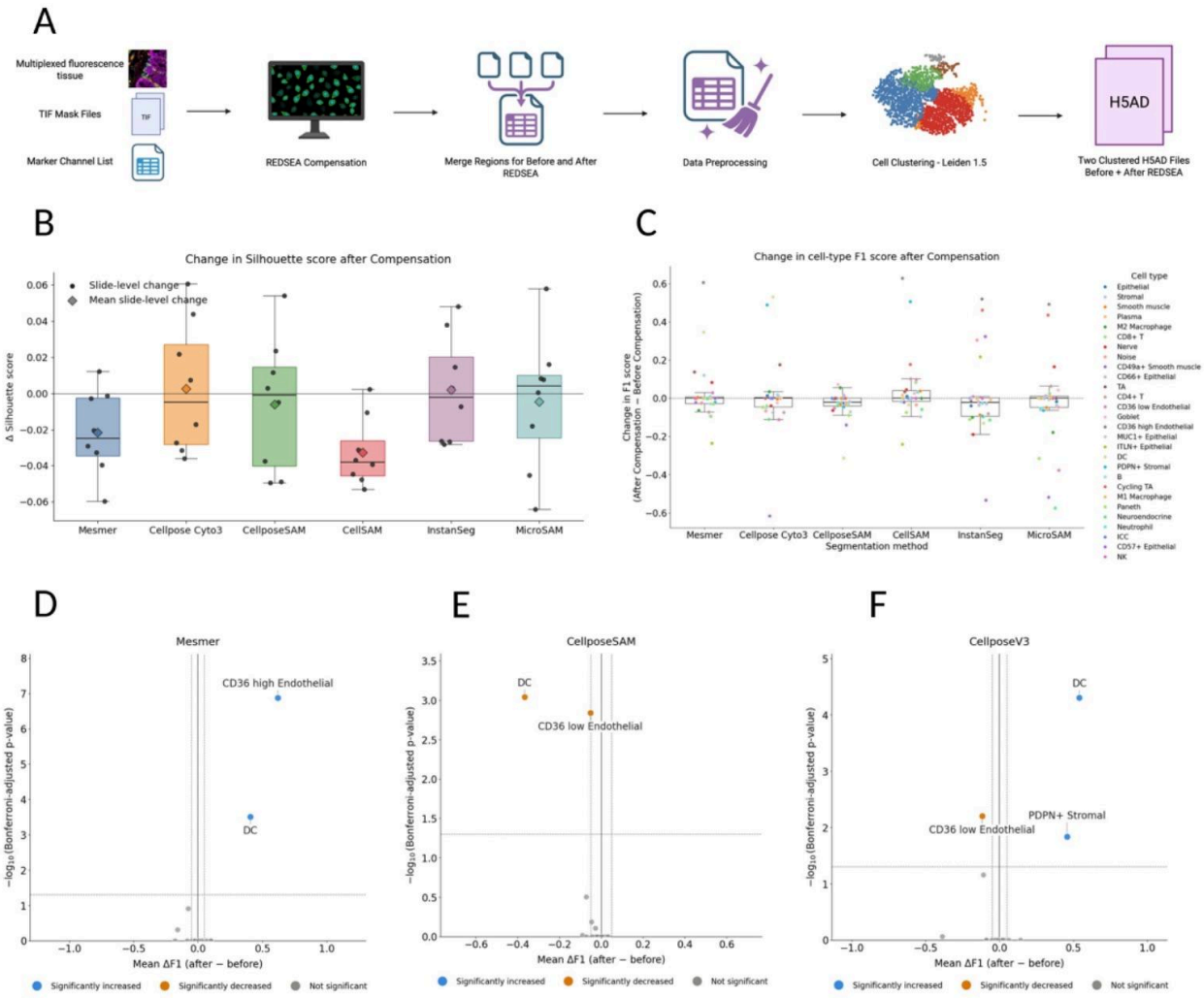


Figure 1. Effects of spillover compensation on downstream cell clustering and cell-type assignment. (A) Workflow for applying spillover compensation to CODEX images segmented using Mesmer, CellposeCyto3, CellposeSAM, CellSAM, InstanSeg, or MicroSAM. Tissue regions were compensated individually using REDSEA; compensated and uncompensated outputs were then merged, processed, clustered using Leiden clustering, and compared with HuBMAP expert cell-type annotations. (B) Change in silhouette score after compensation for each segmentation method. Points show region-level changes, diamonds show the change calculated using the combined dataset for each method, and boxes summarize the distributions of region-level changes. Positive values indicate increased silhouette scores after compensation. (C) Change in cell-type-specific F1 score after compensation across segmentation methods. Points represent individual cell types, and boxes summarize the distributions of cell-type-specific

changes for each method. Positive values indicate increased agreement with the HuBMAP annotations after compensation. (D–F) Cell-type-specific changes in F1 score plotted against permutation-based statistical significance for Mesmer, CellposeCyto3, and CellposeSAM, respectively. Blue points indicate significant positive changes, orange points indicate significant negative changes, and gray points indicate changes that were not statistically significant after adjustment for multiple comparisons. REDSEA, REinforcement Dynamic Spillover EliminAtion.

Marker spillover is a concern in multiplexed fluorescence imaging because signal originating from one cell may be assigned to neighboring cells when segmentation masks include shared or misclassified boundary pixels. This can produce apparent marker expression that does not reflect the biological phenotype of the receiving cell. Because marker intensities are subsequently used for dimensionality reduction, clustering, and cell-type annotation, spillover may alter both cluster structure and the inferred composition of cell populations. Spillover-compensation algorithms attempt to reduce these effects by estimating and removing signal contributed by neighboring cells [13]. With the advent of new segmentation algorithms that use distinct model architectures and produce masks with differing sizes and boundary geometries, we evaluated how spillover compensation impacts downstream clustering and cell-type annotation across six segmentation algorithms: Mesmer, CellposeCyto3, CellposeSAM, CellSAM, InstanSeg, and MicroSAM [11], [14]–[18] (Figure 1A).

We first evaluated spillover compensation using a previously annotated healthy human intestine CODEX multiplexed imaging dataset [3]. We chose this dataset since the cell type labels have been extensively evaluated and because the intestine represents a variety of cell types (e.g., immune, epithelial, and stromal) as well as microenvironments (dense and sparse cell areas) [3]. We compared the outputs from compensated and uncompensated data with the reference labels for each cell segmentation algorithm.

To evaluate the effect of compensation on overall clustering, we calculated silhouette scores separately for each tissue region. Spillover compensation did not consistently improve silhouette scores and produced substantial regional variation. CellposeCyto3 and InstanSeg showed slightly positive changes, whereas MicroSAM, Mesmer, CellSAM, and CellposeSAM showed negative changes (Figures 1B and S1A). Most regional silhouette scores remained below zero under both conditions, indicating that cell phenotypes form continuous or partially overlapping distributions rather than compact, well-separated groups [33], [34]. These data quantify a redistribution of cluster structure following compensation rather than a consistent shift in a particular direction.

Similar results were found using the Davies–Bouldin and Calinski–Harabasz metrics (Figure S1B–C). Davies–Bouldin scores generally increased following compensation, indicating greater within-cluster dispersion relative to separation from the most similar neighboring cluster (Figure

S1B) [35]. Likewise, the Calinski–Harabasz score increased for Mesmer and CellposeCyto3 but decreased for CellposeSAM, CellSAM, InstanSeg, and MicroSAM in the whole-method comparison (Figure S1C) [36]. Together, these results show that the downstream effects of spillover correction depended on both the segmentation method and the tissue region.

Beyond multidimensional cluster structure, we leveraged the reference cell type annotations to evaluate whether compensation altered the agreement between unsupervised Leiden clusters and expert cell-type annotations. We used Leiden clustering, assigned each cluster the reference-truth cell type represented by the greatest number of cells within that cluster, and calculated cell-type-specific F1 scores by comparing these assignments with the HuBMAP expert annotations. The direction and magnitude of these F1-score changes differed by both cell type and segmentation method (Figure 1C). Most F1-score changes were relatively close to zero, indicating limited changes in annotation agreement following compensation; however, several cell types showed larger positive or negative shifts (Figure 1D–F).

Dendritic cells showed particularly large variation. Mesmer and CellposeCyto3 exhibited positive changes in dendritic-cell F1 score, whereas CellposeSAM exhibited a negative change (Figure 1D). We observed increases in cells assigned to the dendritic-cell population for Mesmer and CellposeCyto3 and a decrease for CellposeSAM (Figure S2A). Similarly, Mesmer and CellposeCyto3 displayed marked reductions in the number of cells assigned to the CD8-positive T-cell population after compensation, whereas CellposeSAM showed a smaller spatial change (Figures 1E–F and S2B).

These findings indicate that correcting spillover does not necessarily translate into a uniform increase in clustering metrics or cell-type annotation agreement across segmentation algorithms. Differences in segmentation-mask size, boundary geometry, the amount of spillover present before correction, regional cellular density, and cell-type-specific marker distributions may influence whether compensation produces a measurable change in downstream cluster structure. Overall, spillover compensation did not consistently increase cluster separation or cell-type assignment agreement in this healthy human intestine dataset, and its measured effects depended on the segmentation algorithm, tissue region, and cell type examined.

3.2 Benchmarking unsupervised cell type clustering methods

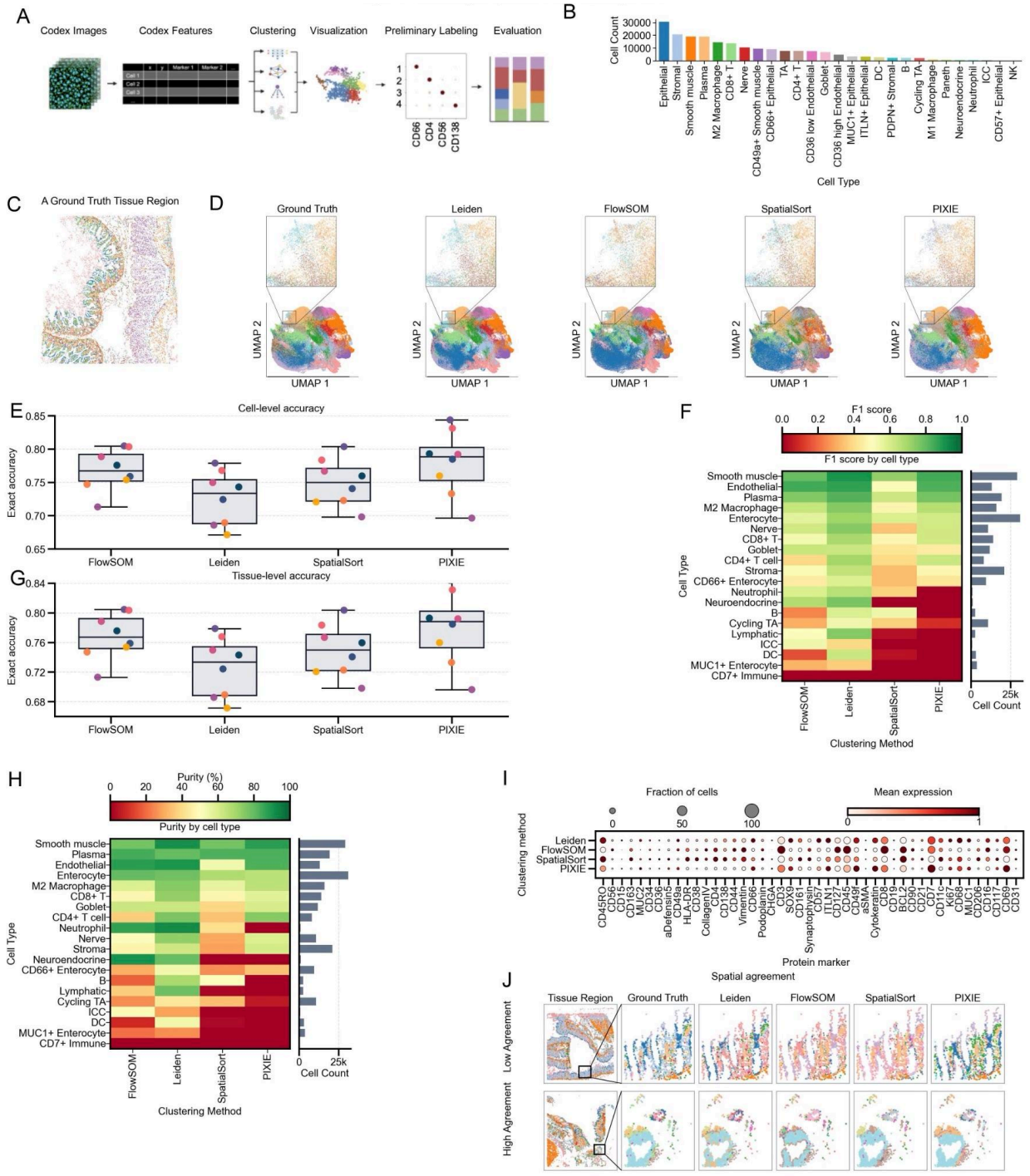


Figure 2. Benchmarking unsupervised cell-type clustering methods. (A) Workflow from CODEX images and cell-level features through clustering, visualization, preliminary majority-label assignment, and evaluation against reference cell types. (B) Reference cell counts by cell type in the healthy human intestine dataset. (C) Reference spatial cell-type map for one tissue region. (D) UMAP embeddings colored by reference or clustering-derived cell-type assignments for Leiden, FlowSOM, SpatialSort, and PIXIE; boxed regions are enlarged above

[37]. (E) Cell-level accuracy by clustering method across tissue regions. (F) F1 scores by cell type and clustering method, with reference cell counts at right. (G) Tissue-level accuracy by clustering method. (H) Cell-type recovery by clustering method, with reference cell counts at right. (I) Protein-marker expression profiles of cells assigned to CD8+ T-cell-mapped clusters by each clustering method. (J) Reference and method-derived spatial cell-type maps for representative regions with low or high clustering agreement. Cell-type colors in (C), (D), and (J) correspond to the legend in Figure 2B.

Clustering cell types based on protein-marker expression is difficult because some cell types are rare, some share similar marker profiles, and neighboring, separately segmented cells can have different biological identities. Therefore, spatial proximity does not necessarily indicate a shared cell type, particularly for methods that incorporate neighborhood information. As a result, clustering methods may group different cell types together or split one cell type into multiple clusters. To understand how the underlying unsupervised method affects these outcomes, we compared Leiden, FlowSOM, SpatialSort, and PIXIE to determine how well each method recovered known cell types.

To generate standardized inputs for clustering, we processed raw CODEX images from a healthy human intestine dataset [3]. We used an established preprocessing pipeline after cell segmentation and generated a single-cell feature table containing spatial coordinates and protein marker intensities [38] (Figure 2A). The reference cell type annotations represented in this dataset included both abundant and rare cell types (e.g., epithelial, NK), subphenotypes (e.g., epithelial, CD66+ epithelial), and cell types only present in certain locations (e.g., Paneth), providing a robust set of challenges for unsupervised clustering (Figures 2B–C and S3A–B).

We applied fixed, method-specific configurations for each clustering method; these configurations were not selected using reference-label purity. To facilitate comparison across methods, we set the number of clusters to 300 for each method. The final assignments contained 300 clusters for Leiden, FlowSOM, and PIXIE. SpatialSort was configured for 300 clusters, of which 246 were occupied in the final assignment (Figures 2D, S3C, and S3D).

To quantify recovery of annotated cell types, we calculated cell-type F1 scores and cell- and tissue-level accuracy across the eight imaged regions (Figures 2E–H and S3E). Performance varied across clustering methods, cell types, and regions, and no method was uniformly best across all comparisons. These differences indicate that clustering recovery was not uniform across annotated cell types and that some cell types were more difficult to recover than others.

Cell-type recovery also differed across methods and reference populations (Figure 2H). Some abundant populations with distinctive marker profiles showed strong recovery, whereas rare populations and those with overlapping marker profiles remained more difficult to separate.

These patterns further show that method performance depended on both the evaluation metric and the cell type considered.

CD8+ T cells were selected as a representative test case because they are an important intestinal immune population and can be difficult to separate from epithelial-associated regions, particularly when CD8+ intraepithelial lymphocytes are spatially intermixed with epithelial cells. All four clustering methods were evaluated using the same segmented cells and the same per-cell marker-expression measurements. For each method, clusters were mapped to cell types according to their global majority reference label, and the marker-expression dot plot was generated using the cells belonging to clusters mapped to CD8+ T cells (Figure 2I). Therefore, differences among the profiles do not result from differences in segmentation masks or changes in the marker measurements of individual cells. Instead, they reflect differences in which cells each clustering method grouped into CD8+ T-cell-mapped clusters and, consequently, the composition of those clusters. The resulting marker profiles differed across methods, consistent with differences in the cells grouped into CD8+ T-cell-mapped clusters. Figure 2I should therefore be interpreted as a visualization of cluster composition rather than evidence that clustering changed the marker expression of individual cells.

To determine whether clustering errors were spatially uniform across the tissue, we compared the cluster labels from tissue regions with low and high agreement with the reference cell type annotations (Figure 2J). In the low-agreement region, clustering accuracy varied substantially across methods. In contrast, all methods showed higher accuracy in the high-agreement region. These results are consistent with Figures 2E and 2G, respectively. These differences indicate that clustering performance was not spatially uniform across the tissue. However, the source of this regional variation may reflect multiple factors, including local cell-type composition, cell-type separability, segmentation or imaging artifacts, and staining noise. These results show that clustering agreement depends on region-specific properties of the image-derived single-cell data, rather than being uniform across the tissue.

Overall, these results show that unsupervised clustering can recover many major CODEX cell types, but performance varies by clustering method, cell-type abundance, marker specificity, tissue region, and evaluation metric.

3.3 Benchmarking large language model-assisted cell-type annotation

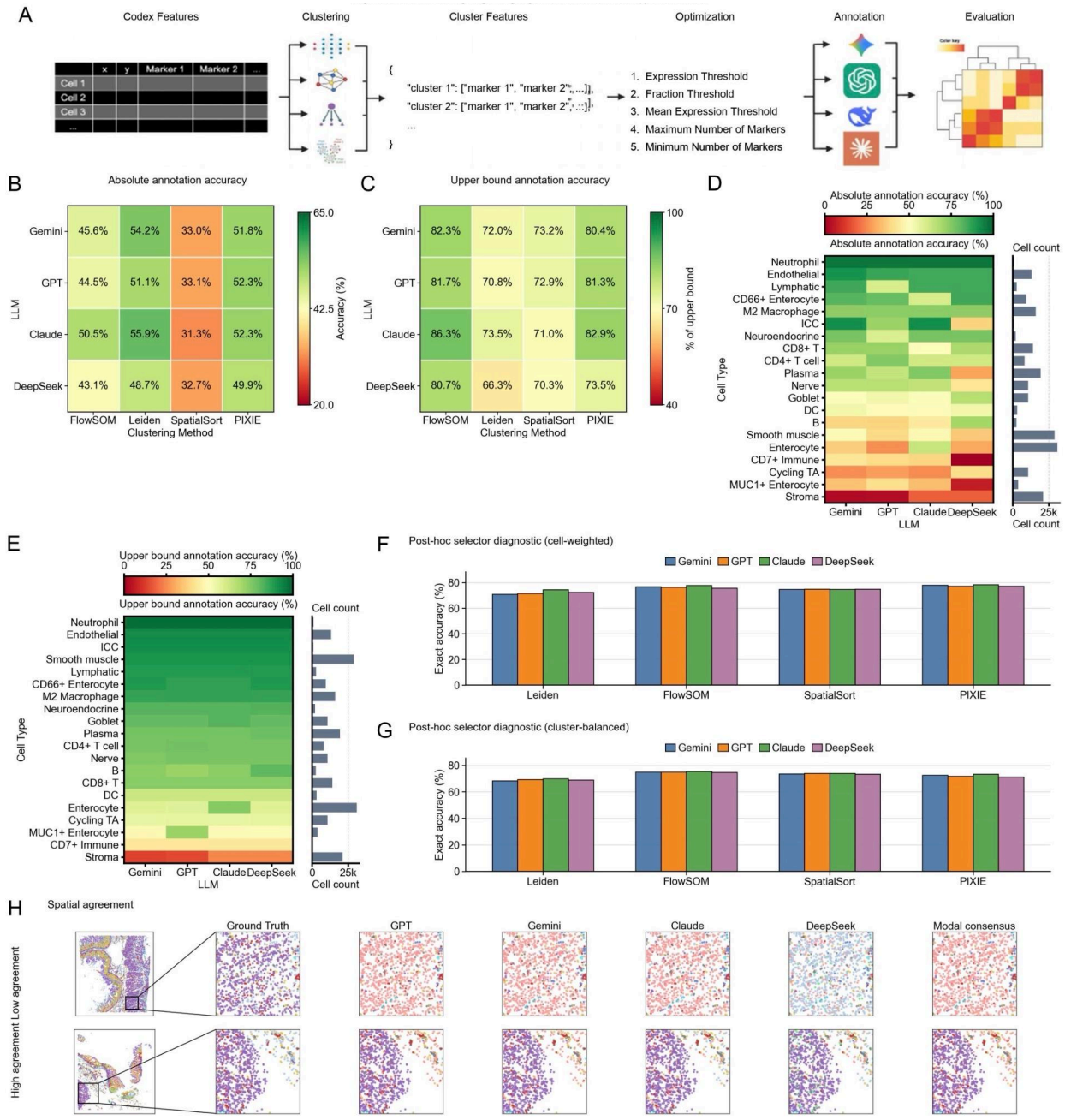


Figure 3. Benchmarking large language model cell-type annotation. (A) Workflow for converting cell-level CODEX features into clustering-derived marker summaries, retrospectively selecting marker-summary parameters, assigning labels with GPT-5.6-sol, Claude Opus 5, Gemini 3.1 Pro Preview, or DeepSeek V4 Pro, and evaluating predictions against reference labels. (B) Absolute cell-level annotation accuracy across the 16 raw method-model maps. (C) Annotation accuracy relative to the clustering upper bound. (D) Cell-type-specific absolute cell-level annotation accuracy for the four models applied to Leiden clusters. (E) Cell-type-specific annotation accuracy relative to the clustering upper bound. Bars at right in (D)

and (E) show reference cell counts. (F) Cell-weighted, truth-informed post-hoc source-union selector diagnostic. (G) Cluster-balanced, truth-informed post-hoc source-union selector diagnostic. (H) Reference and model-predicted spatial maps for representative low- and high-agreement regions. The modal-label map is descriptive and is not an independently validated consensus pipeline. Cell-type colors correspond to Figure 2B.

Beyond accurate separation of clusters, a central challenge is converting clustered marker-expression profiles into accurate and reproducible cell-type labels [23], [27]. Automated cell-type annotation methods have been developed for single-cell transcriptomic data, including marker-based and reference-based approaches, and recent work has shown that GPT-4 can annotate scRNA-seq cell types using marker-gene information [24], [26], [27], [30]. However, it remains unclear whether LLMs can reliably annotate cell types from spatial proteomic CODEX cluster summaries, where marker-threshold selection, clustering structure, model choice, and region-specific properties of the image-derived data may each affect performance. Therefore, we evaluated whether LLM annotation could recover CODEX reference cell-type labels and identified the factors associated with annotation accuracy.

To evaluate LLM cell-type labeling of CODEX-derived clusters, we assessed how retrospectively selected marker summaries, clustering method, and cell type were associated with annotation accuracy. This analysis was designed to distinguish two sources of error. The first was clustering error, in which cells from different reference cell types were assigned to the same cluster. The second was annotation error, in which the LLM assigned an incorrect cell-type label to a cluster whose majority cell type was recoverable from the clustering output. We therefore measured both absolute annotation accuracy and upper-bound annotation accuracy, as defined in Equations (1)–(4). The upper-bound metric estimated the highest cell-level annotation accuracy achievable if each cluster were assigned its reference-majority cell type, allowing us to separate errors caused by mixed cluster composition from errors caused by incorrect LLM label assignment.

To evaluate whether LLMs could annotate CODEX-derived cell clusters, we developed a pipeline that converted clustering outputs into structured marker-based inputs and tested factors that may affect annotation accuracy (Figure 3A). The LLM input was provided as a JSON object in which each key represented a cluster number and its corresponding value was an array containing that cluster's defining marker list. The LLM returned one cell-type label per cluster. That label was assigned to all cells in the cluster and compared with each cell's reference label to calculate absolute cell-level annotation accuracy.

Marker-summary parameters were retrospectively selected using the full reference-labeled cohort and are reported in Table S2.

Using the retrospectively selected marker-summary parameters reported in Table S2, absolute cell-level annotation accuracy across the 16 raw method-model maps ranged from 31.3% to 55.9%, while accuracy relative to the clustering upper bound ranged from 66.3% to 86.3% (Figure 3B–C). These are full-cohort accuracies and therefore include the 4.77% Noise reference class. Performance varied by method-model pair and cell type; no model or clustering method was uniformly best across all comparisons.

Cell-type-specific results further showed that some populations were annotated more accurately than others, consistent with differences in marker specificity and cluster composition (Figure 3D–E).

Spatial maps showed greater disagreement among model predictions in the representative low-agreement region than in the high-agreement region (Figure 3H). The modal-label visualization summarizes agreement descriptively but does not replace model-specific evaluation or expert review.

3.4 End-to-end evaluation of the Leiden clustering and Claude annotation pipeline

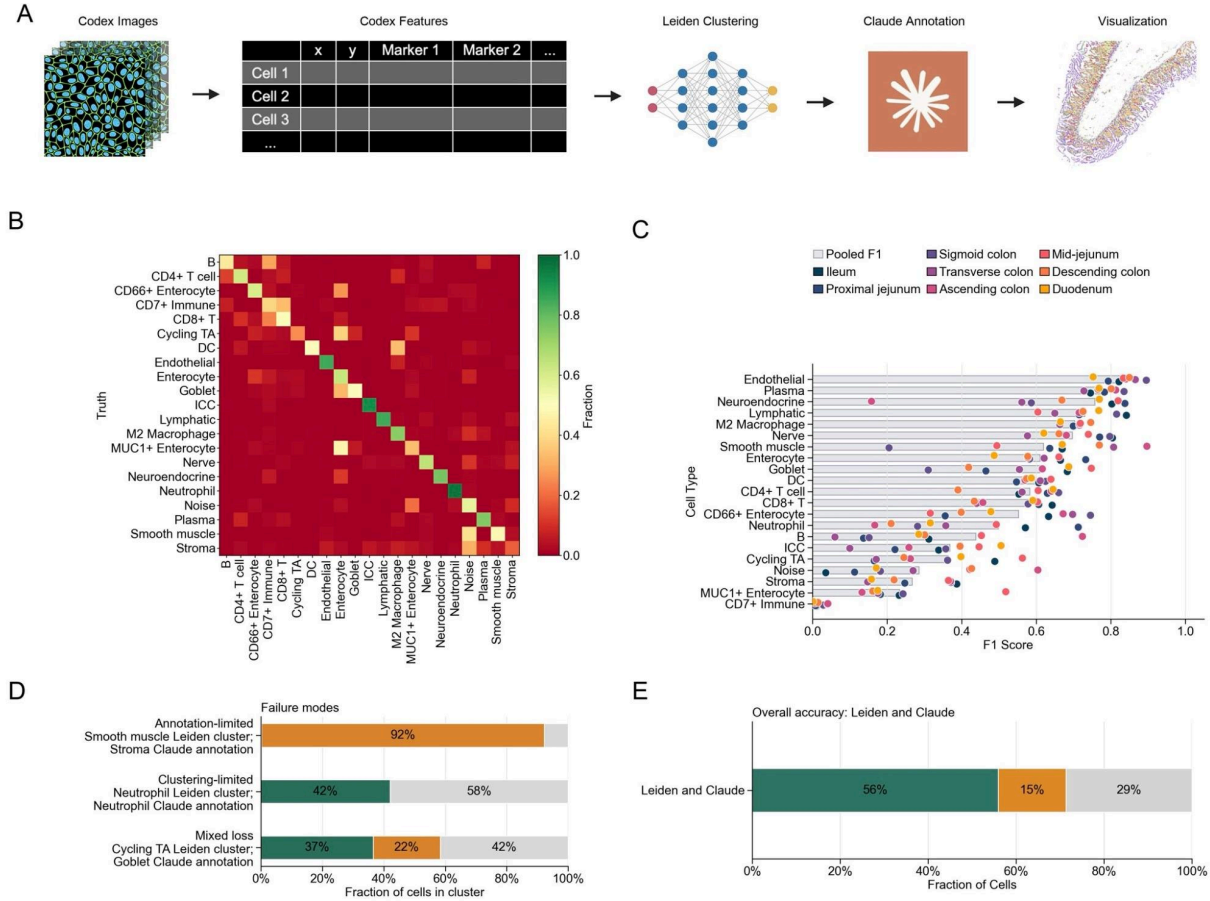


Figure 4. End-to-end evaluation of the Leiden clustering and Claude annotation mapping. (A) Workflow from CODEX images and cell-level feature extraction through Leiden clustering,

Claude-based cluster annotation, and spatial visualization. (B) Row-normalized confusion matrix comparing reference and Leiden–Claude-predicted cell-type labels. Rows indicate reference cell types, columns indicate predicted labels, and color represents the fraction of cells within each reference cell type. (C) Cell-type-specific F1 scores for the Leiden–Claude mapping; points show pooled and tissue-region-specific values. (D) Representative Leiden clusters illustrating annotation-limited, clustering-limited, and mixed failure modes. Stacked bars show cells correctly annotated by the final mapping (green), incorrectly annotated after clustering (orange), or incorrectly clustered (gray). "Cluster" indicates the reference-majority cell type, and "Annotation" indicates the Claude-predicted label. (E) Overall fractions of cells in the three outcome categories across the complete dataset. In (D)–(E), cells belonging to a cluster's reference-majority population were considered correctly annotated when Claude selected the majority label and incorrectly annotated after clustering when Claude selected another label; non-majority cells were considered incorrectly clustered.

After separately evaluating clustering and LLM annotation performance, we next asked whether the strongest method from each stage could be combined into a single pipeline for cell type annotation. This question is important because CODEX and other multiplexed imaging generate high-dimensional single-cell spatial data that require both clustering and annotation before tissue organization can be interpreted [1], [37]. Leiden was chosen as the best overall-performing clustering method because it provided the most consistent cell-level recovery across annotation conditions, whereas PIXIE performed better primarily at the cluster level and FlowSOM outperformed Leiden only under selected conditions. Claude Opus 5 was selected for the combined pipeline because it provided the highest absolute annotation accuracy when paired with Leiden clustering. We combined Leiden clustering with Claude annotation and evaluated the resulting pipeline using cluster-to-label mapping, confusion matrices, F1 scores, and cluster purity analysis (Figure 4A).

We characterized end-to-end Leiden–Claude performance using a row-normalized confusion matrix and cell-type-specific F1 scores (Figure 4B–C). Performance varied across cell types and tissue regions, with stronger recovery for some well-resolved populations and lower F1 scores for more ambiguous, rare, noisy, or marker-overlapping populations.

To determine whether absolute accuracy was limited by Leiden clustering or Claude annotation, we categorized cells as correctly annotated, incorrectly annotated after clustering, or incorrectly clustered.

To calculate these categories, we first identified the reference-majority cell type in each Leiden cluster. This majority label represented the cell type that could recover the largest number of cells if the cluster received a single correct annotation. Cells whose reference label matched their cluster's majority label were therefore considered recoverable by clustering. Among these cells,

those for which the Claude-assigned label also matched the majority label were counted as correctly annotated, whereas those in clusters assigned a different Claude label were counted as incorrectly annotated after clustering. Cells whose reference label did not match their cluster's majority label were counted as incorrectly clustered because they could not be recovered through majority-vote labeling of that cluster. Each percentage was calculated by dividing the number of cells in the category by the total number of cells. The correctly annotated and incorrectly annotated-after-clustering categories together represented the cell-level clustering upper bound, while the incorrectly clustered category represented the remaining cells.

Representative clusters showed distinct failure modes, including an annotation-limited smooth muscle cluster that Claude labeled Stroma, a clustering-limited neutrophil cluster that Claude labeled Neutrophil, and a mixed Cycling TA cluster that Claude labeled Goblet (Figure 4D). These examples show that cluster composition constrains the maximum recoverable accuracy, while incorrect Claude label assignment can further reduce recovery for otherwise recoverable cells.

Across all cells, the Leiden–Claude mapping correctly annotated 56% of cells, while 15% were incorrectly annotated after clustering and 29% were incorrectly clustered (Figure 4E). Overall, Leiden cluster composition defined the maximum recoverable accuracy, but Claude annotation determined whether that potential accuracy was achieved.

The combined analysis illustrates how cluster composition and model label selection interact in practice. High-purity Leiden clusters can be correctly annotated when Claude selects their reference-majority label, but even relatively pure clusters can be incorrectly annotated after clustering when the assigned label is inconsistent with their marker profile. These findings show that end-to-end performance is constrained by both cluster composition and LLM label assignment, supporting uncertainty-aware outputs and expert review for mixed, rare, low-confidence, or unexpected populations.

4. Discussion

In this study, we compared different factors affecting cell-type annotation in spatial proteomics single-cell data. Spillover compensation altered clustering and cell-type assignment, but its effects varied across segmentation methods, tissue regions, and cell populations. Across the four clustering methods, no method was best on every measure. Performance varied across clustering methods, cell types, tissue regions, models, and evaluation metrics, with no method uniformly best across all comparisons. Retrospectively selected marker summaries and model-specific conditions also showed variable annotation performance.

The clustering results support earlier work showing that preprocessing and clustering choices can strongly affect cell-type identification in CODEX data [10]. All final clustering methods targeted 300 clusters, although SpatialSort produced 246 occupied clusters. Performance depended on the evaluation metric and analysis goal rather than identifying one universally best method. These differences should be considered when choosing among graph-based, self-organizing-map, spatial Bayesian, and pixel-based methods [19–23].

Performance also differed by cell type and tissue region. Cell types with clear marker profiles, including several endothelial, plasma, macrophage, and T-cell populations, were easier to recover as expected. Rare cell types and populations with overlapping markers were also more difficult for expert annotators to distinguish. The CD8+ T-cell analysis showed a similar pattern. Clustering methods with better recovery kept a clearer T-cell marker profile, while weaker methods included more markers from nearby or similar cell types. Moreover, the same tissue region with noisy data was also difficult for every clustering method and every LLM. In contrast, all methods performed better in areas with low noise. This pattern suggests that local staining quality, mixed signal between nearby cells, tissue composition, and marker quality can limit performance. Changing the clustering method alone may not solve these problems. For this reason, overall accuracy should be reported together with cell-type-level and region-level results. Areas where several methods disagree may also be useful targets for image and marker review.

The LLM results show that marker-summary construction is an important part of the annotation process. The four LLMs differed across clustering methods and cell types, showing that model identity and input representation both influenced annotation performance.

The end-to-end Leiden–Claude analysis shows how clustering and annotation errors build on each other. Clustering accounted for the larger fraction of errors in this mapping, but Claude also made annotation errors. The prompt should allow the model to report confidence, supporting and conflicting markers, and “unknown” or “mixed” labels when evidence is weak. Low-confidence clusters can then be reviewed by experts using both marker profiles and spatial context.

Clustering was the larger source of error because cluster composition set the maximum accuracy that Claude could reach. However, Claude still made important errors of its own. For example, Claude labeled a relatively pure smooth muscle cluster as Stroma (Figure 4D). This shows that cells in a relatively pure cluster can still be incorrectly annotated after clustering when the model interprets its markers incorrectly. The current prompt required Claude to choose one label from a fixed list, even when the evidence was weak or mixed. Future systems should allow the model to report confidence, list supporting and conflicting markers, and return "unknown" or "mixed." Low-confidence clusters could then be sent to an expert for review of their markers and spatial location. High-confidence clusters could require less manual review. This approach would reduce repetitive annotation work while keeping expert review for difficult or unusual cases.

This type of pipeline could be useful for large spatial atlases. As HuBMAP and related projects add more tissues, donors, regions, and cells, fully manual annotation will become harder to scale and standardize [8], [9]. Automated methods could label clear cell populations, save a record of each decision, and direct experts toward rare, mixed, or unexpected clusters. The goal is not to remove experts from the process. Instead, automation could help experts spend more time on the cases that require biological judgment.

This study also has several limitations. The benchmark used 220,082 cells from eight regions from one donor with healthy intestinal tissue, and results may not generalize to other donors, diseases, tissues, or marker panels. Reference labels may contain errors or uncertain boundaries, and harmonization reduces biological detail. Because Noise was retained as a scored reference class, misclassification of Noise cells contributed to the direct annotation error rate. However, the clustering upper bound was determined by within-cluster reference-label composition, not simply by the presence of Noise cells. Marker-summary selection and source selection were performed retrospectively on the full cohort using the same reference labels used for evaluation. The clustering configurations were not selected using reference-label purity.

5. Conclusion

Together, these results show that accurate and scalable CODEX annotation depends on multiple connected decisions. Spillover compensation should be evaluated within the specific imaging, segmentation, regional, and biological context rather than assumed to improve results uniformly. Clustering and model performance reflected method- and model-specific tradeoffs rather than a single universally best choice. Careful marker-summary construction can support LLM annotation, but cluster composition sets an upper limit on end-to-end recovery and even relatively pure clusters can be mislabeled.

For users starting a new CODEX dataset, a reproducible clustering and model configuration should be validated on a manually reviewed subset or independent data before broader use. LLM

labels should be treated as helpful initial annotations rather than replacements for expert review. Expert review should focus on low-confidence, mixed, rare, unexpected, or spatially inconsistent populations. Allowing models to report confidence and return “unknown” or “mixed” labels could further reduce annotation burden while maintaining biological oversight.

Acknowledgements

The authors have no acknowledgments to declare.

Author contributions

Deutsch Z, Surguladze N, Miao Y, Hickey JW: Study conception and design; review and interpretation of the data analyses; manuscript writing, review, and editing. Surguladze N: Analyses related to Figure 1. Deutsch Z: Analyses related to Figures 2–4. Hickey JW: Study supervision and funding acquisition. All authors read and approved the final manuscript.

Conflicts of interest

The authors declare no conflicts of interest.

Ethical Approval

This study involved secondary analysis of deidentified human tissue imaging data previously collected and made publicly available through the Human BioMolecular Atlas Program (HuBMAP). No new participants were recruited and no new tissue was collected. The original tissue collection complied with all relevant ethical regulations and was approved by the Washington University Institutional Review Board and the Stanford University Institutional Review Board. No additional ethical approval was required for the present secondary analysis.

Consent to Participate

No participants were newly enrolled in this study. For the original tissue collection, written informed consent was obtained from the next of kin of all deceased organ donors.

Consent for Publication

Not applicable. No identifiable participant information is included in this manuscript.

Availability of Data and Materials

Raw CODEX images and related datasets are available through the Human BioMolecular Atlas Program (HuBMAP) Data Portal under HuBMAP Collection HBM692.JRZB.356 (https://doi.org/10.35079/HBM692.JRZB.356). Processed, quantified, and annotated single-cell CODEX data are available through Dryad (https://doi.org/10.5061/dryad.pk0p2ngrf). Additional expert-annotated CODEX datasets, segmentation outputs, and phase-contrast and fluorescence microscopy images used in the segmentation analyses are available through the Duke Research Data Repository (https://doi.org/10.7924/r4r505).

All code used for data analysis and figure generation is available through the project's GitHub repository (https://github.com/HickeyLab/LLM-Spatial-omics-Clustering).

Funding Information

This work was supported by the U.S. National Institutes of Health (grant nos. 3U54AG07593, 3OT2OD033759-01S4, 3OT2OD033759-01S5, 1U01-AI186999-01, and 1U54-AI191253-01); the National Science Foundation CAREER Award (grant no. 2440733); the V Foundation (grant no. V2025-019); the Human Frontier Science Program Early Career Grant (grant no. RGEC26/2025); and start-up support from the Duke University Department of Biomedical Engineering to J.W.H.

References

1. Goltsev Y, Samusik N, Kennedy-Darling J, Bhate S, Hale M, Vazquez G, et al. Deep profiling of mouse splenic architecture with CODEX multiplexed imaging. *Cell*. 2018;174(4):968-981.e15. DOI: 10.1016/j.cell.2018.07.010.

2. Black S, Phillips D, Hickey JW, Kennedy-Darling J, Venkataraaman VG, Samusik N, et al. CODEX multiplexed tissue imaging with DNA-conjugated antibodies. *Nat Protoc.* 2021;16(8):3802-3835. DOI: 10.1038/s41596-021-00556-8.

3. Hickey JW, Becker WR, Nevins SA, Horning A, Perez AE, Zhu C, et al. Organization of the human intestine at single-cell resolution. *Nature*. 2023;619(7970):572-584. DOI: 10.1038/s41586-023-05915-x.

4. Lu G, Hickey JW, Haist M, Qin X, Zhao E, Naveed A, et al. Single-cell spatial pharmacobiology identifies conserved stromal barriers to therapeutic antibody delivery in human solid tumors. *Nat Biotechnol.* 2026. DOI: 10.1038/s41587-026-03152-x.

5. Haist M, Baertsch MA, Reticker-Flynn NE, Lu G, Kempchen TN, Chu P, et al. Lymph node colonization induces tissue remodeling via immunosuppressive fibroblast-myeloid cell niches supporting metastatic tolerance. *Cancer Cell*. 2026;44(3):604-623.e9. DOI: 10.1016/j.ccell.2026.01.003.

6. Strasser MK, Gibbs DL, Gascard P, Bons J, Hickey JW, Pan D, et al. Concerted changes in epithelium and stroma: a multi-scale, multi-omics analysis of progression from Barrett's esophagus to adenocarcinoma. *Dev Cell*. 2025;60(20):2807-2824.e7. DOI: 10.1016/j.devcel.2025.06.034.

7. Hickey JW, Haist M, Horowitz N, Caraccio C, Tan Y, Rech AJ, et al. T cell-mediated curation and restructuring of tumor tissue coordinates an effective immune response. *Cell Rep.* 2023;42(12):113494. DOI: 10.1016/j.celrep.2023.113494.

8. HuBMAP Consortium. The human body at cellular resolution: the NIH Human BioMolecular Atlas Program. *Nature*. 2019;574(7777):187-192. DOI: 10.1038/s41586-019-1629-x.

9. Börner K, Blood PD, Silverstein JC, Ruffalo M, Satija R, Teichmann SA, et al. Human BioMolecular Atlas Program (HuBMAP): 3D Human Reference Atlas construction and usage. *Nat Methods*. 2025;22(4):845-860. DOI: 10.1038/s41592-024-02563-5.

10. Hickey JW, Tan Y, Nolan GP, Goltsev Y. Strategies for accurate cell type identification in CODEX multiplexed imaging data. *Front Immunol.* 2021;12:727626. DOI: 10.3389/fimmu.2021.727626.

11. Greenwald NF, Miller G, Moen E, Kong A, Kagel A, Dougherty T, et al. Whole-cell segmentation of tissue images with human-level performance using large-scale data annotation and deep learning. *Nat Biotechnol.* 2022;40(4):555-565. DOI: 10.1038/s41587-021-01094-0.

12. Miao Y, Surguladze N, Lerner J, Poysungnoen K, Ariano K, Li Y, et al. Foundation cell segmentation models performance on live microscopy and spatial-omics data [Preprint]. 2026. DOI: 10.64898/2026.04.18.719315.

13. Bai Y, Zhu B, Rovira-Clavé X, Chen H, Markovic M, Chan CN, et al. Adjacent cell marker lateral spillover compensation and reinforcement for multiplexed images. *Front Immunol.* 2021;12:652631. DOI: 10.3389/fimmu.2021.652631.

14. Stringer C, Pachitariu M. Cellpose3: one-click image restoration for improved cellular segmentation. *Nat Methods*. 2025;22(3):592-599. DOI: 10.1038/s41592-025-02595-5.

15. Pachitariu M, Rariden M, Stringer C. Cellpose-SAM: superhuman generalization for cellular segmentation [Preprint]. 2025. DOI: 10.1101/2025.04.28.651001.

16. Marks M, Israel U, Dilip R, Li Q, Yu C, Laubscher E, et al. CellSAM: a foundation model for cell segmentation. *Nat Methods*. 2025;22(12):2585-2593. DOI: 10.1038/s41592-025-02879-w.

17. Goldsborough T, Philps B, O'Callaghan A, Inglis F, Leplat L, Filby A, et al. InstanSeg: an embedding-based instance segmentation algorithm optimized for accurate, efficient and portable cell segmentation [Preprint]. 2024. DOI: 10.48550/arXiv.2408.15954.

18. Archit A, Freckmann L, Nair S, Khalid N, Hilt P, Rajashekar V, et al. Segment Anything for Microscopy. *Nat Methods*. 2025;22(3):579-591. DOI: 10.1038/s41592-024-02580-4.

19. Traag VA, Waltman L, van Eck NJ. From Louvain to Leiden: guaranteeing well-connected communities. *Sci Rep.* 2019; 9(1):5233. DOI: 10.1038/s41598-019-41695-z.

20. Van Gassen S, Callebaut B, Van Helden MJ, Lambrecht BN, Demeester P, Dhaene T, et al. FlowSOM: using self-organizing maps for visualization and interpretation of cytometry data. *Cytometry A*. 2015;87(7):636-645. DOI: 10.1002/cyto.a.22625.

21. Couckuyt A, Rombaut B, Saeys Y, Van Gassen S. Efficient cytometry analysis with FlowSOM in Python boosts interoperability with other single-cell tools. *Bioinformatics*. 2024;40(4):btae179. DOI: 10.1093/bioinformatics/btae179.

22. Lee E, Chern K, Nissen M, Wang X, IMAXT Consortium, Huang C, et al. SpatialSort: a Bayesian model for clustering and cell population annotation of spatial proteomics data. *Bioinformatics*. 2023;39(Suppl 1):i131-i139. DOI: 10.1093/bioinformatics/btad242.

23. Liu CC, Greenwald NF, Kong A, McCaffrey EF, Leow KX, Mrdjen D, et al. Robust phenotyping of highly multiplexed tissue imaging data using pixel-level clustering. *Nat Commun.* 2023;14(1):4618. DOI: 10.1038/s41467-023-40068-5.

24. Ianevski A, Giri AK, Aittokallio T. Fully automated and ultra-fast cell-type identification using specific marker combinations from single-cell transcriptomic data. *Nat Commun.* 2022;13(1):1246. DOI: 10.1038/s41467-022-28803-w.

25. Zhang AW, O'Flanagan C, Chavez EA, Lim JLP, Ceglia N, McPherson A, et al. Probabilistic cell-type assignment of single-cell RNA-seq for tumor microenvironment profiling. *Nat Methods*. 2019;16(10):1007-1015. DOI: 10.1038/s41592-019-0529-1.

26. Aran D, Looney AP, Liu L, Wu E, Fong V, Hsu A, et al. Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. *Nat Immunol.* 2019;20(2):163-172. DOI: 10.1038/s41590-018-0276-y.

27. Abdelaal T, Michielsen L, Cats D, Hoogduin D, Mei H, Reinders MJT, et al. A comparison of automatic cell identification methods for single-cell RNA sequencing data. *Genome Biol.* 2019;20(1):194. DOI: 10.1186/s13059-019-1795-z.

28. Osumi-Sutherland D, Xu C, Keays M, Levine AP, Kharchenko PV, Regev A, et al. Cell type ontologies of the Human Cell Atlas. *Nat Cell Biol.* 2021;23(11):1129-1135. DOI: 10.1038/s41556-021-00787-7.

29. Brbić M, Cao K, Hickey JW, Tan Y, Snyder MP, Nolan GP, et al. Annotation of spatially resolved single-cell data with STELLAR. *Nat Methods*. 2022;19(11):1411-1418. DOI: 10.1038/s41592-022-01651-8.

30. Hou W, Ji Z. Assessing GPT-4 for cell type annotation in single-cell RNA-seq analysis. *Nat Methods*. 2024;21(8):1462-1465. DOI: 10.1038/s41592-024-02235-4.

31. Ye W, Ma Y, Xiang J, Liang H, Luo J, Li Y, et al. Evaluation of cell type annotation reliability using a large language model-based identifier. *Commun Biol.* 2025;8(1):1360. DOI: 10.1038/s42003-025-08745-x.

32. Xiao T, Hua D, Wang Y, Lu X, Zhang C. Benchmarking large language models for cell typing in single-cell RNA-seq. *Brief Bioinform.* 2025;26(6):bbaf677. DOI: 10.1093/bib/bbaf677.

33. Rousseeuw PJ. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J Comput Appl Math.* 1987;20:53-65. DOI: 10.1016/0377-0427(87)90125-7.

34. Gagolewski M, Bartoszuk M, Cena A. Are cluster validity measures (in)valid? *Inf Sci.* 2021;581:620-636. DOI: 10.1016/j.ins.2021.10.004.

35. Davies DL, Bouldin DW. A cluster separation measure. *IEEE Trans Pattern Anal Mach Intell.* 1979;PAMI-1(2):224-227. DOI: 10.1109/TPAMI.1979.4766909.

36. Caliński T, Harabasz J. A dendrite method for cluster analysis. *Commun Stat.* 1974;3(1):1-27. DOI: 10.1080/03610927408827101.

37. McInnes L, Healy J, Saul N, Großberger L. UMAP: uniform manifold approximation and projection. *J Open Source Softw.* 2018;3(29):861. DOI: 10.21105/joss.00861.

38. Windhager J, Zanotelli VRT, Schulz D, Meyer L, Daniel M, Bodenmiller B, et al. An end-to-end workflow for multiplexed image processing and analysis. *Nat Protoc.* 2023;18(11):3565-3613. DOI: 10.1038/s41596-023-00881-0.

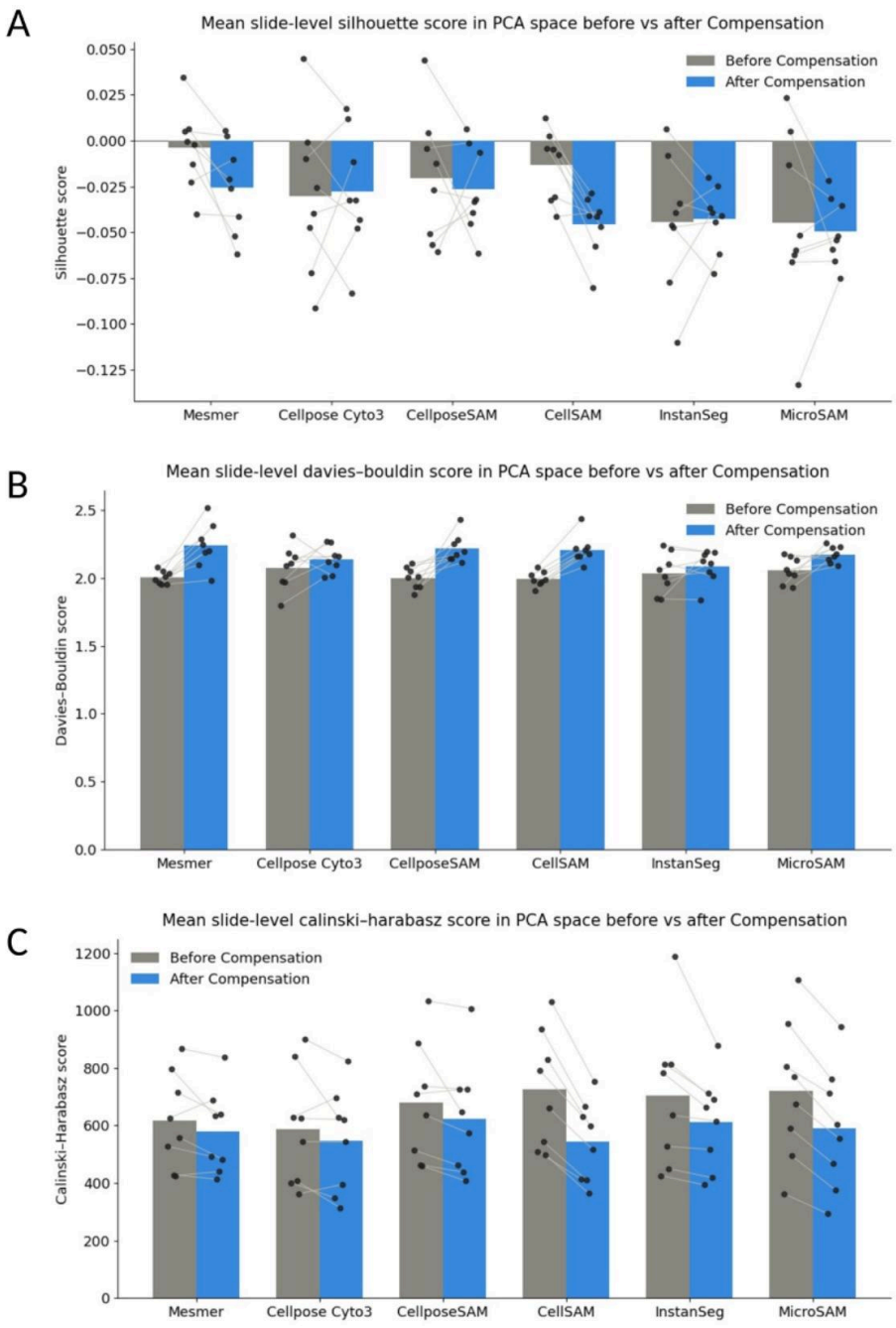


Figure S1. Cluster-quality metrics before and after spillover compensation across segmentation methods. (A) Slide-level silhouette scores, (B) Davies–Bouldin scores, and (C) Calinski–Harabasz scores calculated in principal component analysis (PCA) space for data segmented using Mesmer, CellposeCyto3, CellposeSAM, CellSAM, InstanSeg, or MicroSAM. Gray and blue bars show the mean scores before and after REDSEA compensation, respectively. Each point represents one tissue region (n = 8), and lines connect paired measurements from the same region. Higher silhouette and Calinski–Harabasz scores indicate improved cluster separation, whereas lower Davies–Bouldin scores indicate improved cluster separation.

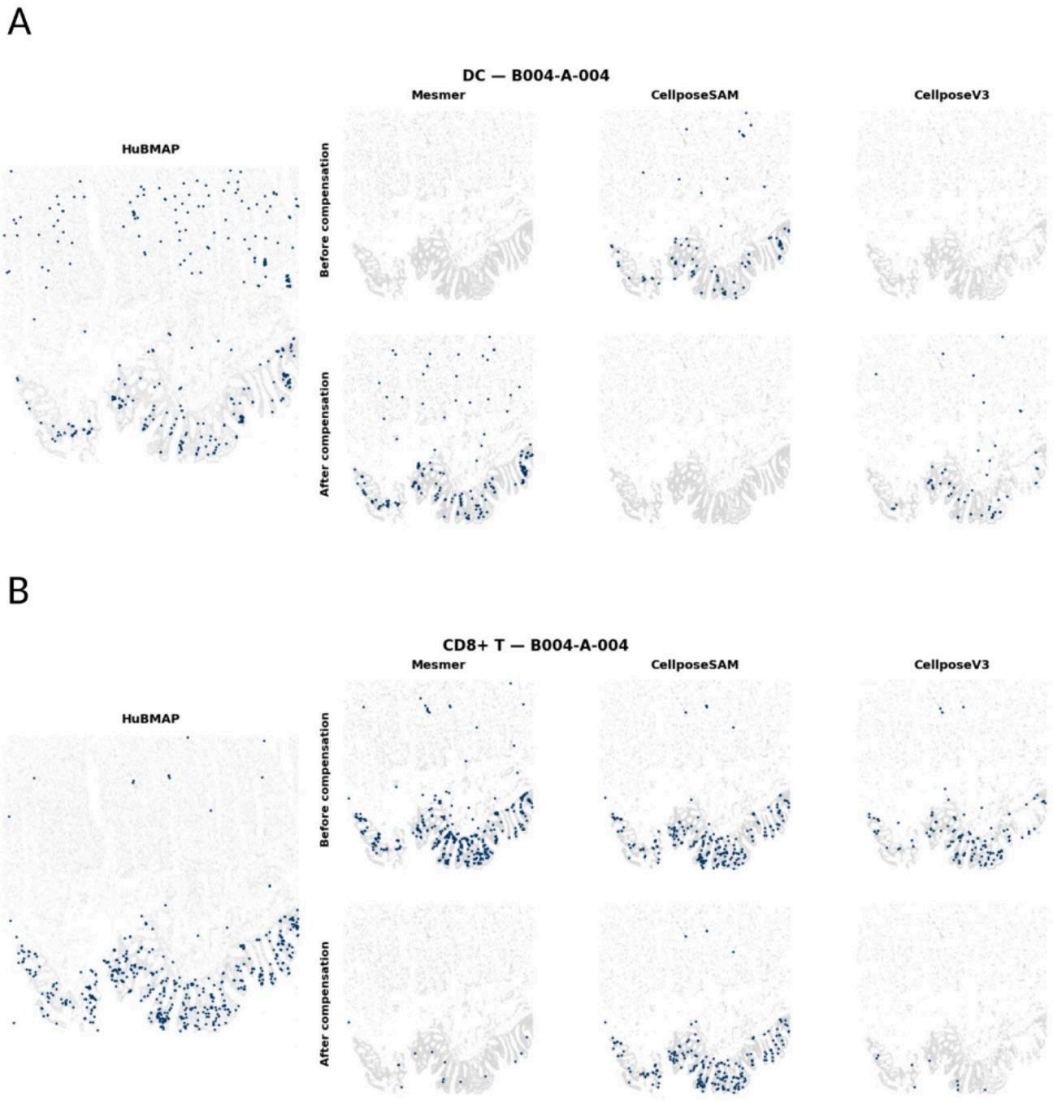


Figure S2. Spatial changes in dendritic-cell and CD8+ T-cell assignments after spillover compensation. Representative spatial maps from tissue region B004-A-004 compare HuBMAP expert annotations with cluster-derived cell-type assignments from data segmented using Mesmer, CellposeSAM, or CellposeCyto3 before and after REDSEA compensation. (A) Dendritic cells. Compensation increased the number of cells assigned as dendritic cells for Mesmer and CellposeCyto3 but decreased the assigned population for CellposeSAM. (B) CD8+ T cells. Compensation substantially reduced the number of cells assigned as CD8+ T cells for Mesmer and CellposeCyto3, whereas CellposeSAM showed a smaller spatial change. Dark blue indicates the specified cell type, and light gray indicates all other cells.

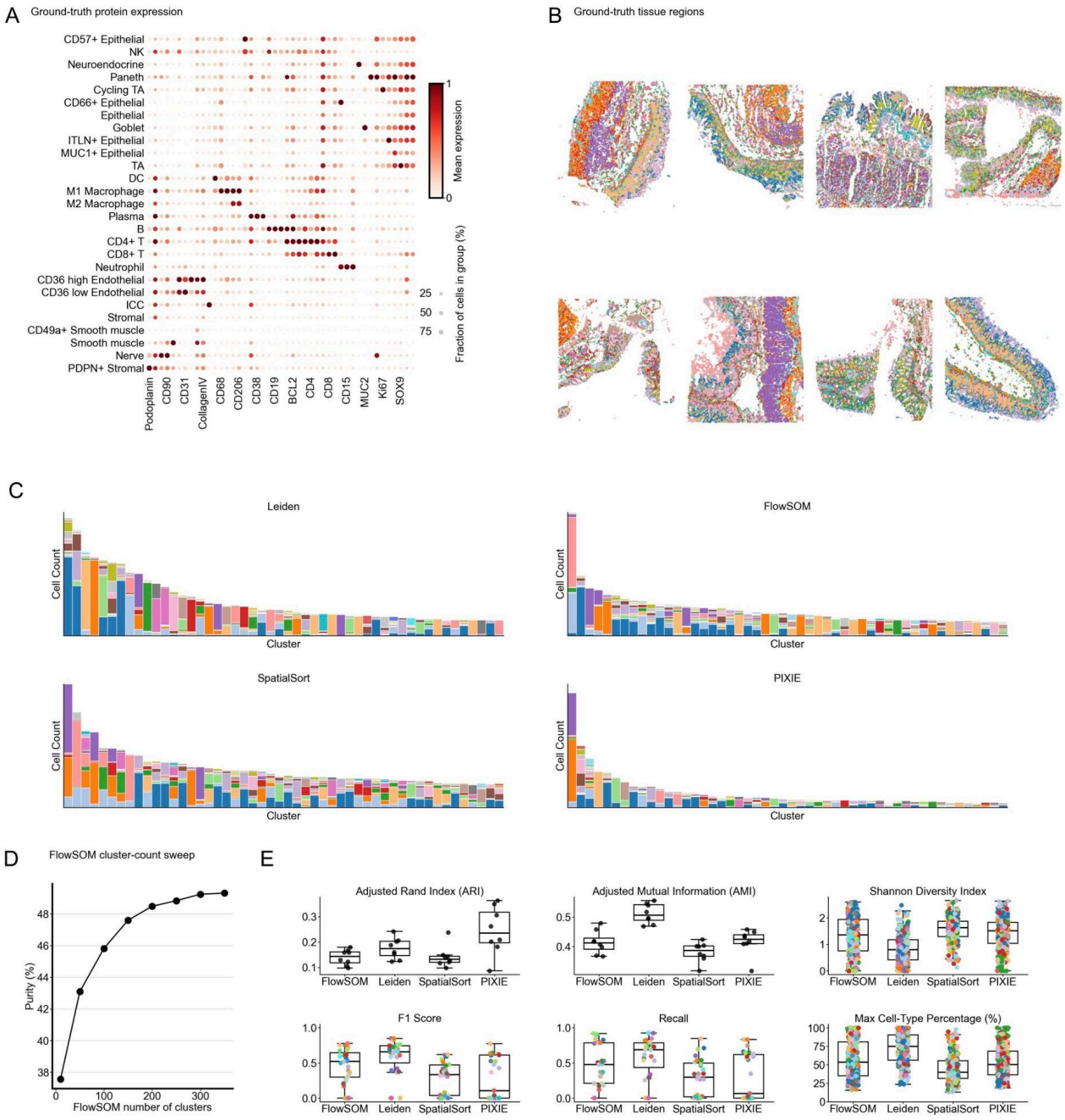


Figure S3. Reference cell-type characteristics and comparative evaluation of clustering methods. (A) Protein-expression profiles across reference cell types. Dot size represents the fraction of cells expressing each marker, and color represents mean marker expression within the cell type. (B) Spatial distribution of reference cell-type annotations across the eight imaged tissue regions. (C) Reference cell-type composition of clusters generated by Leiden, FlowSOM, SpatialSort, and PIXIE. Clusters are ordered by decreasing cell count; colored segments indicate the reference cell types represented within each cluster. (D) Cluster-number sensitivity analysis showing weighted cluster purity as the number of clusters increased. Weighted cluster purity was defined as the fraction of cells matching their cluster's majority reference label. (E) Adjusted

Rand index and adjusted mutual information computed separately for each sample (black dots); per-cluster Shannon diversity, where lower values indicate less heterogeneous clusters; per-cell-type F1 score and recall after majority-vote assignment of each cluster to its most frequent reference label; and per-cluster purity, defined as the percentage of cells assigned to the cluster's most frequent reference label. Colored points denote cell type: cluster-majority label for Shannon diversity and purity, and reference label for F1 and recall. Boxes show the interquartile range, center lines show medians, and whiskers extend to 1.5 times the interquartile range.

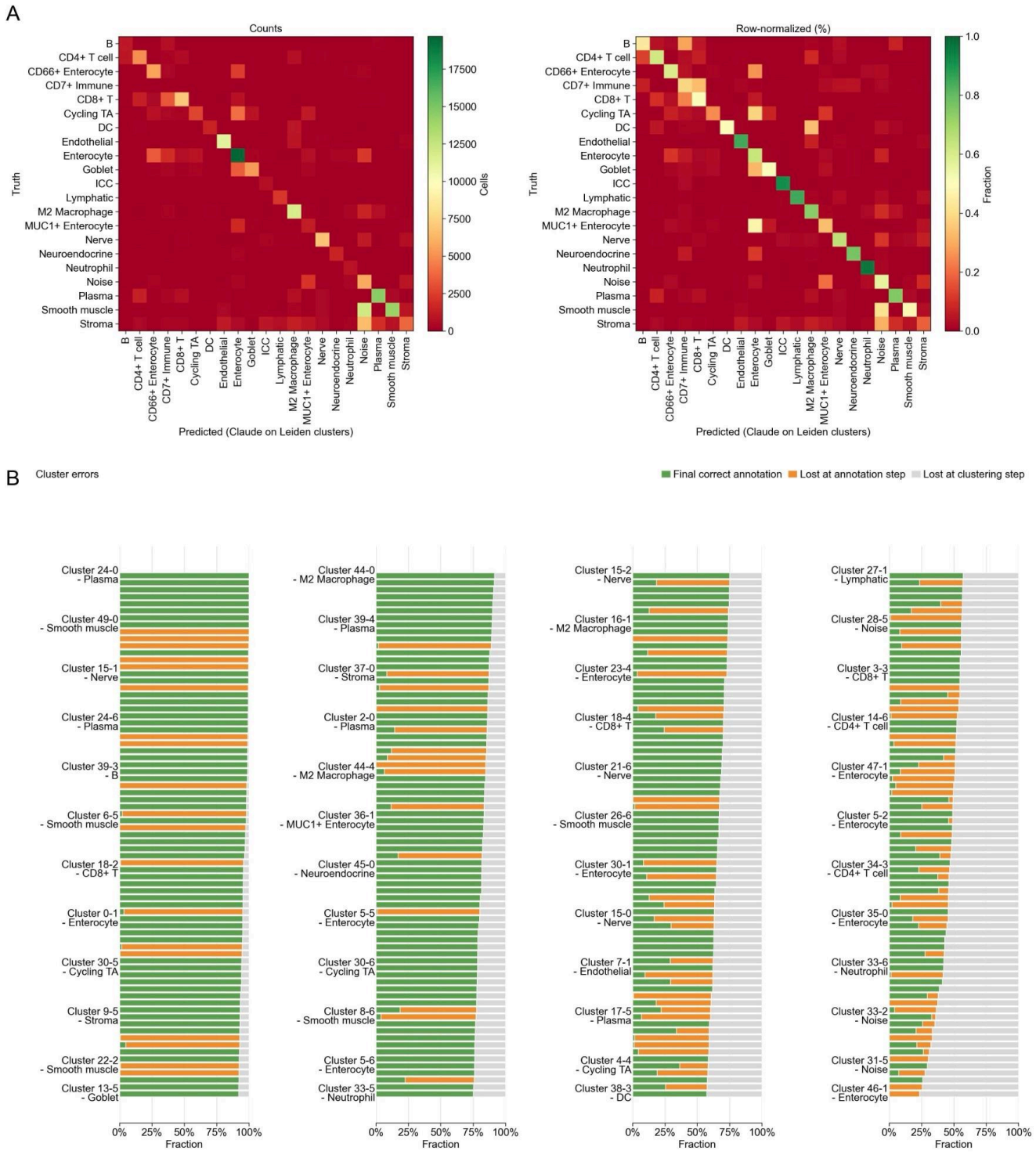


Figure S4. Cell-level confusion and cluster-level error decomposition for Claude Opus 5 annotation of Leiden clusters. (A) Raw cell-count confusion matrix comparing reference cell-type labels (rows) with Claude-predicted labels assigned to Leiden clusters (columns). Color represents the number of cells associated with each reference–prediction pair. (B) Accuracy decomposition for each of the 300 Leiden clusters, ordered by decreasing clustering upper

bound. Cluster labels indicate the cluster identifier and Claude-assigned cell type. Green represents observed correctness, defined as the fraction of cluster cells whose reference label matched the Claude-assigned label; orange represents incorrect annotation after clustering relative to the cluster's majority-label upper bound; and gray represents incorrect clustering attributable to within-cluster cell-type heterogeneity.

Clustering method	Clusters	Parameters
FlowSOM	300	MiniSom=32x32; meta-clusters=300; sigma=1.0; learning_rate=0.3; iterations=10000; seed=42
Leiden	300	K320 local-Leiden hierarchy; Ward-style coarsening to K300; 47 markers; arcsinh(marker / 5); FOV-wise StandardScaler; seed=42
SpatialSort	300	configured K=300; occupied clusters=246; n_neighbors=24; precision_scale=0.65; trace iterations=8; DMH iterations=1; seed=42
PIXIE	300	TIFF pixel MiniSom=10x10; pixel Ward meta-clusters=20; cell MiniSom=24x24; cell meta-clusters=300; sigma=2.0; learning_rate=0.3; iterations=5000; seed=42

Table S1. Final clustering configurations and hyperparameters used in the comparative benchmark. The table reports the number of clusters and selected hyperparameters for FlowSOM, Leiden, SpatialSort, and the PIXIE cell-level adaptation, including map dimensions, metacluster counts, neighborhood and dimensionality-reduction settings, spatial-model parameters, iteration counts, and random-seed handling. MDI, mean decrease in impurity; PCA, principal component analysis; SOM, self-organizing map.

A

V3 candidate ranges

Parameter	FlowSOM	Leiden	SpatialSort	PIXIE-style
Expression threshold	0.25-0.39	0.30-0.44	0.35-0.49	0.10-0.24
Fraction threshold	0.25-0.39	0.50-0.64	0.35-0.49	0.50-0.64
Mean-expression threshold	0.10-0.20	0.02-0.10	0.02-0.10	0.02-0.10
Maximum markers	2-4	3-5	5-7	3-5
Fallback minimum markers	2	2	3	4

B

V3 candidate-grid cardinality

Method	Expression values	Fraction values	Mean-expression values	Maximum markers	Fallback minimum	Combinations
FlowSOM	8	8	6	3	1	1,152
Leiden	8	8	5	3	1	960
SpatialSort	8	8	5	3	1	960
PIXIE-style	8	8	5	3	1	960

C

V3 fine-grid search space

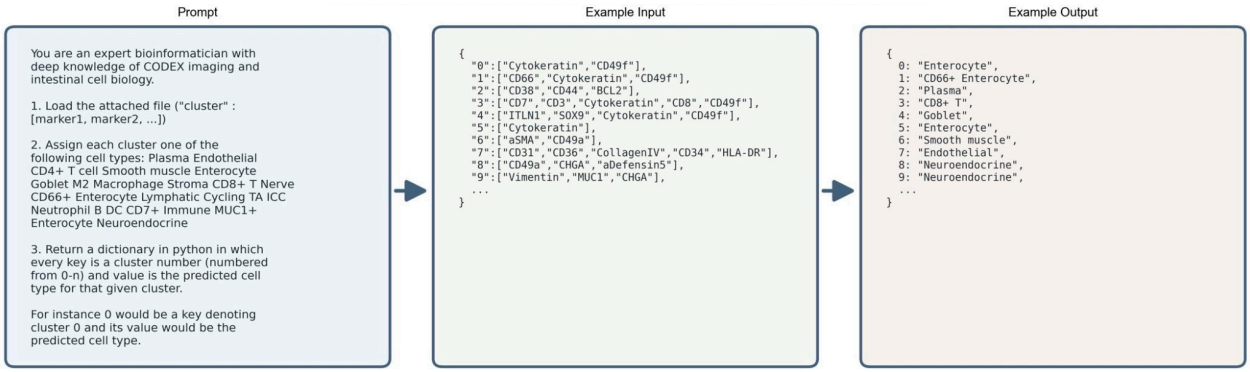
Method	Expression threshold	Fraction threshold	Mean-expression threshold	Maximum markers	Fallback minimum markers	Parameter combinations
FlowSOM	0.25-0.39 (increment 0.02)	0.25-0.39 (increment 0.02)	0.10-0.20 (increment 0.02)	2-4	2	1,152
Leiden	0.30-0.44 (increment 0.02)	0.50-0.64 (increment 0.02)	0.02-0.10 (increment 0.02)	3-5	2	960
SpatialSort	0.35-0.49 (increment 0.02)	0.35-0.49 (increment 0.02)	0.02-0.10 (increment 0.02)	5-7	3	960
PIXIE-style	0.10-0.24 (increment 0.02)	0.50-0.64 (increment 0.02)	0.02-0.10 (increment 0.02)	3-5	4	960

D

Selected V3 parameters used for benchmark

Method	Expression threshold	Fraction threshold	Mean-expression threshold	Maximum markers	Fallback minimum markers	Parameter combinations
FlowSOM	0.31	0.29	0.14	3	2	1,152
Leiden	0.30	0.58	0.02	4	2	960
SpatialSort	0.39	0.41	0.02	6	3	960
PIXIE-style	0.10	0.56	0.02	4	4	960

Table S2. Retrospective selection of cluster marker-summary parameters. (A) Candidate ranges by clustering method. (B) Candidate-grid cardinalities. (C) Fine-grid search spaces. (D) Selected parameters used for the benchmark. Parameters were selected retrospectively using the same full reference-labeled cohort used for evaluation and should not be interpreted as independently validated settings.



Supplementary Note 1. LLM prompt templates and representative inputs and outputs for cell-type annotation. Exact prompt templates are provided for GPT-5.6-sol, Claude Opus 5, Gemini 3.1 Pro Preview, and DeepSeek V4 Pro. Inputs consisted of cluster identifiers, clustering methods, optimized lists of defining protein markers, a fixed 20-class cell-type vocabulary, and immunology and intestinal-epithelium task context. Each model was instructed to assign exactly one permitted cell-type label to each cluster and return the assignments as a JSON object. Representative model inputs and corresponding outputs are included.