Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Efficient reference composition sensitivity analysis for single-cell annotation

Xiaowei Chu

xiaoweichu873@gmail.com

North China University of Science and Technology https://orcid.org/0009-0007-4333-8334

Research Article

Keywords: single-cell RNA sequencing, reference composition, cell-type annotation, optimal transport, shared curvature, implicit differentiation, sensitivity analysis

Posted Date: November 6th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11637549/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Efficient reference composition sensitivity analysis for single-cell annotation

Xiaowei Chu[1,2], Caixia Lyu[2*]
[1] School of Clinical Medicine, North China University of Science and Technology, Tangshan, China
[2] Department of Geriatric Cardiology, Hebei General Hospital, Shijiazhuang, China
* Corresponding author: Caixia Lyu
Email: clyu@ic.ac.uk; lvcaixia@126.com

Abstract

Background: Single-cell labels can depend on the cell-type proportions represented in a reference atlas. Evaluating this dependence requires repeated model fits and informative composition scenarios. We developed an optimal transport workflow that accelerates these checks and directs them toward sensitive assignments.

Findings: A shared baseline curvature factor supports repeated transport correction, while implicit composition derivatives guide perturbations and prioritize cells for review. With references of up to 4,800 cells, quadratic prediction with shared curvature was faster than batched Anderson-accelerated Sinkhorn in 10 of 11 complete comparisons, with a median speedup of 2.55 at the prescribed numerical precision. In 23,658 held-out cells from 40 liver-cancer and lung donors, directed search increased donor-mean sensitive-cell detection from 7.39% to 11.65% in liver myeloid cells, 4.81% to 6.75% in lung epithelium and 4.21% to 4.78% in lung mesenchyme at the same scenario budget. A 16-configuration analysis identified stronger search gains for native multiclass annotations than for merged or binary labels. Across eight native-label configurations, sensitivity ranking recovered 70.63 randomly confirmed sensitive cells per 100 reviewed, versus 58.25 using the winning-score margin and 50.25 using confidence. Selective-annotation improvement was modest and confined to lung epithelium.

Conclusions: Shared curvature reduces repeated composition-checking costs, and derivative-guided search improves detection yield. Composition sensitivity also adds review-priority information beyond the baseline winning-score margin. These diagnostics help identify reference-dependent assignments for biological review.

Keywords: single-cell RNA sequencing; reference composition; cell-type annotation; optimal transport; shared curvature; implicit differentiation; sensitivity analysis

Background

Reference-based single-cell annotation supplies the cell identities used to compare populations, transcriptional states and disease-associated programs. Reference atlases differ in their sampling of cell types, donors and biological conditions. These differences can influence transferred labels even when a query cell has a high annotation score. Reference-composition bias and its consequences for compartment-level biological signals have been demonstrated in xenotransplantation atlases [1]. Multimodal measurements provide complementary evidence for reviewing cell identity [2]. A useful analysis should therefore identify which assignments depend on the assumed reference proportions and make that dependence visible at the cell level.

In balanced optimal transport (OT), reference weights constrain the mass assigned to reference cells and their annotated classes [3,4]. Recent work separates the effect of these marginal capacities from expression geometry and shows that composition-matched marginals can improve label transfer [3]. Here we address a complementary computational problem: with cells and expression distances fixed, how efficiently can alternative reference compositions be evaluated, and which cells should receive priority when reviewing the resulting sensitivity?

In practice, this task arises after an analyst has selected a reference and wants to assess fine labels under alternative class weights. We use lung epithelial annotations to evaluate sensitive-cell detection and selective retention, and liver-cancer myeloid annotations to examine composition sensitivity alongside difficult cross-patient fine-cluster transfer. These applications make the cell-level diagnostic useful before downstream biological comparisons.

Two costs arise in this analysis. First, each candidate composition requires a transport solve. Second, an uninformative set of candidate compositions may miss assignments that change along other directions. Improved initialization [5], implicit Sinkhorn differentiation [6,7,8] and Newton-type optimization [9,10,11] provide mathematical tools for the first cost. Local boundary linearization, as used in adversarial robustness methods such as DeepFool [12], motivates efficient proposals for the second. We apply these tools to reference class weights rather than to a cell's expression measurements.

Our workflow shares a baseline curvature factor across corrected transport solutions and uses the associated implicit derivatives for both directed stress tests and cell prioritization (Figure 1). We evaluate repeated-solve time, sensitive-cell detection at matched scenario budgets, and the yield of a fixed review list as separate outcomes. Four public studies provide immune-cell benchmarks and donor-separated liver and lung analyses. Experiments with reference size, annotation granularity and perturbation radius characterize the settings in which the workflow is useful. Calibrated selective annotation and missing-class challenges assess how these sensitivity readouts relate to conventional confidence.

Methods

Public data and donor held out annotation tasks

We analyzed the Kang immune-cell dataset (GSE96583) [13] and the Hao CITE-seq dataset (GSE164378) [2]. The processed Kang matrix contained 24,673 cells from eight donors and eight author-provided cell types. Control cells from one alternating half of the sorted donor list supplied reference cells, and interferon-β-stimulated cells from the other half supplied queries. Exchanging the donor halves defined the second task. These tasks include both donor and condition differences.

For Hao, donors P1–P4 were selected by identifier. After removing the coarse "other" category and doublet labels, these donors supplied 66,055 cells across seven retained coarse classes, including other T cells. P1 and P2 were each held out as queries, with the other three donors providing the reference pool. The public data also supplied 228 protein channels. Protein measurements were excluded from the RNA model and subsequently used to describe composition-sensitive assignments.

Each of the four tasks contained 3,000 query cells and an initial reference of 1,200 cells. Reference sampling reserved up to five available cells per class; queries were sampled without class balancing. The four query sets contained 12,000 distinct cells from two studies. Class labels supplied by the original authors were used for reference annotation and descriptive assessment of query agreement.

Expression geometry and reference composition scenarios

Counts were normalized to 10,000 per cell and log-transformed. Using reference donors only, we selected the 3,000 highest-variance genes detected in at least ten reference cells. Reference standard deviations were floored at 0.1, scaled values were clipped to ±10, and a 30-component principal-component analysis (PCA) was fitted to the sampled reference. Query cells were transformed with the fitted quantities. Squared Euclidean costs were divided by their median. Entropic regularization strengths were $\varepsilon = 0.05$ and 0.1.

Reference-donor class fractions were estimated with a pseudocount of 0.5 per class. Dirichlet(1,…,1) weights generated 256 mixtures of these donor fractions. Local scenarios moved 25% of the way from the baseline class composition to each donor mixture; full-mixture scenarios used the donor mixtures directly. Query labels did not

determine these scenarios. The initial screen evaluated 16, 64 and 256 local scenarios and 256 full-mixture scenarios. Biological summaries used the completed 256-scenario analyses. These scenarios define the range being examined rather than a sampling-based confidence interval.

Transport annotation and cell level stability measures

Let C be the fixed n × m query-to-reference cost matrix, a and b positive query and reference marginals, and Y the m × q matrix of one-hot reference class labels. Both marginals sum to one. Balanced entropic OT minimizes the transport cost plus an entropy penalty under the specified marginals [4]. The coupling G and class-probability matrix P have the forms

$$G = diag(u)\, K\, diag(v), \quad K_{ij} = exp(-C_{ij}/\varepsilon), \quad P = diag(a)^{-1}\, GY.$$

For baseline reference weights b_0 and class proportions $p_0 = Y^T b_0$, we define $W_{jk} = b_{0j} Y_{jk}/p_{0k}$ and set $b(p) = Wp$. Thus, a scenario changes the total weight of each class while preserving relative weights within that class. All scenario proportions remain positive. This is weight perturbation of a fixed reference set; the reference-expansion experiment described below changes the cells themselves.

For each query cell, the workflow returns corrected class probabilities for every scenario, the fraction of scenarios whose highest-probability label differs from the baseline label, the maximum absolute probability change and the largest within-class probability range. The “any switch” fraction counts cells that change labels in at least one scenario. Because balanced OT constrains class mass, these summaries locate the cells and assignments through which a change in reference composition is expressed.

Reusing baseline curvature across transport solves

The method builds on implicit differentiation of Sinkhorn scaling [6,7,8] and the classical chord, or modified Newton, iteration [9]. At a tightly converged baseline coupling G_0, we construct and factor the target-side Schur matrix

$$S_0 = diag(b_0) - {G_0}^T\, diag(a)^{-1}\, G_0 + b_0\, {b_0}^T.$$

The rank-one term fixes the constant-potential ambiguity. The factorization is shared across scenarios. Optional first- and second-order implicit derivatives provide Taylor initializers in the low-dimensional class-composition displacement. Only the distinct mixed second-order terms are stored; the implementation avoids a tensor indexed by every cell pair and every pair of composition directions.

For the reference log-scaling g, set $v = \exp(g)$, $u(g) = a/(Kv)$ and $h(g) = v \odot K^T u(g)$, with elementwise division and multiplication. The query marginal is normalized exactly at each evaluation. Each chord step solves

$$S_0\, \Delta g = b(p) - h(g).$$

The update is recentered and initially limited to a maximum log-scaling change of four. Up to 12 step halvings seek a sufficient decrease in the L1 marginal residual, using a decrease coefficient of 10^{-4}. Scenarios still unfinished after 40 chord iterations, or failing the step search, pass to a common batched Anderson-accelerated Sinkhorn solver with memory two (AA2). The fallback checks residuals every five iterations and permits up to 10,000 iterations. Each accepted solution satisfies the larger of the row and column marginal L1 errors at a tolerance of 10^{-7}.

The principal variants were baseline reuse with AA2, baseline reuse with shared curvature, and quadratic prediction with shared curvature. The initial screen also included quadratic prediction with AA2 and linear prediction with shared curvature. Class readouts use matrix products without storing a separate dense coupling for every scenario. For n queries and m references, the kernel and curvature factor require $O(nm + m^2)$ storage;

constructing the target factor costs approximately $O(nm^2 + m^3)$. The shared work is amortized across the requested scenarios.

Published solver adaptations and reference expansion

The expanded benchmark added the authors' C++ implementations of the safe and sparse Newton method (SSNS) [10] and sparse-plus-low-rank quasi-Newton method (SPLR) [11], from regot-python commit cac1f20472d6cf7dad482d1446e6b923be3725af. We exposed the supported initial-dual argument through a thin Python wrapper and aligned the final stopping comparison with the same full marginal L1 criterion. The adapters also correct a dimension mismatch in the checked gauge-fixed conjugate-gradient interface and its cached-direction slice. Source patches, validation records and upstream licensing files are retained with the benchmark.

After default sparse-direct-solver pilots exceeded the time budget, an explicit amendment enabled the authors' conjugate-gradient option and the common AA2 fallback. Each comparator was allowed up to 40 native iterations per scenario, nearest-composition ordering and initialization from the nearest previously available or baseline potential. SPLR retained its native low-rank linear algebra. The resulting workflows are denoted SSNS + AA2 and SPLR + AA2. Their exposed interfaces permit potential reuse but do not expose cross-call Hessian factors or quasi-Newton histories. The comparison therefore includes differences in batching and reusable solver state.

For each study's first task, reference sets were expanded to 2,400 and 4,800 real cells sampled without replacement. Each expanded set retained the original 1,200 reference cells. The 3,000 queries, selected genes, fitted normalization and PCA transform, and original distance scale remained fixed. New cells were projected through the existing transform. Uniform baseline cell weights were used at each size, and local scenarios were recentered on that size's class composition. Adding real cells changes both reference geometry and composition, so this experiment evaluates practical reference expansion.

Timing memory and numerical validation

The expanded grid comprised 12 configurations, five workflows and three planned fresh-process repeats per combination, for 180 records. Runs were sequential, used the same CPU affinity and one numerical thread, and followed a deterministic cyclic method order. Audit time included method-specific preparation, prediction, correction, fallback and class readout. Common expression processing, matrix preparation and the baseline fit were excluded from audit time and recorded separately. Baseline fits used a marginal tolerance of 10^{-11}. Each complete audit had a 180-second budget; after an initial timeout, the two remaining planned repeats were marked as not run. Medians and finite speed ratios were reported only for complete three-repeat comparisons. This benchmark was exploratory, with the original plan and amendments preserved.

Process resident set size (RSS) was sampled every 10 ms. Memory summaries report the largest observed audit RSS across attempted repeats, including loaded inputs and working arrays. An observation before timeout describes the attempted run. All timings used an Intel Xeon Platinum 8573C CPU with Python 3.12 and no GPU.

Independent POT solutions [14] were computed for the first, middle and final scenario of each expanded-benchmark configuration, to a full marginal L1 error below 10^{-10}. The resulting 36 configuration–scenario reference solutions supported 456 comparisons with completed methods and repeats. Where AA2 completed, every scenario was additionally compared with its output. Validation ran after timing finished. The earlier 32-configuration screen and stricter-tolerance checks are reported separately in the Supplementary information.

Biological interpretation of sensitivity profiles

We summarized switching by author-provided cell type and compared switching frequency with baseline class-probability entropy for ranking disagreements with author labels, using the area under the receiver operating characteristic curve (AUROC). For Hao T cells, the protein contrast was the mean log(1 + count) across CD4-1/CD4-2 channels minus the corresponding mean across CD8/CD8a channels. Its distributions were compared descriptively between stable and switching cells within author-annotated CD4 and CD8 classes. Proteins were withheld from RNA fitting, although the original author annotations used multimodal information.

Donor-separated human tissue validation

We extended evaluation to liver myeloid cells from GSE149614 [15] and lung epithelial and mesenchymal cells from GSE135893 [16]. The extension included 10 liver-cancer patients and 30 lung donors; the two lung compartments share donors and are not independent cohorts. Previously archived basic quality-control and per-capture scDblFinder results were reused (Supplementary methods). Only non-tumor liver and primary tumor samples were included in the liver analysis. Five donor folds assigned three folds to reference construction, one to calibration and one to testing in each rotation. Lung folds were stratified by control versus pulmonary-fibrosis status. Paired liver tissues remained together. Every donor served as a test donor once; no donor occupied multiple roles within a rotation.

For each reference split, we sampled 384 reference cells, reserving at least four per available class and balancing donor contributions within class. Each query sample contributed at most 500 uniformly sampled cells. Reference-only normalization statistics, 2,000 high-variance genes and 30 PCA components defined median-scaled cosine costs, with $\varepsilon = 0.05$. Baseline class proportions were estimated without query labels from 15 inverse-distance-weighted nearest reference neighbors, adjusted for the reference sampling fractions and donor-averaged class frequencies, with a 0.5 class pseudocount. This tissue configuration differs from the original immune-cell benchmark and is not assigned its speedup ratio. Author labels were evaluation targets for disagreement, not experimentally established cell-level truth.

Derivative-guided composition stress testing

We parameterized positive class proportions as $p(z) = \mathrm{softmax}(\log p_0 + z)$, with centered log weights z. The implicit class-probability derivative was transformed by the chain rule using $\mathrm{diag}(p_0) - p_0 p_0^T$. For each query cell, we computed its baseline winning-class score minus each competing-class score and divided this gap by the norm of its log-weight gradient. The competitor with the smallest ratio defined a unit direction opposite that gradient. This linearized construction proposes a nearby label boundary; it does not certify a globally minimal change. The score-gap-to-gradient-norm construction follows the local boundary geometry underlying methods such as DeepFool [12]. Here a single baseline linearization defines directions in centered log composition, and each proposed scenario is subsequently solved to the common numerical tolerance. We do not iterate perturbations in expression space.

Each test sample supplied 32 uniformly selected anchor cells, or all cells when fewer were available. Each anchor generated one directed scenario. The matched control used the same number of isotropic random directions projected onto the centered log-weight subspace and normalized to unit length. The primary radius was 0.10; radius 0.20 was a secondary analysis. These are log-weight L2 radii, not per-class percentage-change bounds. Every proposal was solved to marginal L1 tolerance 10^{-7} using the same shared-curvature workflow and fallback. We separately counted intended-anchor changes and the fraction of all query cells changing labels in at least one scenario. The original four immune-cell tasks were also tested at these radii using unchanged embeddings, 1,200 references, 3,000 queries and $\varepsilon = 0.05$. This is a matched-scenario-budget comparison; direction-selection overhead was not converted into a new wall-time speed ratio.

Operating range and cell review prioritization

We conducted an exploratory applicability analysis using 1,000 uniformly sampled queries from each original immune-cell task, keeping these queries fixed across configurations. Each task used four configurations: 300 references with native labels, 1,200 references with native labels, 1,200 references with merged T-cell and monocyte labels, and 1,200 references with T versus non-T labels. The 300-reference subset reserved at least two cells per native class where available and sampled the remaining cells without replacement. The original RNA embedding and full-task distance scale were retained. Merging labels changes the composition coordinates being perturbed; the configurations therefore assess different annotation questions.

For each of the 16 configurations, three random seeds generated 32 uniformly sampled anchors and 32 centered isotropic random directions. Directed and random scenarios used log-weight L2 radii 0.03, 0.10 and 0.30. Nested prefixes of 8, 16 and 32 scenarios assessed the effect of the search budget. All scenarios used $\varepsilon = 0.05$, first-order prediction, shared-curvature correction and a marginal tolerance of 10^{-7}. Repeated directions were retained. There were 9,216 solved perturbation scenarios; nested prefixes were not counted as additional solves. Timing in this extension was diagnostic and was not added to the original speed benchmark.

For review prioritization, we ranked cells by the smallest linearized composition radius, by one minus maximum class probability, by normalized entropy, and by the negative winning-score margin (the largest class probability minus the second largest). We evaluated the first 100 of 1,000 cells, with stable sorting in the archived query order to resolve ties. The evaluation set comprised cells changing labels under the union of 96 random scenarios across three seeds at each radius, generated independently of all ranking scores. The margin-only comparator was added during manuscript review using unchanged probabilities, scenarios, evaluation sets and review budgets; no scores were fitted or retuned. It tests the incremental utility of the derivative-informed ranking over a conventional small-gap rule. A random-plus-directed union was retained descriptively. Neither sampled union is exhaustive. Reference sizes reuse the same queries, so the eight native-label configurations are not eight independent cohorts.

One large-radius directed scenario per configuration was checked against an independent POT solution. Failures and partial outputs were retained under a 240-second configuration budget. A precheck corrected the mapping of Kang T-cell label names before the merged and binary analyses were rerun; earlier mapping outputs were isolated. The applicability plan preceded these runs but followed the initial positive findings, and is treated as exploratory.

Selective annotation and missing-reference-class checks

For each tissue query, the baseline and 16 random local scenarios supplied sensitivity features: eight centered log-weight draws each at standard deviations 0.05 and 0.10. A confidence-only logistic model used one minus maximum class probability, entropy and negative winning margin. The augmented model additionally used switching fraction, maximum probability shift, negative worst-scenario winning margin and negative log linearized radius. Standardization and L2-regularized logistic regression ($C = 1$) were fitted on calibration donors only, with equal total weight per donor. The primary comparison retained the lowest-risk 80% of each test sample; coverage curves from 50% to 100%, uncalibrated scores and geometry-augmented variants were also retained. The classifiers ranked assignments for review and did not relabel cells.

In a separate challenge, one class was removed from the reference. Selection used reference donors only: the most abundant class represented in at least three reference donors and by at least ten sampled reference cells. Expression features, projections and cost scaling were retained, while baseline composition was re-estimated. Withheld-label detection was compared with ordinary confidence and mean distance to the five nearest references. Calibration models were not trained on the synthetic missing-class targets. Other naturally unsupported labels were excluded from this controlled comparison. Rejection and uncertainty-aware annotation

have established precedents [17,18]; the question here was whether composition-specific features add information to the same fitted OT classifier.

We averaged paired tissue results within each liver patient, then computed donor-mean differences within each compartment. Percentile intervals used 2,000 paired donor-bootstrap draws. They are exploratory, conditional on the fitted folds and models, without multiplicity correction or re-fitting inside each bootstrap. The two lung compartments were not pooled as independent cohorts. The frozen analysis plan, all alternative scores and incomplete or opposite outcomes were retained. Central finite differences and independent POT solves checked the extension numerically. We additionally summarized the existing coverage curves by disease condition without refitting. These exploratory subgroup intervals used 5,000 donor-bootstrap draws; groups with fewer than five donors received descriptive summaries only.

Diagnostic assessment of tissue labels and perturbation sizes

We examined the frozen liver test outputs by tissue condition, reporting agreement with the original C-numbered myeloid clusters for both OT and the existing 15-neighbor baseline. The latter used the same expression geometry and reference-frequency adjustment as the marginal initialization; it was not retuned as a new annotation competitor. Reference support was counted in the corresponding training pool and sampled reference. Confusion counts and donor concentration of each cluster located difficult transfers. These post-review diagnostics used author labels only for evaluation and did not modify fitted models, selected cells or the primary outcomes.

We translated the log-weight perturbations into changes in actual class proportions, reporting total variation, one half of the summed absolute class-proportion changes, and the largest absolute change in one class. For context, each reference donor's QC-retained author-label proportions were compared with the donor-mean proportions in its reference fold. Repeated fold appearances were treated descriptively. Equal log-weight L2 radii need not produce equal total variation; the directed-versus-random comparison is matched in log-weight space. Observed donor variation contextualizes the scale without defining a calibrated uncertainty distribution.

Results

Reference composition sensitivity varies across immune cell populations

Across completed tasks and regularization settings, 4.3–8.5% of query cells changed their highest-probability label in at least one of 256 local composition scenarios. Under full donor mixtures, the corresponding range was 16.1–33.5%. These results show that the choice of reference class weights affects a subset of assignments even when the same cells and RNA geometry are retained.

Sensitivity was uneven across cell types and donors (Figure 2). In Hao P1 at $\varepsilon = 0.05$, 116 of 525 author-annotated CD8 T cells (22.1%) switched labels under local scenarios, compared with 31 of 1,114 CD4 T cells (2.8%). In Hao P2, the corresponding values were 52 of 606 CD8 T cells (8.6%) and 34 of 990 CD4 T cells (3.4%). These profiles identify where composition dependence is concentrated within an analysis, rather than reducing a reference to a single overall stability score. Maximum probability changes and probability ranges retain shifts that do not cross a hard-label boundary.

Protein measurements contextualize composition sensitive annotations

The paired protein data retained the expected group-level separation of CD4 and CD8 T cells in both Hao query donors (Figure 3). In P1, median CD4-minus-CD8 protein scores were 2.292 and 2.045 in stable and switching CD4 T cells, respectively, and −3.370 and −3.325 in the corresponding CD8 T-cell groups. In P2, the medians were 2.338 and 2.313 for CD4 T cells and −3.355 and −3.412 for CD8 T cells. Thus, composition-sensitive

RNA assignments coexisted with recognizable class-associated protein patterns. These observations illustrate how the sensitivity output can guide a review using an additional assay.

Composition sensitivity and baseline uncertainty had different behavior. Baseline agreement with author labels ranged from 72.1% to 91.7%. Entropy ranked annotation disagreements better than switching frequency in 13 of 15 completed task-by-regularization-by-scenario-family settings. The exceptions were full-mixture analyses in Kang fold 2, with switching AUROC values of 0.932 and 0.940 versus entropy values of 0.854 and 0.822 at ε = 0.05 and 0.1. These two settings used the same query cells. The profiles therefore provide information specifically about dependence on reference weights and are best interpreted alongside ordinary uncertainty measures.

Shared curvature accelerates repeated sensitivity analysis

The expanded benchmark accounted for every planned record: 152 runs completed, 10 timed out and 18 were not repeated after an earlier timeout. Quadratic prediction with shared curvature completed all 36 of its planned runs. Across the 11 configurations with complete AA2 comparisons, it was faster in 10, with a median AA2-to-method time ratio of 2.55 and a range of 0.60–8.10 (Figure 4A; Table S1).

For 64 local scenarios at ε = 0.05 and 1,200 reference cells, the Kang fold-1 median audit time decreased from 52.611 s with AA2 to 7.770 s with quadratic prediction and shared curvature, a 6.77-fold speedup. Hao P1 decreased from 5.254 to 1.014 s, a 5.18-fold speedup. The corresponding shared-curvature variants without prediction took 9.392 and 0.961 s. Across all 12 paired configurations, adding quadratic prediction improved on shared curvature alone in eight, with a median time ratio of 1.10 (Figure 4B). Curvature reuse supplied the main computational benefit; the incremental value of prediction depended on the workload.

The small, easier Kang fold-1 configuration at ε = 0.1 and 16 scenarios favored AA2: its median was 0.889 s, compared with 1.479 s for quadratic prediction with shared curvature. All observed repeats, including slower ones, were retained. This configuration identifies a practical setting in which additional preparation is not sufficiently amortized.

Reference expansion defines the computational operating range

With 2,400 references and 64 scenarios, quadratic prediction with shared curvature required 28.696 s on Kang and 2.476 s on Hao, compared with AA2 medians of 73.261 and 7.371 s. At 4,800 references, it completed in 68.577 and 6.432 s, respectively. AA2 completed the Hao task in 15.984 s; its first Kang attempt exceeded the 180-second budget, so that comparison has no finite speed ratio (Figure 5; Table 1).

The SSNS + AA2 workflow completed all three repeats in eight configurations, and SPLR + AA2 did so in seven. The quadratic shared-curvature workflow was faster in every complete comparison with either adaptation. At 1,200 references and 64 scenarios, SSNS + AA2 took 143.830 s on Kang and 117.601 s on Hao. SPLR + AA2 completed two of three Kang repeats and had a Hao median of 156.551 s. At the two expanded reference sizes, both adaptations timed out on their first attempts and the remaining repeats were not run. These measurements describe the implemented repeated-audit workflows, including their initialization, native correction, fallback and readout costs.

The observed peak audit RSS for quadratic prediction with shared curvature increased from 242.7 MiB on Kang and 238.7 MiB on Hao at 1,200 references to approximately 825.7 MiB in both 4,800-reference tasks. Common baseline fits at the largest size took 39.107 s on Kang and 3.716 s on Hao, measured once separately from the repeated audits. The reference-expansion results therefore quantify both the time benefit after baseline fitting and the memory cost of dense curvature reuse.

Corrected outputs retain the prescribed numerical precision

Every completed expanded-benchmark run met the common 10^{-7} marginal criterion. Across 456 comparisons with the 36 tighter-tolerance POT reference solutions, the maximum absolute class-probability difference was 7.19538×10^{-6}, with no label disagreements in the checked scenarios. Across all scenarios that could be paired with AA2, the largest probability difference was 1.41306×10^{-5} and the label-disagreement fraction was zero. No nonfinite native outputs were recorded among completed final comparator runs. These checks support using the accelerated solutions to obtain the same annotation readouts at the evaluated precision.

Directed scenarios increase the yield of composition stress tests

The tissue extension completed 156 sample analyses, comprising 78 calibration and 78 test analyses. Test samples contained 8,031 liver myeloid, 11,701 lung epithelial and 3,926 lung mesenchymal cells. At radius 0.10, directed search increased donor-mean all-query switching from 7.39% to 11.65% in liver myeloid cells, from 4.81% to 6.75% in lung epithelium and from 4.21% to 4.78% in lung mesenchyme (Figure 6A; Table S4). Paired differences were 4.26, 1.93 and 0.58 percentage points, respectively, with exploratory 95% intervals of 3.39–4.94, 1.35–2.53 and 0.22–0.97. Improvement occurred in 10 of 10, 26 of 30 and 13 of 30 donors; one donor in each lung compartment favored random directions and the others were tied. The larger radius retained the same aggregate direction of benefit.

The four original immune-cell tasks also favored directed scenarios (Figure 6B). At radius 0.10, the numbers of sensitive query cells were 151 versus 118 in Kang fold 1, 200 versus 155 in Kang fold 2, 323 versus 292 in Hao P1 and 268 versus 224 in Hao P2. Across the 12,000 queries, directed search identified 942 cells versus 789 at the same 32-scenario budget, an increase of 19.4%. At radius 0.20, the corresponding totals were 1,910 and 1,570. These are additional confirmed model-label changes, not additional experimentally verified annotation errors.

Search gains depend on annotation granularity and scenario budget

All 16 applicability configurations completed, comprising 9,216 corrected perturbation scenarios (Figure 7A; Table 3; Table S7). With native labels, radius 0.10 and 32 scenarios, the mean number of sensitive cells detected per 1,000 queries increased from 72.92 to 91.33 with 300 references and from 70.75 to 84.33 with 1,200 references. These are relative increases of 25.3% and 19.2%, respectively, averaged over four tasks and three seeds. At the same radius, directed search with eight scenarios yielded 75.58 and 65.58 cells for the two reference sizes, compared with 53.75 and 52.75 for random search. This establishes a detection benefit at finite scenario budgets without assigning a wall-time speedup to the proposal step.

Coarser annotations reduced the incremental search benefit. With 1,200 references and merged labels, 32 directed scenarios found a mean of 54.67 sensitive cells versus 52.67 using random directions at radius 0.10. In the T versus non-T setting, both strategies found 40.50 cells on average; every evaluated seed and budget yielded the same hit count. Centered two-class log weights have only one direction axis, which random proposals can cover with either sign. Larger radii increased the number of detected changes, but changing taxonomy also changes the perturbation problem, so absolute sensitivity counts are not measures of comparative annotation accuracy.

Composition derivatives prioritize cells for stability review

The derivative-informed ranking increased the yield of a fixed review list against the random-only evaluation scenarios (Figure 7B; Table 4). Across eight native-label task-by-reference configurations at radius 0.10, the first 100 cells contained a mean of 70.625 confirmed sensitive cells when ranked by linearized radius, versus 58.250 using the winning-score margin, 50.250 using confidence and 35.875 using entropy. The radius ranking exceeded the margin comparator in all eight configurations, by 4–20 cells per list. Its mean increment was 12.375 cells per 100 reviewed, a 21.2% relative increase over margin ranking; the increment over confidence

was 20.375 cells, or 40.5%. This isolates an advantage of the derivative-informed rule beyond a small baseline score gap. Native-label mean yields also favored radius ranking at radii 0.03 and 0.30 (Table S9). These outcomes measure model-label changes under sampled compositions, not experimentally established annotation errors.

Every applicability scenario met the common marginal criterion. Across the 16 independent POT checks, the maximum absolute probability difference was 8.21e-06 and no checked labels disagreed. Full scenario counts, alternative radii, review rankings and numerical diagnostics are retained with the applicability results.

Sensitivity provides a compartment specific selective annotation benefit

At matched 80% fine-label coverage, adding sensitivity features to calibrated confidence reduced donor-mean disagreement with author labels in lung epithelium from 5.604% to 5.227%, a difference of −0.376 percentage points (exploratory 95% interval −0.588 to −0.182; Figure S1A). Sixteen donors improved, one worsened and thirteen were unchanged. The corresponding differences were −0.019 percentage points in liver myeloid cells and −0.288 in lung mesenchyme, with intervals spanning zero (Table S5). These results support a modest compartment-specific contribution, rather than a general increase in annotation accuracy. Baseline liver fine-label agreement remained low, approximately 50% averaged across test samples.

In the 12 IPF donors, disagreement at 80% retained coverage decreased from 9.531% to 9.026% (difference −0.505 percentage points; exploratory 95% interval −0.794 to −0.247). Models were unchanged from the held-out analysis. IPF subgroup gains were smaller at 50–60% and 90% coverage (Table S8 and accompanying condition_coverage.csv). Figure S2 presents compartment-level curves across all donors. This exploratory analysis locates the observed benefit rather than validating a new coverage threshold.

Controlled reference-class removal exposed another boundary. For lung epithelium, nearest-reference distance achieved a donor-mean withheld-class detection AUROC of 0.926, compared with 0.553 for switching frequency. No consistent sensitivity advantage over distance emerged across compartments (Figure S1B; Table S6). Sensitivity to composition therefore did not provide a general detector of cells absent from the reference. The study did not test improvement of downstream disease differential-expression or pathway conclusions.

All completed tissue scenarios met the 10^{-7} marginal criterion, including those requiring fallback. Central-difference checks in one held-out sample per compartment gave relative derivative errors between 1.49×10^{-9} and 1.06×10^{-7}. Nine independent POT comparisons had a maximum class-score difference of 1.93×10^{-6} and no label disagreements. These checks distinguish derivative proposal quality from numerical solution error.

Liver diagnostics locate difficult donor held out cluster transfers

Condition-stratified diagnostics showed that low liver agreement was concentrated in tumor samples (Figure S3; Table S10). Mean agreement across donors was 74.17% in non-tumor liver (eight donors, 3,436 cells) and 31.38% in tumors (ten donors, 4,595 cells). The existing nearest-neighbor baseline gave 75.74% and 31.76%, respectively. All 8,031 test cells had an author-label category present in the corresponding reference; each represented query class had at least four sampled reference cells and support from at least three training donors. Thus, absence of a label category did not explain this failure.

The fine labels were 14 author-provided myeloid clusters. In the QC-retained input, 93.3% of C38 cells and 89.8% of C39 cells came from a single donor for each cluster, despite low-level representation in other donors. Frequent tumor disagreements were C39 to C6 (255 cells), C38 to C6 (253), C6 to C41 (227) and C41 to C6 (202). These findings localize the difficulty to cross-donor fine-cluster transfer shared by both annotation baselines. They support using the liver analysis to examine model sensitivity, while providing no evidence that sensitivity scoring repairs the underlying label transfer.

The radius-0.10 directed scenarios produced median total-variation changes of 1.68%, 2.12% and 2.40% in liver myeloid, lung epithelial and lung mesenchymal compositions, respectively (Table S11). The largest single-class

changes across these scenarios were 2.78, 3.04 and 3.19 percentage points. Random directions at the same log-weight radius produced smaller median total-variation changes of 1.00%, 0.90% and 1.45%. Median reference-donor deviations from their fold mean were 39.97%, 35.42% and 29.23%, respectively. The primary stress tests therefore probe local changes relative to the observed between-donor composition differences, with direction-dependent changes in actual proportions.

Discussion

This work connects computational reuse with a concrete single-cell analysis task: identifying annotations that depend on reference class proportions. Shared curvature reduces the cost of related solves, derivative-directed scenarios expose more label changes within a fixed log-weight scenario budget, and the linearized radius concentrates randomly confirmed sensitive cells near the top of a review list. The added margin-only comparison shows that the ranking benefit extends beyond selecting cells with nearly tied leading labels. These outcomes address different stages of the same analysis and are evaluated separately.

Practical use in lung and liver annotation

An analysis starts from a fitted OT annotation, a fixed reference cell set and a stated range of alternative class proportions. Corrected scenario solutions identify which cells switch, their competing labels and the composition changes that elicit them. These outputs can be reviewed alongside the baseline label, confidence, marker genes and independent measurements where available. The Hao protein comparisons illustrate this complementary evidence: cells with changing RNA-model labels can retain recognizable class-associated protein patterns. Composition sensitivity identifies assignments to investigate; the biological identity is assessed using the additional evidence.

For lung epithelium, the practical question is whether the same cell retains its subtype assignment when reference epithelial proportions change. At 32 scenarios and log-weight radius 0.10, directed search increased donor-mean sensitive-cell detection from 4.81% to 6.75%. Adding sensitivity features to calibrated confidence also reduced disagreement with author labels from 5.604% to 5.227% among the 80% of cells retained for fine annotation. These results support identifying reference-dependent assignments and prioritizing their review. The resulting cells and competing labels provide a focused starting set for marker inspection before comparing epithelial subtypes across disease groups; the selective-retention gain is modest and does not constitute automatic label correction.

For hepatocellular carcinoma myeloid cells, the aim is to locate composition-sensitive fine labels while also recognizing clusters that transfer poorly across patients. Directed search increased donor-mean sensitive-cell detection from 7.39% to 11.65% at the same scenario budget. The accompanying confusion and reference-support analyses showed that low tumor agreement persisted for both OT and nearest-neighbor annotation even when every tested label category was represented in the reference. A practical review would therefore examine the competing fine clusters, their marker support and their representation across donors. Where a fine identity remains unsupported, an analyst could retain a broader lineage assignment pending further evidence. The current results localize difficult assignments but do not demonstrate repaired liver label transfer.

Methodological scope

Reference-composition auditing has been developed in xenotransplantation [1], and marginal-capacity correction has been studied for OT label transfer [3]. Uncertainty-aware rejection and conformal annotation address reliable labeling [17,18], while a recent preprint uses OT coupling ambiguity to flag unresolved cells in multi-omics integration [19]. Our evaluated contribution is a reusable computational workflow for composition-specific sensitivity: cross-scenario factorization reuse, derivative-guided scenario selection and a local composition-based review score. The present comparisons establish its numerical behavior and added utility in the tested tasks; they are not head-to-head comparisons with every annotation or rejection framework.

The mathematical components have established foundations. Implicit first- and second-order derivatives, warm starts and chord iteration are described in prior numerical work [5,6,7,8,9,10,11]; score-gap linearization is related to local decision-boundary methods [12]. Other OT acceleration strategies include sparse Newton steps, spectral methods, learned amortization and temperature-path methods [20,21,22,23]. The shared-curvature ablation shows that reusable correction work, rather than quadratic prediction alone, supplies the main time benefit here. Directed search uses the same composition derivatives for a different purpose: spending a finite set of corrected solves on informative perturbations.

The operating-range results guide practical use. Reference expansion to 4,800 cells demonstrates that dense curvature reuse can remain beneficial, with a quadratic storage cost. Small and easy workloads may favor AA2 because extra preparation is not amortized. Directed search adds most in the tested multiclass problems and loses its advantage in the binary direction space. Composition-based ranking outperformed confidence, entropy and winning margin on the native-label sensitivity endpoint. Selective-annotation improvement was restricted to lung epithelium; liver tumor fine-cluster transfer remained difficult for both OT and nearest neighbors. Reference distance was stronger for the lung missing-class challenge. These diagnostics serve complementary annotation-review purposes.

The evidence covers four public studies and is conditional on their expression representations, references and perturbation ranges. Author labels are benchmark targets, and the Hao proteins contributed to the original multimodal annotations despite being excluded from our RNA fitting. Tissue intervals condition on fitted folds, lung compartments share donors, and applicability configurations reuse queries. The post-review margin ablation and liver diagnostics reuse frozen outputs rather than adding independent cohorts. Log-weight-matched search can produce different changes in actual proportions across directions, as explicitly quantified. A sampled set of label changes does not establish an exhaustive robustness radius. Independent biological labels and downstream disease endpoints remain necessary to determine whether targeted review improves biological conclusions.

Prospective applications

A concrete future application is heart failure with preserved ejection fraction (HFpEF). Human HFpEF single-nucleus RNA-seq has reported reduced macrophage clearance programs [24]. A composition audit could ask whether this association persists when assignments among neighboring myeloid populations change. Each corrected annotation would feed patient-level cell-abundance estimates and donor-level pseudobulk expression analyses, keeping the patient set, covariates and prespecified gene sets fixed. Comparing disease-effect directions and magnitudes across scenarios would reveal which signals depend on the annotation and which merit orthogonal validation. The composition range should be specified independently of the desired disease result. This HFpEF application was not evaluated in the present study.

Other prospective uses include prioritizing cells for protein or spatial validation and embedding composition checks in automated single-cell annotation workflows. Such a module could return sensitivity flags, alternative labels and review priorities whenever an OT annotation is used to support downstream analysis. Repeated checks can share the existing factorization while the cells, cost matrix and regularization remain fixed; changing the reference cells or expression representation requires rebuilding the numerical structures. These extensions would need evaluation against their own biological and computational endpoints.

Conclusions

Shared curvature and implicit composition derivatives support efficient sensitivity analysis of OT-based single-cell annotation. Repeated solves become faster, finite-budget stress tests detect more reference-dependent assignments, and derivative-informed ranking adds sensitivity-detection value beyond the winning-score margin. The observed multiclass gains and tissue-specific results define a practical scope for using these diagnostics alongside conventional annotation evidence.

Availability of data and software

The source studies are available under GSE96583 and GSE164378. The accompanying code-and-results archive contains processed-input provenance and checksums, selected cell identifiers and embeddings, scenario definitions, the Python implementation, comparator sources and patches, raw timing and memory records, numerical validation tables and figure scripts. The principal interface is CompositionAudit(cost, labels, epsilon).run(proportions); order=2 selects quadratic prediction, while the default order=0 uses shared curvature without prediction. Reproduction instructions for the expanded benchmark are provided in FOLLOWUP_REPRODUCIBILITY.md. The tissue extension uses GSE149614 and GSE135893. Its archive includes protocol.json, protocol_original_extension.json, donor partitions, selected cell and feature identifiers, per-scenario probabilities, complete score comparisons, numerical checks and scripts 02_validation.py through 06_integrate_manuscript.py. README.md documents restoration of the archived annotated count checkpoints and reproduction of the added analyses. The applicability extension adds a fixed 16-configuration protocol, all 9,216 scenario outputs, input checksums, random-only review evaluations, condition-stratified coverage tables and reproduction scripts. The revised figures and manuscript-building source accompany these results. The review-completion package adds the margin-only ablation at every evaluated radius and taxonomy, frozen test-probability inputs for liver diagnostics, confusion and reference-support tables, realized proportion changes, a standalone installation file, a deterministic numerical example with expected outputs, and manuscript/figure regeneration scripts. Its README gives the file layout and exact commands.

The tested environment used Python 3.12, NumPy 2.3.5, SciPy 1.17.0, POT 0.9.6.post1, scikit-learn 1.8.0, pandas 2.2.3, anndata 0.11.4 and h5py 3.16.0. The comparator extension was built for x86-64 Linux; other systems require rebuilding. Public repository and archive DOI: [authors to insert before journal submission]. The review package contains a proposed MIT license for author confirmation; third-party code retains its original licensing. Source code and an executable example are supplied for review; no public deposition is claimed. The new tissue validation used an isolated Python 3.12 environment with anndata 0.12.2 and h5py 3.14.0; NumPy, SciPy, POT, scikit-learn and pandas matched the versions listed above. These versions are archived separately from the original timing environment.

Declarations

Ethics approval and consent: This study reanalyzed existing public processed data and involved no new recruitment, specimen collection or animal experiments. The source publications describe the original studies' ethics and consent procedures.

Competing interests: [Authors to confirm.]

AI assistance: Generative AI assisted literature review, implementation, numerical analyses, figure preparation and manuscript drafting. [Authors to confirm the tool details and final review statement before submission.]

References

1. Mou L, Lu Y, Wu Z, Zhang Q, Pu Z. Reference composition biases automated cell annotation and obscures biological signals in xenotransplantation single-cell atlases. Xenotransplantation. 2026;33(5):e70171. https://doi.org/10.1111/xen.70171
2. Hao Y, Hao S, Andersen-Nissen E, et al. Integrated analysis of multimodal single-cell data. Cell. 2021;184:3573–3587.e29. https://doi.org/10.1016/j.cell.2021.04.048
3. Yalabadi AK, Amirabad AD, Moskalev A, et al. Disentangling marginal capacity from geometry in optimal-transport label transfer. arXiv [Preprint]. 2026;2610.14853. https://arxiv.org/abs/2610.14853
4. Cuturi M. Sinkhorn distances: Lightspeed computation of optimal transport. Advances in Neural Information Processing Systems. 2013;26:2292–2300. https://arxiv.org/abs/1306.0895

5. Thornton J, Cuturi M. Rethinking initialization of the Sinkhorn algorithm. Proceedings of Machine Learning Research. 2023;206:8682–8698. https://proceedings.mlr.press/v206/thornton23a.html

6. Eisenberger M, Toker A, Leal-Taixé L, Bernard F, Cremers D. A unified framework for implicit Sinkhorn differentiation. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022:509–518. https://arxiv.org/abs/2205.06688

7. Li X, Lu F, Tao M, Ye FXF. Robust first- and second-order differentiation for regularized optimal transport. SIAM Journal on Scientific Computing. 2025;47:C630–C654. https://doi.org/10.1137/24M1674030

8. Bünger F. On uniqueness, analyticity, and first- and second-order derivatives of Sinkhorn-type DAD scalings. Linear Algebra and its Applications. 2026;737:173–192. https://doi.org/10.1016/j.laa.2026.02.004

9. Kelley CT. Iterative Methods for Linear and Nonlinear Equations. Philadelphia: SIAM; 1995. https://doi.org/10.1137/1.9781611970944

10. Tang Z, Qiu Y. Safe and sparse Newton method for entropic-regularized optimal transport. Advances in Neural Information Processing Systems. 2024;37:129914–129943. https://proceedings.neurips.cc/paper_files/paper/2024/hash/ea8620683340facbd5f754dd169e0980-Abstract-Conference.html

11. Wang C, Qiu Y. The sparse-plus-low-rank quasi-Newton method for entropic-regularized optimal transport. Proceedings of Machine Learning Research. 2025;267:64202–64221. https://proceedings.mlr.press/v267/wang25ck.html

12. Moosavi-Dezfooli SM, Fawzi A, Frossard P. DeepFool: A simple and accurate method to fool deep neural networks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016:2574–2582. https://openaccess.thecvf.com/content_cvpr_2016/html/Moosavi-Dezfooli_DeepFool_A_Simple_CVPR_2016_paper.html

13. Kang HM, Subramaniam M, Targ S, et al. Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. Nature Biotechnology. 2018;36:89–94. https://doi.org/10.1038/nbt.4042

14. Flamary R, Courty N, Gramfort A, et al. POT: Python Optimal Transport. Journal of Machine Learning Research. 2021;22(78):1–8. https://jmlr.org/papers/v22/20-451.html

15. Lu Y, Yang A, Quan C, et al. A single-cell atlas of the multicellular ecosystem of primary and metastatic hepatocellular carcinoma. Nature Communications. 2022;13:4594. https://doi.org/10.1038/s41467-022-32283-3

16. Habermann AC, Gutierrez AJ, Bui LT, et al. Single-cell RNA sequencing reveals profibrotic roles of distinct epithelial and mesenchymal lineages in pulmonary fibrosis. Science Advances. 2020;6:eaba1972. https://doi.org/10.1126/sciadv.aba1972

17. Theunissen L, Mortier T, Saeys Y, Waegeman W. Uncertainty-aware single-cell annotation with a hierarchical reject option. Bioinformatics. 2024;40:btae128. https://doi.org/10.1093/bioinformatics/btae128

18. López-De-Castro M, García-Galindo A, González-Gomariz J, Armañanzas R. Conformal inference for reliable single cell RNA-seq annotation. Bioinformatics. 2025;41:btaf521. https://doi.org/10.1093/bioinformatics/btaf521

19. Piran M, Vafaee F. Optimal transport reveals unresolved cells in single cell multi-omics. bioRxiv [Preprint]. 2026. https://doi.org/10.64898/2026.10.15.759426

20. Tang X, Shavlovsky M, Rahmanian H, et al. Accelerating Sinkhorn algorithm with sparse Newton iterations. International Conference on Learning Representations. 2024. https://arxiv.org/abs/2401.12253

21. Chhaibi R, Gratton S, Vaiter S. Faster computation of entropic optimal transport via stable low frequency modes. arXiv [Preprint]. 2025;2506.14780. https://arxiv.org/abs/2506.14780

22. Amos B, Luise G, Cohen S, Redko I. Meta optimal transport. Proceedings of Machine Learning Research. 2023;202:791–813. https://proceedings.mlr.press/v202/amos23a.html

23. Kemertas M, Jepson AD, Farahmand A. Efficient and accurate optimal transport with mirror descent and conjugate gradients. Transactions on Machine Learning Research. 2025. https://arxiv.org/abs/2307.08507

24. Hahn VS, Chaffin M, Simonson B, et al. Single-Cell Analysis of Human Heart Failure With Preserved Ejection Fraction. Circ Res. 2026;139(2):e327433. https://doi.org/10.1161/CIRCRESAHA.125.327433

Tables

Table 1. Audit times after genuine reference expansion

Task and size	AA2	Shared curvature	Quadratic + curvature	SSNS + AA2	SPLR + AA2
K1 1,200 references	52.611	9.392	7.770	143.830	2/3 complete
K1 2,400 references	73.261	28.374	28.696	0/3 complete	0/3 complete
K1 4,800 references	0/3 complete	87.047	68.577	0/3 complete	0/3 complete
H1 1,200 references	5.254	0.961	1.014	117.601	156.551
H1 2,400 references	7.371	2.197	2.476	0/3 complete	0/3 complete
H1 4,800 references	15.984	7.221	6.432	0/3 complete	0/3 complete

All times are seconds. Each task contains 3,000 queries and 64 local scenarios at $\varepsilon = 0.05$. Numeric entries are medians of three completed fresh-process runs and include method-specific preparation, prediction, correction, fallback and class readout. Common preparation and baseline fitting are separate. Incomplete entries give the number completed out of three planned runs; skipped repeats after an initial timeout are included in that denominator. K1, Kang fold 1; H1, Hao P1.

Table 2. Numerical agreement in the expanded benchmark

Assessment	Result
Completed audits meeting the common marginal tolerance	152
Independent configuration–scenario POT reference solutions	36
Method–repeat comparisons with those reference solutions	456
Maximum absolute class-probability difference from POT	7.19538×10^{-6}
Maximum label-disagreement fraction from POT	0
Maximum probability difference across all scenarios paired with AA2	1.41306×10^{-5}
Maximum label-disagreement fraction across those AA2 pairs	0

Reference solutions use a full marginal L1 tolerance below 10^{-10}; evaluated workflows use 10^{-7}. Repeated comparisons are numerical checks, not independent biological samples.

Table 3. Search yield across annotation granularities

Reference cells	Annotation	Random	Directed	Relative gain

Reference cells	Annotation	Random	Directed	Relative gain
300	Native	72.92	91.33	25.3%
1200	Native	70.75	84.33	19.2%
1200	Merged	52.67	54.67	3.8%
1200	T / non-T	40.50	40.50	0.0%

Mean sensitive cells per 1,000 queries. Four tasks and three random seeds per row; 32 scenarios at radius 0.10. Counts are paired within a configuration. Random seeds and repeated reference sizes are not independent biological replicates.

Table 4. Sensitive cells recovered in a fixed review list

Task	References	Confidence	Entropy	Winning margin	Linear radius
Hao P1	300	59	43	82	97
Hao P1	1200	64	52	87	100
Hao P2	300	65	50	66	85
Hao P2	1200	51	40	63	83
Kang fold 1	300	36	23	39	52
Kang fold 1	1200	34	23	35	39
Kang fold 2	300	45	25	46	55
Kang fold 2	1200	48	31	48	54
Mean	—	50.250	35.875	58.250	70.625

Each list contains 100 of 1,000 cells. The evaluation set is the union of label changes under 96 random directions at radius 0.10, generated without ranking scores. Smaller winning margin and smaller linearized radius indicate higher priority. All comparisons use identical cells and sampled sensitivity targets; stable ties use archived query order. Native labels are used throughout. Larger counts represent more sampled composition-sensitive cells found, not more biological errors corrected.

Reference composition sensitivity analysis

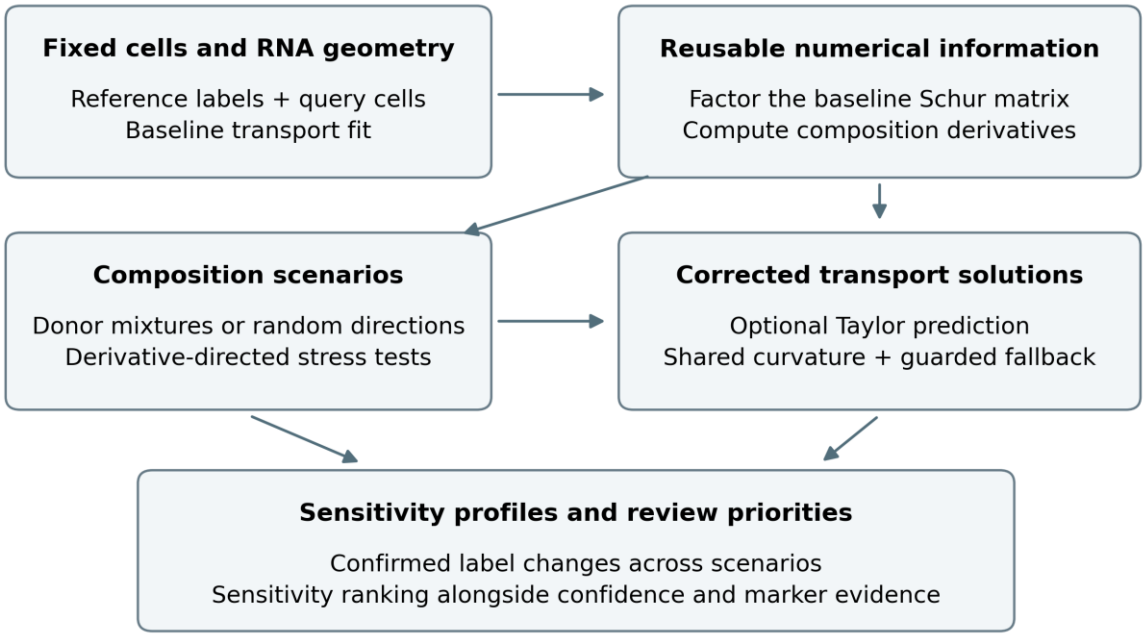


Figure 1. Shared numerical information supports repeated composition analysis and review. Fixed query and reference cells define the expression geometry. Baseline curvature is factored once and reused for correction. Implicit composition derivatives support Taylor initialization, directed scenario proposals and a local sensitivity ranking. Proposed scenarios are fully corrected before label changes are reported. Donor mixtures and random scenarios provide complementary stress-test inputs; marker or protein evidence is used in subsequent biological review.

Composition sensitivity differs across immune cell populations

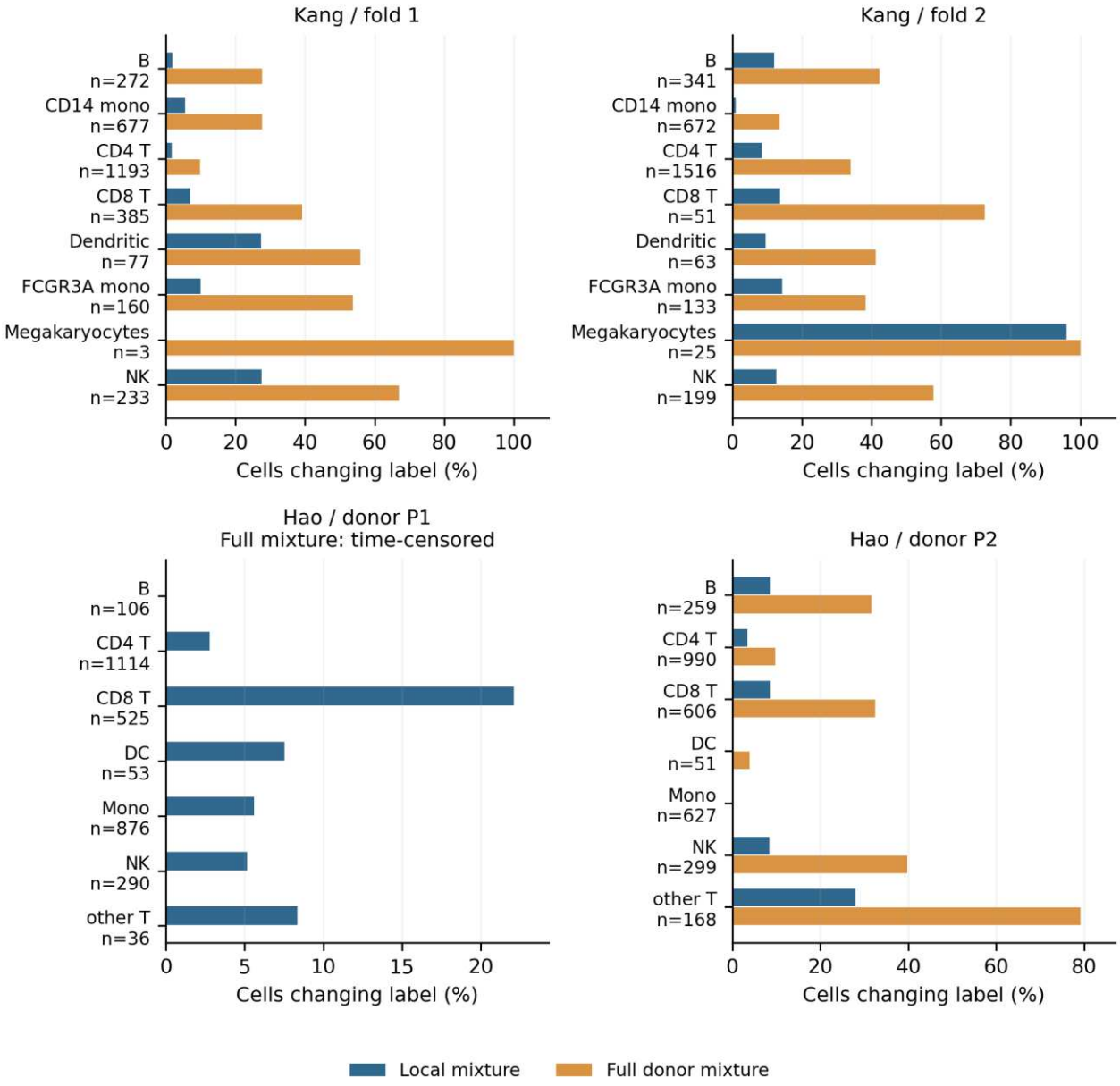


Figure 2. Composition sensitivity differs across immune cell populations. Bars show the percentage of cells in each author-annotated class whose highest-probability label changes in at least one of 256 scenarios at ε = 0.05. Blue denotes local mixtures and orange full donor mixtures. Cell counts are printed beside each class. The full-mixture Hao P1 analysis did not complete within the budget and is not plotted. Each panel has its own horizontal scale. Very small groups, including three Kang fold-1 megakaryocytes, are descriptive.

Protein context and annotation uncertainty

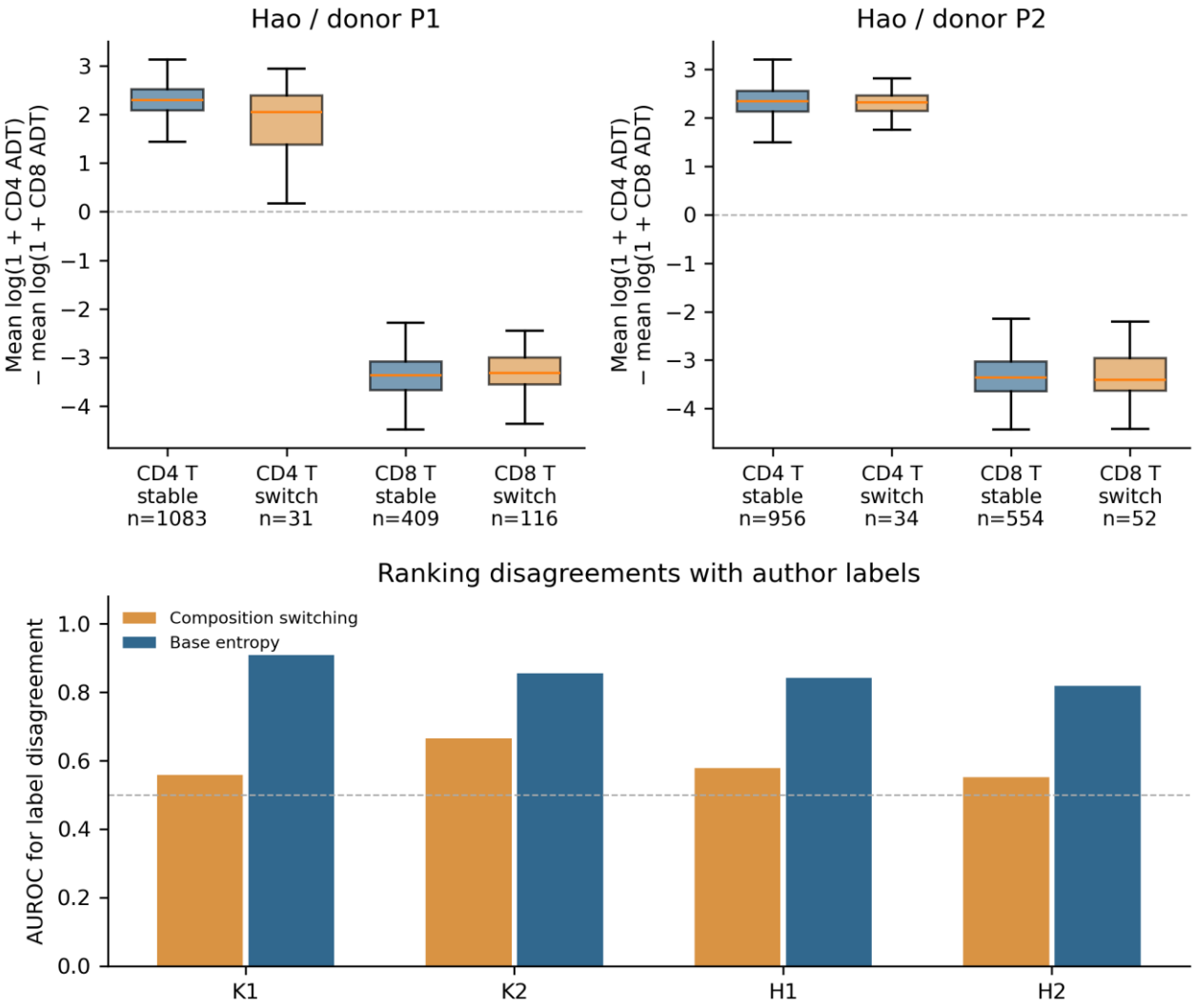


Figure 3. Protein context and annotation uncertainty. Upper panels show CD4-minus-CD8 protein scores for stable and switching author-annotated T cells in Hao P1 and P2, under local mixtures at ε = 0.05. Boxes show interquartile ranges, center lines medians and whiskers the 1.5-interquartile-range convention; outliers are hidden. The zero line is a descriptive contrast, not a calibrated cell-type gate. Lower bars compare switching frequency and baseline entropy for ranking disagreements with author labels in the four local-mixture tasks. K1/K2 denote Kang folds and H1/H2 Hao P1/P2. Proteins were withheld from RNA fitting but contributed to the original multimodal labels.

Acceleration and the incremental contribution of quadratic prediction

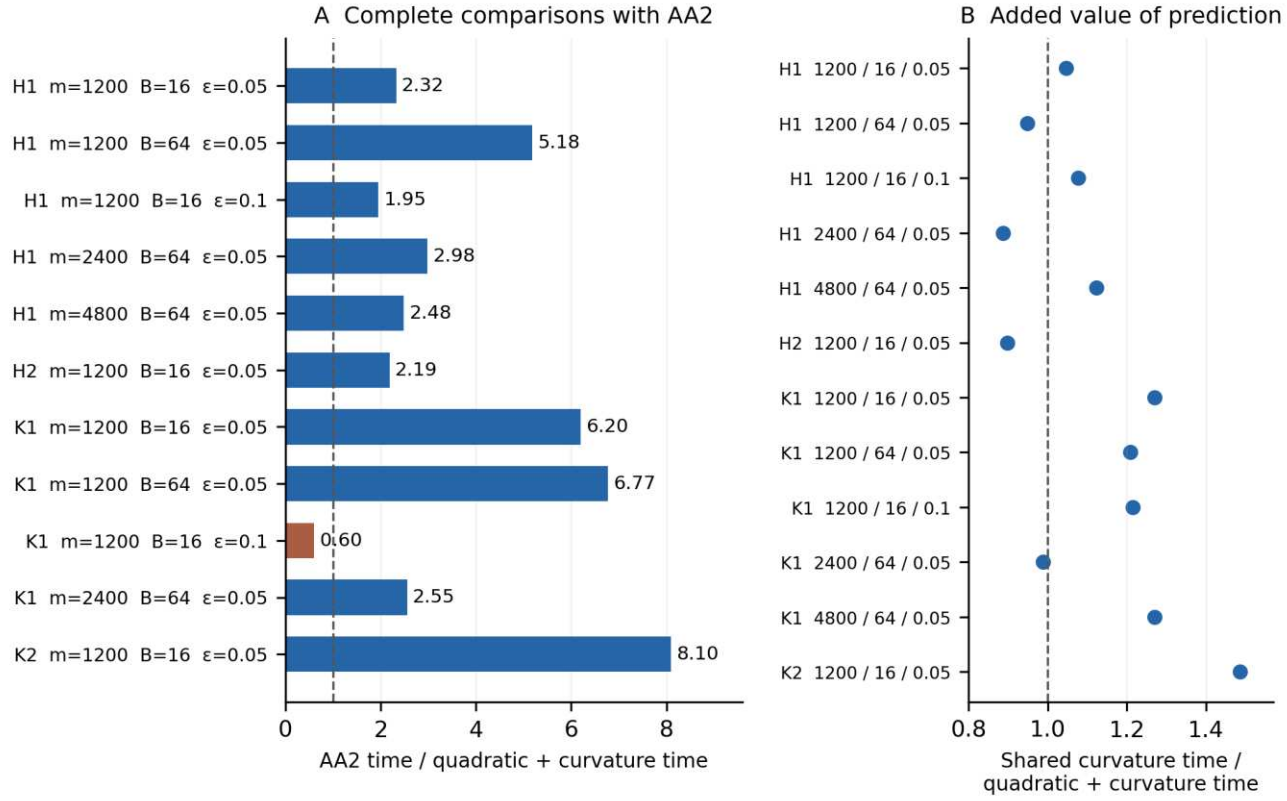


Figure 4. Acceleration and the incremental contribution of quadratic prediction. A, ratios of median AA2 time to median quadratic shared-curvature time for all 11 complete paired configurations. Ratios above one favor the proposed workflow; the slower configuration is shown in brown. Kang fold 1 with 4,800 references lacks a complete AA2 comparison and is excluded from the ratio plot. B, ratios of shared-curvature time without prediction to time with quadratic prediction, for all 12 configurations. Ratios above one indicate a prediction benefit. Labels identify task, reference count m, scenario count B and ε; panel B uses the compact order m / B / ε. Each median comes from three fresh-process runs.

Runtime and memory under genuine reference expansion

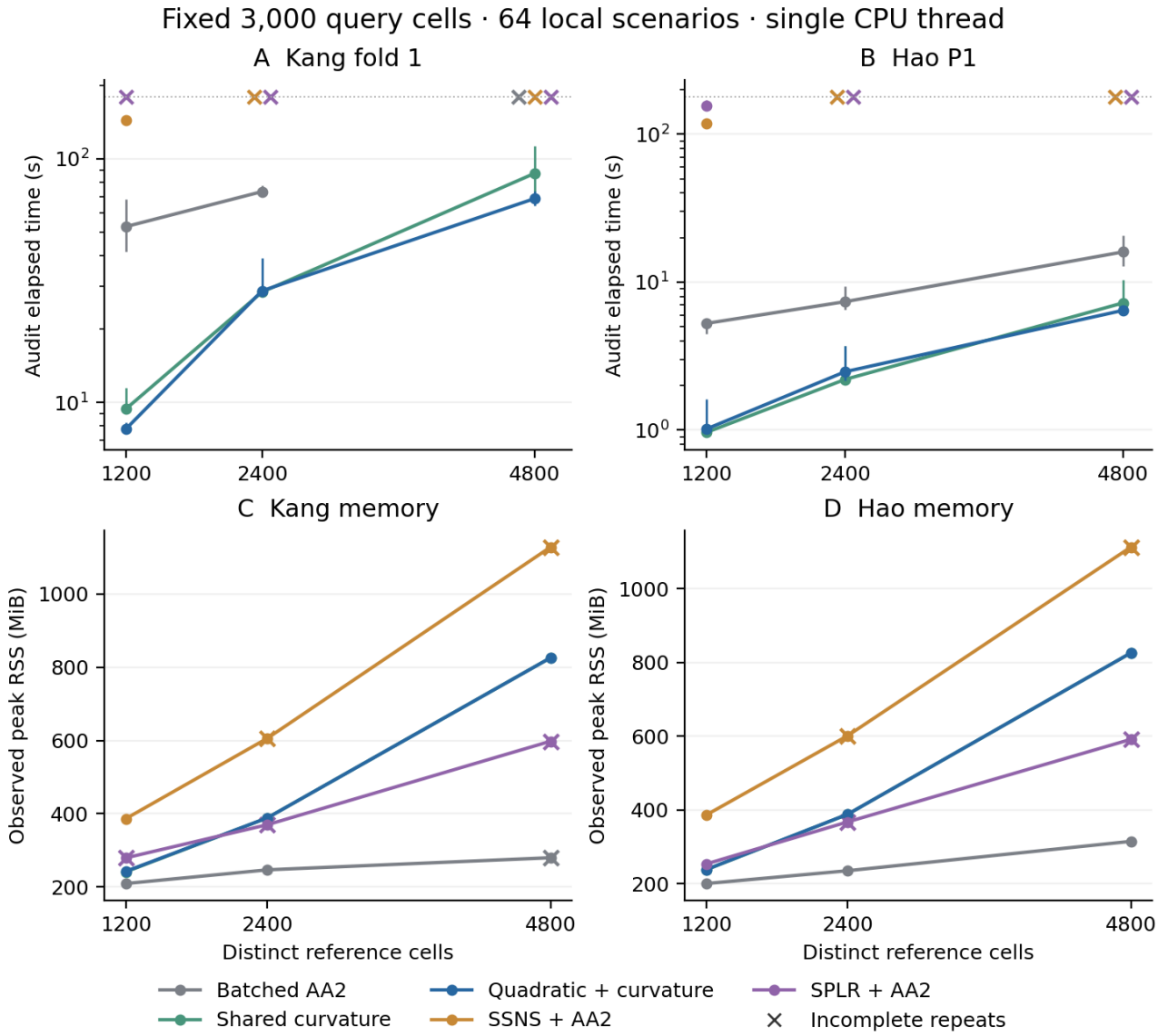


Figure 5. Runtime and memory under genuine reference expansion. Each task retains 3,000 query cells, with 64 local scenarios and ε = 0.05. A and B, median audit times and observed repeat ranges on a logarithmic axis. Crosses at 180 s indicate configurations lacking three completed repeats, rather than measured completion times; overlapping crosses are offset slightly. C and D, maximum observed audit RSS across attempted repeats. Memory before timeout is an incomplete-run observation. The blue curve is quadratic prediction with shared curvature and the green curve shared curvature without prediction. SSNS + AA2 and SPLR + AA2 are the safeguarded adaptations described in Methods.

Derivative-guided composition stress tests

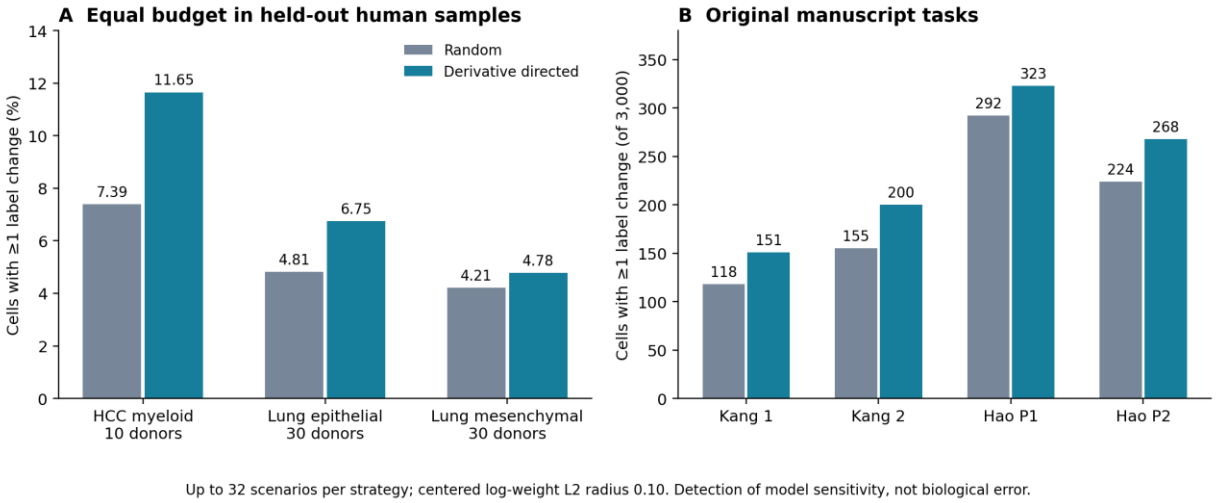


Figure 6. Equal-budget detection of composition-sensitive annotations. A, donor-mean percentage of all query cells changing labels under at least one corrected scenario, at centered log-weight L2 radius 0.10. Each strategy uses one scenario per uniformly chosen anchor, up to 32 per sample. Liver tissues are averaged within patient; lung compartments share the same 30 donors. B, sensitive-cell counts in the four unchanged original tasks, with 3,000 queries and 32 scenarios each. Gray, random directions; teal, derivative-directed directions. These outcomes measure model dependence on reference weights, not biological misclassification. Paired intervals and radius 0.20 results are in the accompanying tables.

Operating range and review prioritization

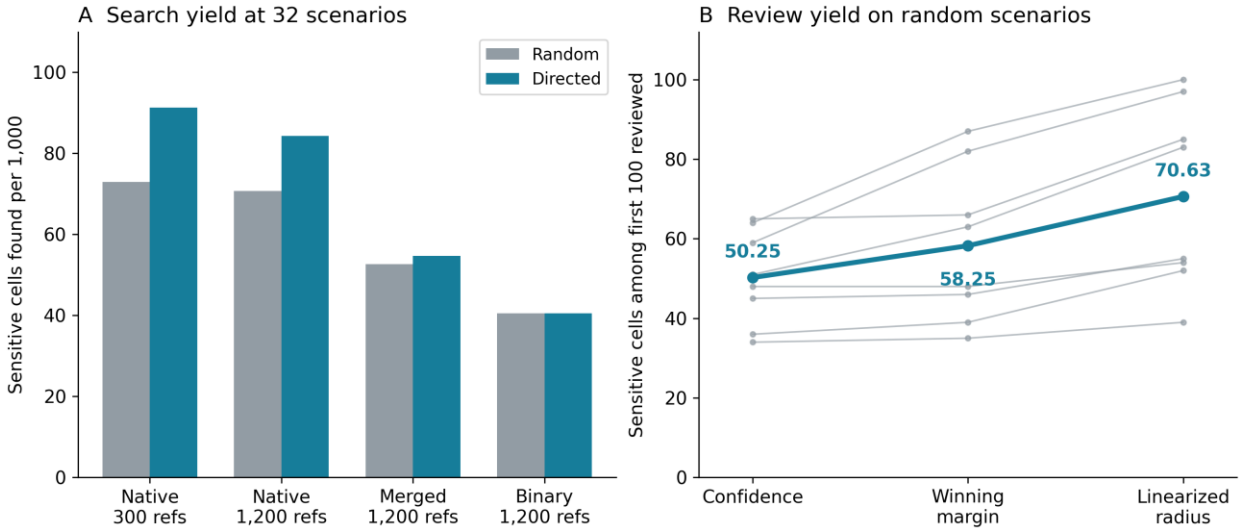


Figure 7. Annotation granularity and cell review priorities. A, mean sensitive-cell counts per 1,000 queries at 32 scenarios and log-weight radius 0.10, averaged over four immune-cell tasks and three seeds. Gray, random directions; teal, derivative-directed directions. Merged and binary taxonomies define different perturbation problems. B, random-scenario sensitive cells recovered among the first 100 reviewed using confidence, winning-score margin or linearized composition radius. Thin gray lines pair eight native-label task-by-reference configurations; the thick teal line shows their mean. The margin comparison uses frozen probabilities and the same evaluation set. Reference sizes reuse queries. Counts describe sensitivity detection, not biological errors.

Supplementary information

Supplementary methods on prediction and reproducibility

For composition displacement $\delta = p - p_0$, the implementation uses zero-sum reference directions $D = W - b_0 1^T$, for which $D\delta = W\delta$. The first derivatives of the query and reference log-scalings solve the differentiated marginal equations. Eliminating the query block gives the cached Schur system in Methods. Mixed second derivatives use the same factorization with right-hand sides formed from products of first derivatives. The predictor sums linear terms, half-weighted diagonal quadratic terms and distinct off-diagonal mixed terms. All finite predictions are corrected before probabilities are reported. METHOD_DERIVATION.md supplies the equations, gauge conventions, local probability derivatives and storage analysis; finite-difference checks are included with the code.

The initial exploratory screen evaluated 32 task–regularization–scenario configurations and five numerical variants, with 480 planned timing records. It retained 465 completed runs, five failed or time-budget-limited attempts and ten omitted repeats after an earlier failure. Across 31 completed paired configurations, the median AA2-to-quadratic-shared-curvature time ratio was 2.90 (range 0.62–8.53). Four single-thread task workers shared the host during that screen. Separate earlier isolated 64-scenario runs produced 5.26-fold and 4.89-fold speedups on Kang fold 1 and Hao P1. These earlier timings are not pooled with the later sequential expanded benchmark summarized in the main text.

The initial screen included 93 POT checks with maximum probability difference 1.75925×10^{-5} and no label differences in those checks. A stricter secondary experiment used a workflow tolerance of 10^{-9} and gave a maximum probability difference of 6.99×10^{-8} in completed checks. The Hao P1 full-mixture 64-scenario strict configuration remained unfinished. Across the initial screen's timed variants, the largest recorded probability difference was 1.89×10^{-4} and the largest differing-label fraction was 1.3×10^{-6} across cell–scenario pairs. These finite-tolerance results are retained separately from the expanded benchmark.

At $\varepsilon = 0.05$, the initial Hao P1 full-mixture 256-scenario workload exceeded the per-call budget for all attempted variants. Its missing complete biological profile is retained as an unfinished result. Individual diagnostic POT solves did converge. This earlier large-perturbation boundary is distinct from the later successful 64-local-scenario expansion experiments.

The expanded benchmark retains the original protocol, the amendment prompted by default-solver timeouts and the comparator source patch. Pre-correction CG results were isolated. Four records with missing parent-side memory-sampling fields were retained as instrumentation diagnostics and rerun under the same algorithm parameters. Two early corrected comparator records lack an individual binary-hash field; their build and execution logs identify the version, and this exception is recorded in release_audit.json. The distributed raw tables, source hashes and environment records document these details without discarding unsuccessful workloads.

Human tissue preprocessing and diagnostic analysis details

The tissue extension reused archived annotated count matrices and per-capture scDblFinder decisions. Basic bounds were at least 1,000 detected genes and mitochondrial fraction ≤25% for lung; at least 200 counts, 200–8,000 genes and mitochondrial fraction ≤10% for liver. Cells marked as doublets were excluded. No ambient-RNA correction was estimated without unfiltered droplet data. Counts were library-normalized to 10,000 and log-transformed. Genes detected in at least five sampled references with mean >0.005 were eligible; mitochondrial and ribosomal genes and the archived program/source-endpoint genes were excluded as frozen in the code. Dispersion was standardized within 20 mean bins. The top 2,000 eligible genes were scaled using reference means and a standard-deviation floor of 0.05, clipped to ±10, then projected through a reference-fitted 30-component PCA. No integrated atlas or batch-corrected latent model was claimed.

All original and tissue directed-search results are retained at both radii, including random-favored donors. The calibrated sensitivity model adds four features to the three confidence features; geometry-augmented variants additionally include nearest-reference distance. All models used the same calibration split and C = 1 without test-label tuning. The target coverage used integer floor(coverage × sample size), with at least one retained cell. Unknown-class AUROC requires both withheld and supported labels in the sample. Four lung mesenchymal donors lacked such an evaluable contrast, leaving 26 for that summary. One incomplete compressed result file was preserved and regenerated deterministically before evaluation. The predeclared protocol was not externally registered.

Complete expanded benchmark

Table S1. All 12 configurations and five workflows

Configuration	AA2	Shared curvature	Quadratic + curvature	SSNS + AA2	SPLR + AA2
K1 m=1,200 ε=0.05 B=16	23.153	4.739	3.734	38.229	43.759
K1 m=1,200 ε=0.1 B=16	0.889	1.796	1.479	49.642	31.584
H1 m=1,200 ε=0.05 B=16	1.445	0.651	0.622	34.692	33.699
H1 m=1,200 ε=0.1 B=16	0.880	0.485	0.450	28.322	27.150
K1 m=1,200 ε=0.05 B=64	52.611	9.392	7.770	143.830	2/3 complete
H1 m=1,200 ε=0.05 B=64	5.254	0.961	1.014	117.601	156.551
K2 m=1,200 ε=0.05 B=16	12.086	2.218	1.493	42.937	47.694
H2 m=1,200 ε=0.05 B=16	1.163	0.477	0.532	30.843	34.035
K1 m=2,400 ε=0.05 B=64	73.261	28.374	28.696	0/3 complete	0/3 complete
K1 m=4,800 ε=0.05 B=64	0/3 complete	87.047	68.577	0/3 complete	0/3 complete
H1 m=2,400 ε=0.05 B=64	7.371	2.197	2.476	0/3 complete	0/3 complete
H1 m=4,800 ε=0.05 B=64	15.984	7.221	6.432	0/3 complete	0/3 complete

Seconds; medians require three completed repeats. Every task has 3,000 queries. K1/K2 are Kang folds 1/2; H1/H2 are Hao P1/P2. m is the reference count and B the number of local scenarios. A 0/3 entry comprises one timeout and two unrun repeats; the SPLR 2/3 entry comprises two completions and one timeout. The per-audit budget is 180 s. Raw repeat ranges and outcome records are supplied in results/followup/timing_summary.csv and timings.csv.

Memory and common preparation costs

Table S2. Maximum observed audit RSS in MiB

Task and size	AA2	Shared curvature	Quadratic + curvature	SSNS + AA2	SPLR + AA2
K1 1,200 references	209.3	241.3	242.7	386.5	279.3†
K1 2,400 references	246.7	388.7	388.7	606.0†	370.0†
K1 4,800 references	279.8†	825.8	825.7	1128.1†	597.8†
H1 1,200 references	200.9	238.3	238.7	386.5	253.9
H1 2,400 references	235.7	388.6	388.7	601.6†	368.0†
H1 4,800 references	315.2	825.6	825.7	1112.4†	591.5†

The maximum includes loaded inputs and method working arrays during attempted audits. † marks configurations without three completed repeats; these observations do not estimate memory required for successful completion. Conditions match Table 1.

Table S3. Common matrix preparation and baseline fitting

Task	References	Distance and matrix preparation s	Baseline fit s
K1	1,200	0.207	3.302
K1	2,400	0.428	0.906
K1	4,800	0.447	39.107
H1	1,200	0.210	0.818
H1	2,400	0.305	1.875
H1	4,800	0.461	3.716

One preparation record per task and size at $\varepsilon = 0.05$; these are not medians of repeated fits. Raw-data download, gene selection, normalization and the previously fitted PCA are outside these timings. Each task has 3,000 queries.

Added validation summaries

Table S4. All-query sensitive-cell detection at radius 0.10

Compartment	Donors	Random (%)	Directed (%)	Difference (pp) [95% interval]
Liver myeloid	10	7.39	11.65	4.26 [3.39, 4.94]
Lung epithelial	30	4.81	6.75	1.93 [1.35, 2.53]
Lung mesenchymal	30	4.21	4.78	0.58 [0.22, 0.97]

Paired donor means. Same scenario budget and numerical tolerance; all query cells are evaluated. Intervals are conditional, exploratory percentile donor-bootstrap intervals. The two lung rows share donors. Full radius 0.20 and anchor-hit results are in paired_donor_comparisons.csv.

Table S5. Disagreement at 80% retained fine-label coverage

Compartment	Confidence (%)	+ Sensitivity (%)	Difference (pp) [95% interval]
Liver myeloid	48.075	48.057	-0.019 [-0.556, 0.546]
Lung epithelial	5.604	5.227	-0.376 [-0.588, -0.182]
Lung mesenchymal	4.558	4.270	-0.288 [-0.895, 0.075]

Reference outcome is disagreement with author labels, not experimentally established error. Both models are calibrated on donors separate from the test donors. Lower is better. No label correction is performed.

Table S6. Controlled withheld-class detection AUROC

Compartment	Confidence	Switch rate	Reference distance
Liver myeloid	0.684	0.687	0.626
Lung epithelial	0.531	0.553	0.926
Lung mesenchymal	0.801	0.796	0.861

Mean across evaluable donors: liver 10, lung epithelial 30, lung mesenchymal 26. Removed classes were selected using reference data only. All alternative scores and known/unknown retention rates are in missing_class_challenge.csv.

Selective annotation and reference-class removal

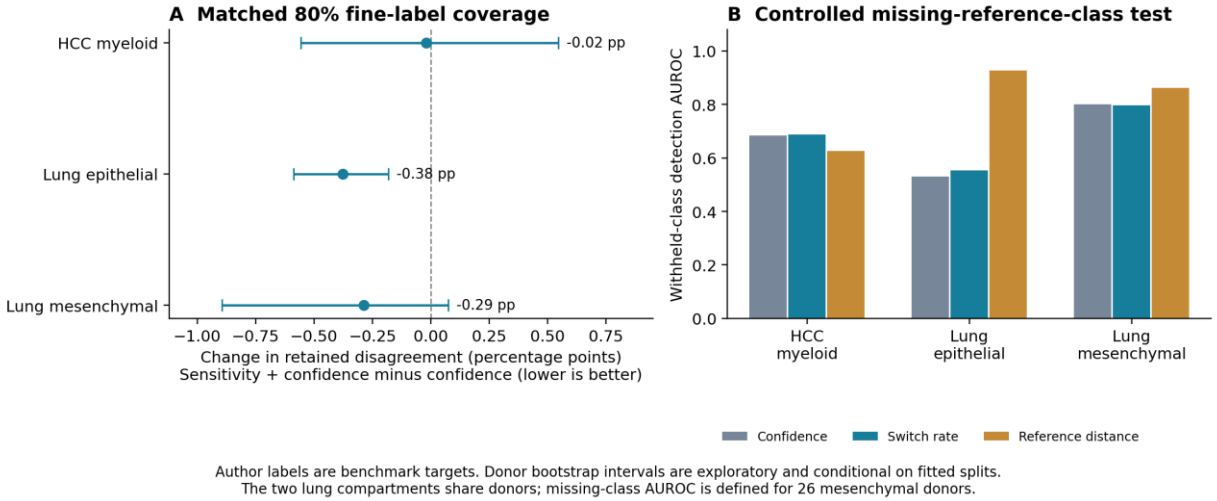


Figure S1. Additional sensitivity features have limited annotation benefits. A, donor-mean change in retained disagreement at 80% fine-label coverage for calibrated confidence plus sensitivity versus calibrated confidence alone. Negative values favor adding sensitivity. Bars show exploratory 95% donor-bootstrap intervals conditional on fitted splits. B, donor-mean AUROC for identifying a class removed from the reference, comparing ordinary confidence, switching rate with entropy tie-break and nearest-reference distance. These outcomes use author labels as benchmark targets. They do not establish experimentally verified correctness or improved disease conclusions.

Applicability results across the full evaluation grid

Table S7. Detection yield at every radius and scenario budget

Reference / labels	Radius	Budget	Random	Directed	Difference
300 / native	0.03	8	15.08	21.83	6.75
300 / native	0.03	16	17.42	23.92	6.50
300 / native	0.03	32	19.75	26.58	6.83
300 / native	0.10	8	53.75	75.58	21.83
300 / native	0.10	16	64.33	83.08	18.75
300 / native	0.10	32	72.92	91.33	18.42
300 / native	0.30	8	159.50	222.83	63.33
300 / native	0.30	16	184.67	240.00	55.33
300 / native	0.30	32	209.92	259.92	50.00
1200 / binary	0.03	8	14.25	14.25	0.00
1200 / binary	0.03	16	14.25	14.25	0.00
1200 / binary	0.03	32	14.25	14.25	0.00
1200 / binary	0.10	8	40.50	40.50	0.00
1200 / binary	0.10	16	40.50	40.50	0.00
1200 / binary	0.10	32	40.50	40.50	0.00
1200 / binary	0.30	8	151.25	151.25	0.00
1200 / binary	0.30	16	151.25	151.25	0.00
1200 / binary	0.30	32	151.25	151.25	0.00
1200 / merged	0.03	8	13.83	16.67	2.83
1200 / merged	0.03	16	15.25	17.50	2.25
1200 / merged	0.03	32	17.25	18.25	1.00
1200 / merged	0.10	8	42.75	48.50	5.75
1200 / merged	0.10	16	47.50	50.58	3.08
1200 / merged	0.10	32	52.67	54.67	2.00
1200 / merged	0.30	8	114.08	140.08	26.00
1200 / merged	0.30	16	129.08	148.92	19.83
1200 / merged	0.30	32	146.00	162.25	16.25
1200 / native	0.03	8	16.83	20.00	3.17
1200 / native	0.03	16	19.33	22.42	3.08
1200 / native	0.03	32	22.17	26.08	3.92
1200 / native	0.10	8	52.75	65.58	12.83
1200 / native	0.10	16	62.67	72.25	9.58
1200 / native	0.10	32	70.75	84.33	13.58
1200 / native	0.30	8	152.75	190.50	37.75
1200 / native	0.30	16	173.00	212.33	39.33
1200 / native	0.30	32	193.50	238.67	45.17

Means over four tasks and three seeds; 1,000 fixed queries per task. Negative individual outcomes remain in the accompanying per-seed tables. Budgets of 8 and 16 are nested prefixes of the same 32-scenario runs. Radius is measured in centered log-weight coordinates.

Coverage dependence and disease subgroups

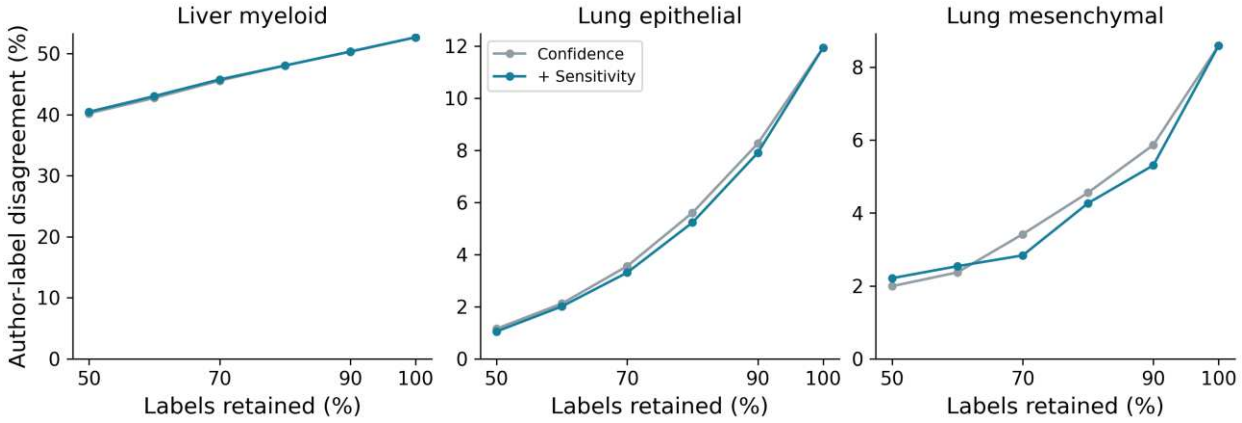


Figure S2. Held-out selective-annotation coverage curves. Each sample retains the lowest predicted-risk fraction under the confidence-only or confidence-plus-sensitivity model. Liver samples are averaged within patient before computing donor means. Lower disagreement is better; author annotations are the evaluation reference. The curves reuse the fitted calibration models without threshold optimization on test labels.

Table S8. Condition-stratified disagreement at 80 percent coverage

Compartment / condition	Donors	Confidence %	+ Sensitivity %	Difference pp
Liver Myeloid / Normal	8	18.021	17.971	-0.051
Liver Myeloid / Tumor	10	65.970	66.052	0.082
Lung Epithelial / Control	10	1.682	1.539	-0.143
Lung Epithelial / ILD	1	5.500	3.500	-2.000
Lung Epithelial / IPF	12	9.531	9.026	-0.505
Lung Epithelial / NSIP	3	3.796	3.463	-0.333
Lung Epithelial / Sarcoidosis	2	3.846	3.698	-0.148
Lung Epithelial / cHP	2	6.166	5.916	-0.250
Lung Mesenchymal / Control	10	3.667	2.833	-0.833
Lung Mesenchymal / ILD	1	0.000	0.000	0.000
Lung Mesenchymal / IPF	12	8.083	8.102	0.019
Lung Mesenchymal / NSIP	3	0.833	0.750	-0.083
Lung Mesenchymal / Sarcoidosis	2	0.000	0.000	0.000
Lung Mesenchymal / cHP	2	0.281	0.140	-0.140

Differences are sensitivity minus confidence. Subgroups reuse held-out predictions and were not used to refit models. Small disease groups are descriptive. Full coverage values and conditional bootstrap intervals are provided in condition_coverage.csv; comparisons are exploratory and not multiplicity-adjusted.

Additional review ranking comparisons

Table S9. Review yield across taxonomies and radii

Labels	Radius	Confidence	Entropy	Margin	Linear radius
Binary	0.03	14.000	14.000	14.000	14.250
Binary	0.10	36.750	36.750	36.750	38.500
Binary	0.30	80.000	80.000	80.000	86.250
Merged	0.03	17.750	16.000	17.500	18.250
Merged	0.10	46.500	43.500	46.250	51.000
Merged	0.30	84.000	83.500	82.750	91.000
Native	0.03	18.750	11.625	22.125	24.000
Native	0.10	50.250	35.875	58.250	70.625
Native	0.30	89.000	75.875	93.250	99.125

Mean sensitive cells among the first 100 of 1,000 reviewed. Native rows average eight task-by-reference configurations; merged and binary rows average four. Each target is defined by 96 random scenarios at the stated radius. The margin comparator was added after review without changing targets, score definitions or existing outputs. Full configuration-level counts, including ties or negative differences, are supplied in review_margin_all.csv.

In native-label configurations at radius 0.10, increments over margin ranking were 13 and 15 cells for Hao P1 (1,200 and 300 references), 20 and 19 for Hao P2, 4 and 13 for Kang fold 1, and 6 and 9 for Kang fold 2. These are paired descriptive configuration comparisons. Shared queries and overlapping donor pools preclude treating the eight rows as eight independent biological replications.

Liver fine cluster transfer diagnostics

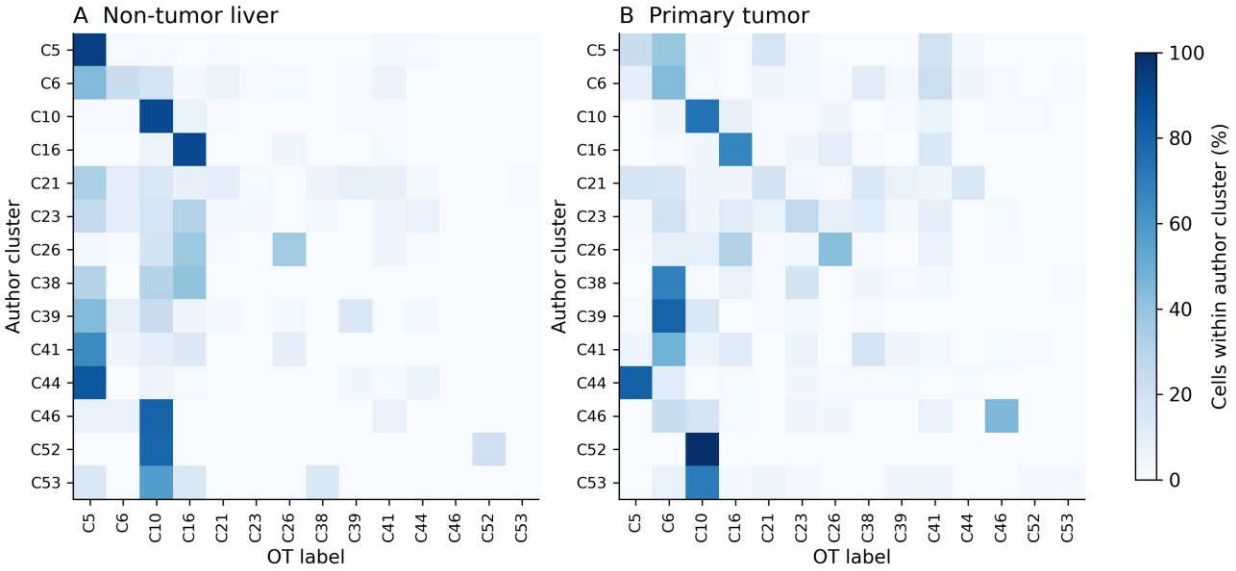


Figure S3. Held-out liver fine-cluster confusion. Rows are the original author-provided myeloid clusters and columns the OT assignments. Percentages are normalized within each author cluster and condition, pooling held-out cells for visualization. A, 3,436 non-tumor cells; B, 4,595 tumor cells. Donor-mean agreement is reported separately in Table S10. C-numbered clusters are retained without assigning unvalidated biological identities.

Table S10. Liver agreement by tissue condition

Condition	Donors	Cells	OT agreement %	kNN agreement %
Non-tumor	8	3436	74.17	75.74
Tumor	10	4595	31.38	31.76

Donor-mean agreement against original fine clusters, using every held-out query before selective retention. The existing 15-neighbor baseline uses the same fitted expression geometry and reference-frequency adjustment. Eight non-tumor and ten tumor donors overlap; they are not 18 independent patients. No tested author-label category was absent from its reference. Full sample results, reference support and confusion counts are included as CSV files.

Realized reference proportion changes

Table S11. Proportion changes at fixed log-weight radii

Compartment	Radius	Directions	Median TV %	Max TV %	Max class change pp
Liver myeloid	0.10	Directed	1.68	2.78	2.78
Liver myeloid	0.10	Random	1.00	2.06	2.04
Liver myeloid	0.20	Directed	3.31	5.55	5.55
Liver myeloid	0.20	Random	1.99	4.11	4.08
Lung epithelial	0.10	Directed	2.12	3.04	3.04
Lung epithelial	0.10	Random	0.90	2.25	2.25
Lung epithelial	0.20	Directed	4.14	6.07	6.07
Lung epithelial	0.20	Random	1.82	4.51	4.50
Lung mesenchymal	0.10	Directed	2.40	3.19	3.19
Lung mesenchymal	0.10	Random	1.45	3.07	3.07
Lung mesenchymal	0.20	Directed	4.80	6.36	6.36
Lung mesenchymal	0.20	Random	2.90	6.13	6.13

TV is one half of the summed absolute changes in class proportions. Max class change is the largest absolute change in any class over the corresponding scenarios, in percentage points. Summaries pool the frozen held-out scenarios and are descriptive; cells, donors and repeated scenarios are not independent replicates. Equal log-weight radii do not imply equal TV distances.

For comparison, the median total variation between an individual reference donor’s author-label proportions and the fold’s donor-mean composition was 39.97% in liver myeloid, 35.42% in lung epithelial, 29.23% in lung mesenchymal. These values use QC-retained cells before query capping, with the same reference class sets. Donors recur across reference folds. This comparison provides a descriptive scale for local perturbations, not a confidence region or a recommendation to adopt the observed donor extremes.