BEYOND CELL REALISM: RECOVERING DISEASE SIGNALS WITH BULK-TO-SINGLE-CELL GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Bulk RNA-seq provides transcriptomic data for large patient cohorts, but averaging across cell populations can obscure disease-associated expression changes within specific cell types. Single-cell profiling can resolve these changes, yet its cost limits cohort-scale measurements, and many existing bulk cohorts lack matched single-cell data. Generating cell-level expression profiles from bulk could extend such analyses to these cohorts. The central question is whether the generated profiles preserve the differences between patient groups that motivate disease analysis. We evaluate this through the recovery of differentially expressed genes (DEGs): genes showing significant expression differences between patient groups within a specified cell type. For example, our breast cancer evaluation compares cancer epithelial expression between triple-negative and estrogen-receptor-positive tumors. Because each cohort poses a different disease question, the relevant patient groups, cell types, and gene-level evidence must be established for each evaluation. We therefore propose an agentic benchmarking workflow that translates a cohort-specific disease question into a traceable DEG-recovery evaluation. It audits patient pairings and disease contrasts, retrieves supporting evidence from publications and supplements, and assembles a reference of disease-relevant DEGs detected in measured cell-type profiles but missed by matched bulk analysis. This design allows the evaluation to be tailored to different cohorts and biological questions while retaining a common patient-level scoring procedure. Furthermore, to reflect the practical use of existing bulk cohorts, models learn from external reference cells and receive only measured bulk from new patients, without target-cell training or calibration. Under this protocol, the tested generative baselines score zero on important DEG recovery in breast cancer and T-cell leukemia. We address this recovery gap with a patient-conditioned generator that combines bulk-derived context with a source-fitted expression shift. Across three paired cohorts comprising 104 participants, our method recovers 25 of 91 important cohort–gene references, compared with nine for a direct adaptation of scSemiProfiler. It recovers 10 of 26 references in breast cancer and six of 11 in T-cell leukemia, while tying that comparator at nine of 54 in sepsis. Together, the benchmark and generator connect bulk-to-single-cell generation to the disease signals needed for patient-cohort analysis.

1 INTRODUCTION

Bulk RNA-seq provides data for large patient cohorts, including about 10,000 tumors across 33 cancer types in pan-cancer studies (Hoadley et al., 2018), but averages across cell populations can obscure disease changes within a specific type. Single-cell profiling resolves this context, yet its cost limits cohort-scale measurements (Wang et al., 2021), and archived cohorts may lack suitable material (Newman et al., 2019). These constraints motivate computational cell-level analysis of existing bulk collections (Wilks et al., 2021; Lachmann et al., 2018).

To extract cell-type information from these collections, reference-based methods already infer cell fractions and sample-specific expression from bulk (Newman et al., 2019; Chu et al., 2022). Generating cell profiles offers another route (Liao et al., 2022; Wang et al., 2024; Xiao & Raue, 2026). How-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

ever, generating recognizable cell types does not by itself establish that disease signals are preserved. This requires testing whether generated profiles retain expression differences between patient groups within each cell type. This raises the question: *can generated profiles recover disease-related expression changes detected in measured cell-type profiles but missed by matched bulk analysis?*

To answer this question, we evaluate the recovery of *differentially expressed genes* (DEGs), whose group differences satisfy fixed significance and effect-size criteria. For example, in Breast, we compare triple-negative and estrogen-receptor-positive tumors within cancer epithelial cells. Genes called in measured cell-type profiles but not matched bulk form the bulk-missed reference: 173 cohort–gene signals across our three cohorts. Recovery requires generated profiles to satisfy the same criteria with the measured direction, using participants as statistical units.

Applying this criterion across cohorts, however, requires defining the contrast, cell population, participant mapping, and disease evidence for each biological question. We therefore propose an *agentic benchmarking workflow* that translates cohort-specific questions into auditable recovery evaluations. Agents reconcile cohort records and retrieve publication and supplement evidence; deterministic code fixes the axes and scores methods against a shared reference. Measured cells establish the effect, while literature and recorded priors establish disease relevance. Reference construction can thus be tailored across questions without changing the scoring procedure; independent execution on new cohorts remains untested.

Furthermore, to reflect the practical use of existing bulk cohorts, models learn from external reference cells and their pseudobulks, then receive only measured bulk from new patients. Target single-cell expression and disease contrasts are reserved for evaluation. Evaluating existing generators under this protocol reveals a recovery gap: tested generative baselines score zero on important DEGs in Breast and T-ALL under the full-cohort rule; Bulk2Space's Breast zero reflects missing required outputs (Section 3.1).

To address this recovery gap, we introduce a patient-conditioned generator. For new patients, learned donor identities are unavailable, while clinical descriptors may be too coarse or encode the disease contrast being evaluated. We therefore use each patient's bulk expression as context without supplying the disease labels under study. A Patient Bulk Encoder (PBE) conditions a flow and count decoder, and Bulk-Conditioned Shift (BCS) adds a source-learned expression adjustment. Both remain fixed during target inference. Across 104 participants, recovery is 10/26, 6/11, and 9/54 in Breast, T-ALL, and Sepsis: 25/91 overall versus 9/91 for scSemiProfiler-direct, with a Sepsis tie. External references constrain this inference but do not uniquely identify the patient's original cells.

We make two contributions:

- **An agentic benchmark for patient-level disease-signal recovery.** We connect cohort-specific questions to traceable gene references and shared participant-level scoring, and assess evidence-retrieval yield.
- **A patient-conditioned generator for bulk-only target inference.** PBE and BCS convey patient information during and after generation, evaluated with generative and mean-prediction controls and matched BCS ablations.

Figure 1 summarizes the two contributions and their information boundaries.

2 Related work

Deconvolution and expression inference. Fraction estimation and sample-specific expression inference are distinct tasks. CIBERSORTx, bMIND, and BayesPrism estimate cell-type expression from bulk (Newman et al., 2019; Wang et al., 2021; Chu et al., 2022); BuDDI models cell-type perturbation effects (Davidson et al., 2025). These methods motivate source-trained mean-prediction controls for our aggregate disease endpoint. Cell generation must be evaluated against that simpler objective, without assuming that sampling is necessary.

Inferring cells from bulk. Bulk2Space generates cell profiles from bulk before spatial reconstruction (Liao et al., 2022). scSemiProfiler combines reference-based inference with active learning to reduce the number of samples requiring single-cell profiling; it also supports external references

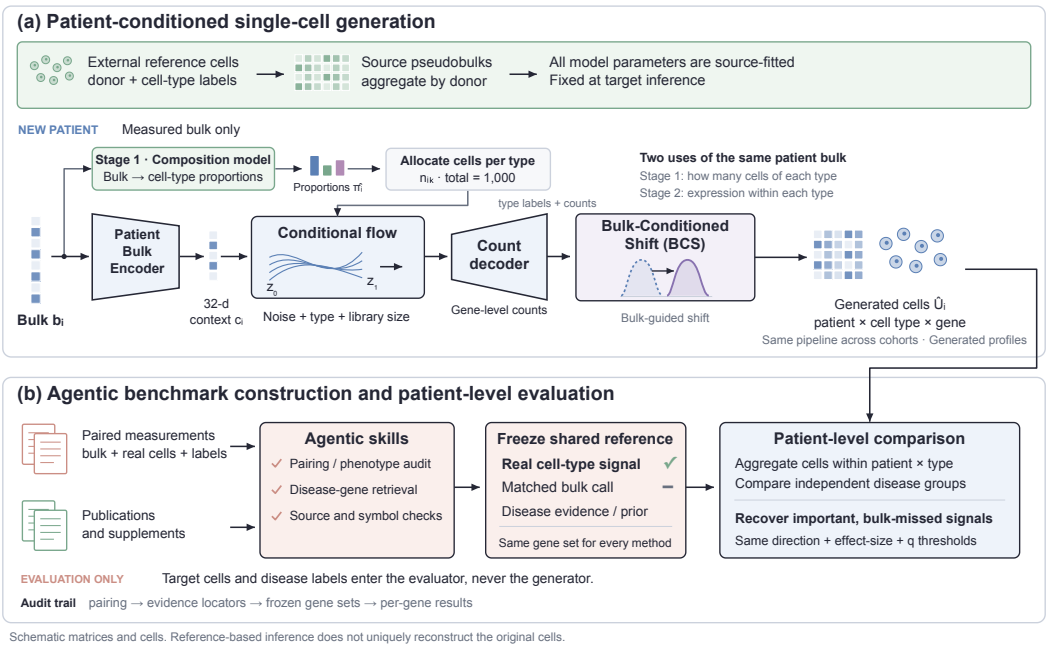


Figure 1: **Overview.** Stage 1 predicts cell-type proportions and allocates cell counts from bulk; the separate bulk encoder conditions Stage 2 expression generation. External reference cells and their pseudobulks train a patient-conditioned generator. At inference, a new patient supplies measured bulk only. The agentic workflow independently assembles paired cohorts and literature evidence for patient-level evaluation. Target single cells and disease labels are used only by the evaluator. The shift curves illustrate the operation of Bulk-Conditioned Shift; they are not measured distributions.

and evaluates disease expression and biomarkers (Wang et al., 2024). The bulk2scDiff preprint uses pseudobulk-conditioned latent diffusion and studies sample specificity, while fitting its cell representation to cohort cells; validation on independently measured bulk remains open (Xiao & Raue, 2026). These approaches differ in the information available at fitting and inference. Our benchmark focuses on physical bulk from new patients, with target single cells withheld from the predictor and used to assess cell-type disease effects.

Patient conditioning. CFGen and scDiffusion generate cells under specified conditions (Palma et al., 2025; Luo et al., 2024). CFGen can condition on donor identity through a learned categorical embedding. A previously unseen patient's identifier has no learned donor-specific representation, and the published model does not infer one from that patient's bulk. We instead compute patient context from measured expression. CFGen also improves cell-type classification on held-out donors through data augmentation; that experiment evaluates a downstream classifier, rather than generating cells matched to an unseen donor's bulk.

Evaluating biological utility. Disease-expression and biomarker analyses appear in scSemiProfiler, and the GGE preprint explicitly studies differential-expression-based evaluation of generative models (Wang et al., 2024; Rubbi et al., 2026). Our contribution is therefore not the use of disease genes or bulk conditioning alone. It is a paired evaluation of disease signals missed by bulk under an external-reference training regime, together with an auditable workflow for constructing cohort-specific gene references and a generator evaluated on that task.

3 METHODS

We first define the patient-level recovery task and its evidence reference, then describe the generator evaluated on that task. Model fitting uses external source cells and source-derived pseudobulks. At target inference, the patient supplies measured bulk; target single-cell expression and phenotype

contrasts enter evaluation only. Cell-type labels come from the source reference. The objective is recovery of within-type patient-group effects, not identification of the original cells or discovery of new cell types.

3.1 PATIENT-LEVEL DISEASE-SIGNAL RECOVERY

A *differentially expressed gene* (DEG) has an expression difference between groups that meets a specified statistical criterion. We compare disease subtypes or cases and controls within one cell population per cohort. Cells are first aggregated within each participant and cell type, so the statistical replicates are participants (Squair et al., 2021). For a cell matrix $U \in \mathbb{R}_{\geq 0}^{N \times G}$, the aggregate profile is

$$h(U) = \log\left(1 + 10^4 \frac{\sum_{c=1}^{N} U_{c,:}}{\sum_{c=1}^{N}\sum_{g=1}^{G} U_{cg}}\right). \tag{1}$$

Counts are summed before normalization and logging. For generated profiles, technical seeds are averaged on the normalized expression scale before logging.

Let C_g^R, C_g^B, and C_g^m indicate whether gene g is called differentially expressed in measured cell-type profiles, physical bulk, or method m, respectively. All methods use the same cohort-specific gene family, participant-level Welch tests, Benjamini–Hochberg adjusted $q \leq 0.05$, and an absolute mean difference of at least 0.5 in natural-log units. The set of signals missed by bulk is $\mathcal{S} = \{g : C_g^R = 1, C_g^B = 0\}$.

Our primary reference prioritizes disease-related genes: $\mathcal{S}_U = \mathcal{S} \cap (\mathcal{L} \cup \mathcal{P})$, where $\mathcal{L}$ contains literature-verified genes and $\mathcal{P}$ contains existing disease priors. A gene is recovered only if the generated profile meets the same thresholds and agrees with the measured effect direction:

$$\mathrm{Recall}_{m,U} = \frac{\sum_{g \in \mathcal{S}_U} \mathbf{1}\{C_g^m = 1,\ \Delta_g^m \Delta_g^R > 0\}}{|\mathcal{S}_U|}. \tag{2}$$

For every method, missing any required participant–focus-type profile incurs zero cohort recovery as an end-to-end failure; otherwise, Eq. 2 applies. This full-cohort reporting rule retains all participants. An assigned zero does not imply a completed differential-expression test.

A bulk-missed signal is below this calling threshold, not necessarily absent from bulk. Direct bulk has zero recovery by definition and is shown to explain the reference, rather than as evidence of model superiority. We also report recovery on the full set $\mathcal{S}$ and on the literature-only subset. The literature criterion establishes disease relevance; the paired measured cells establish the cell-type effect and its direction. Keeping these roles separate prevents a disease citation alone from serving as evidence that a generated effect is correct.

3.2 AGENTIC BENCHMARK CONSTRUCTION

The workflow turns heterogeneous cohort records and literature evidence into a shared evaluation reference. A cohort-audit skill verifies participant–specimen pairs, disease contrasts, cell labels, and source overlap. Retrieval and evidence-audit skills then search publications and supplements, reconcile gene aliases, and attach source locations to each retained claim. Deterministic code fixes the participant and gene axes, applies the statistical tests, and exports both recoveries and exclusions. The workflow separates evidence judgments from numerical scoring; structured records do not by themselves certify a biological interpretation. Algorithm 1 specifies the sequence and its outputs; Appendix D defines the skill interfaces and the boundary between agent judgments and deterministic checks.

The literature audit covers all 173 bulk-missed cohort–gene references, including those no generator recovers. Evidence for human gene-level RNA expression is distinguished from protein-only, splice-only, or incompatible-context observations. The resulting primary references contain 26, 11, and 54 genes for Breast, T-ALL, and Sepsis. Four legacy T-ALL priors lack resolved gene-specific sources; excluding them gives a literature-only reference of 26, seven, and 54 genes. Appendix D documents the evidence criteria and retrospective curation history.

Algorithm 1 Agentic benchmark construction and patient-level evaluation

Input: cohort metadata, paired bulk/cells, source manifest, fixed generated profiles, and priors $\mathcal{P}$.
Output: audited cohort, evidence ledger, evaluation contract, and scores.

1 **Audit cohort:** resolve participant pairs, contrasts, cell labels, and source overlap; attach source locators.
2 **Check admission:** exclude unresolved cohorts until required metadata checks pass.
3 **Build reference:** aggregate measured cells by participant/type; test cell-type and bulk profiles to obtain $\mathcal{S}$.
4 **Retrieve evidence:** search disease/contrast-specific publications and supplements for every gene in $\mathcal{S}$; record context, direction, and source.
5 **Audit evidence:** resolve aliases and source families; separate eligible RNA support from contextual/unresolved evidence; record exclusions and $\mathcal{L}$.
6 **Freeze:** store axes, contrasts, thresholds, ledger, and shared $\mathcal{S}_U = \mathcal{S} \cap (\mathcal{L} \cup \mathcal{P})$.
7 **Evaluate:** verify each method's profiles and provenance; assign zero if any required participant–type profile is missing; otherwise apply the same tests.
8 **Export:** count significant, direction-consistent recovery in $\mathcal{S}_U$; retain per-gene results, denominators, and failure reasons.

Boundary: target cells and the evaluation evidence ledger define evaluation only; they never update the generator. Retrieval includes unrecovered genes.

3.3 Patient-conditioned cell generation

Let b_i denote patient i's log-normalized bulk expression on a fixed 2,000-gene panel. The Patient Bulk Encoder (PBE) maps b_i to a 32-dimensional context c_i, providing a context for unseen patients without fitting a categorical donor embedding. During training, contexts are pseudobulks aggregated from external single cells. At inference, they are computed from independently measured bulk. A separate Stage 1 composition predictor maps the same patient's bulk to cell-type proportions, $\hat{\pi}_i = f_{\mathrm{S1}}(b_i)$. These proportions determine integer cell counts n_{ik} summing to 1,000. Stage 1 specifies *how many* cells of each type to generate; PBE supplies *within-type* patient context to the Stage 2 generator. Both predictors are trained on external source data. Source pseudobulks provide known relationships between aggregate expression and the cells contributing to it; they are training contexts, not substitutes for the independently measured target bulk used in the benchmark.

We adopt the count-autoencoder and conditional flow-matching components of CFGen (Palma et al., 2025; Lipman et al., 2023). A reference cell u is encoded as $z_1 = e_\psi(u)$. The flow learns to transport a Gaussian state z_0 toward this representation, conditioned on cell type k, library size L, and patient context c_i:

$$\mathcal{L}_{\mathrm{FM}} = \mathbb{E}\,\|v_\omega(z_t, t, k, \log L, c_i) - (z_1 - z_0)\|_2^2, \qquad z_t = (1-t)z_0 + tz_1 + \sigma\epsilon. \tag{3}$$

Here t is the interpolation time and ϵ is Gaussian noise. At inference, flow integration produces latent cells, and a negative-binomial count decoder maps them to gene expression at library size L. The bulk encoder supplies patient-specific information alongside the shared cell-type condition. Appendix A gives the training and decoding details.

3.4 Bulk-Conditioned Shift

Conditioning the generator on bulk does not guarantee that all patient-level expression differences survive generation. We therefore apply a source-fitted adapter, Bulk-Conditioned Shift (BCS), that maps bulk context to a cell-type-specific expression adjustment:

$$\delta_{ik}^E = \mathrm{clip}(\eta_E(b_i - \mu_E)P_E A_{E,k}, -1, 1)\,. \tag{4}$$

The source mean μ_E centers the bulk profile, P_E projects it onto source principal components, and the regression matrix $A_{E,k}$ maps that context to a shift for cell type k. The coefficient η_E sets the fixed adjustment strength. The adapter adds this context-dependent component rather than replacing the generated profile with a source cell-type mean. The regression predicts deviations of measured source cell-type profiles from their source mean. Thus PBE supplies context during latent generation, whereas BCS explicitly adds a source-learned gene-level component after decoding. Appendix A specifies the source regression, fitting target, and numerical settings.

The shift is applied to each generated cell in log-normalized expression space. Values are floored at zero, transformed back to the expression scale, and rescaled to preserve the cell's original library total (Appendix A). The adjusted values are transformed expression predictions, not necessarily integer draws from the decoder likelihood. The complete model is therefore **PBE → conditional flow → count decoder → BCS**, denoted *Ours* in comparisons. All fitted parameters remain fixed during target inference.

4 EXPERIMENTS

4.1 COHORTS AND BASELINES

We ask whether the method recovers bulk-missed disease signals, how much BCS contributes with the parent generator held fixed, and whether enhanced evidence retrieval improves reference construction. Mean-prediction controls test the aggregate endpoint without sampling cells.

We evaluate three paired cohorts comprising 104 participants, with one fixed cell population and disease contrast per cohort (Table 1). The cohorts comprise the breast cancer atlas of Wu et al. (2021), ScPCA project SCPCP000003 (Hawkins et al., 2026; Single-Cell Pediatric Cancer Atlas, n.d.), and the paired Sepsis deposits GSE279448/GSE279452 (Ye et al., 2025a;b). The training inventory separates source studies from target accessions. Known donor overlaps are excluded; unresolved cross-study identities are documented in Appendix B.

Table 1: **Three paired cohorts.** The final column counts measured cell-type DEGs missed by matched bulk, before literature filtering. TNBC: triple-negative breast cancer; ER+: estrogen-receptor-positive; ETP: early T-cell precursor leukemia.

Cohort	Focus population	Case / reference	Participants	Bulk-missed
Breast	Cancer epithelial	TNBC 5 / ER+ 9	14	28
T-ALL	Leukemia blast	ETP 28 / other 9	37	39
Sepsis	Monocyte	Sepsis 46 / healthy 7	53	106

We use the official implementations of CFGen, Bulk2Space, and scSemiProfiler (Palma et al., 2025; Liao et al., 2022; Wang et al., 2024). We run CFGen with cell-type conditioning and supply the same frozen Stage 1 cell counts used by our generator. This controls composition while comparing within-type expression generation. Target donor IDs have no source-trained embedding, so donor conditioning is omitted; Appendix C explains this choice. Bulk2Space uses its native cell allocation. In Breast, it generates no cancer epithelial cells for three of the 14 fixed participants in any seed; it receives zero under the full-cohort rule.

Our direct adaptation of scSemiProfiler resets the source checkpoint for each patient and allows adaptation to that patient's bulk only. Because the source references lack matched measured bulk, it omits the original real-bulk ratio calibration. Together with the limited initial training budget, this defines the comparison reported as *scSemiProfiler-direct*. Two source-trained mean predictors, a bulk-residual ridge regressor and a donor-contrast network, predict patient–cell-type aggregates without cell sampling. Appendix C describes each implementation and its information access.

4.2 RECOVERY OF DISEASE SIGNALS MISSED BY BULK

Ours recovers 10 of 26 important reference genes in Breast and six of 11 in T-ALL, whereas the generative baselines score zero under the full-cohort recovery rule (Table 2). In Sepsis, it recovers nine of 54, tying scSemiProfiler-direct. Across cohorts, the model recovers 25 of 91 cohort–gene references (27.5%), compared with nine (9.9%) for scSemiProfiler-direct: a 2.78-fold difference in observed recovery. CFGen recovers none. Bulk2Space scores zero throughout: assigned for missing Breast outputs, and measured in T-ALL and Sepsis.

Figure 2 follows the same genes from measured bulk to measured and generated cell-type profiles. Several reference genes have bulk effects below the calling threshold but meet the significance, effect-size, and direction criteria after generation. All methods use the same highlighted reference, axes, and thresholds within a cohort.

Table 2: **Recovery of important disease signals missed by bulk.** Entries show full-cohort recovery counts and recall; bold marks maxima and ties. † Assigned zero: Bulk2Space lacks focus-type outputs for three of 14 Breast participants. The T-ALL reference includes four legacy priors with unresolved sources; the literature-only sensitivity is reported in Appendix D.

Method	Breast /26	T-ALL /11	Sepsis /54
CFGen (cell-type)	0 (0.0%)	0 (0.0%)	0 (0.0%)
Bulk2Space	0 (0.0%)†	0 (0.0%)	0 (0.0%)
scSemiProfiler-direct	0 (0.0%)	0 (0.0%)	**9 (16.7%)**
Ours	**10 (38.5%)**	**6 (54.5%)**	**9 (16.7%)**

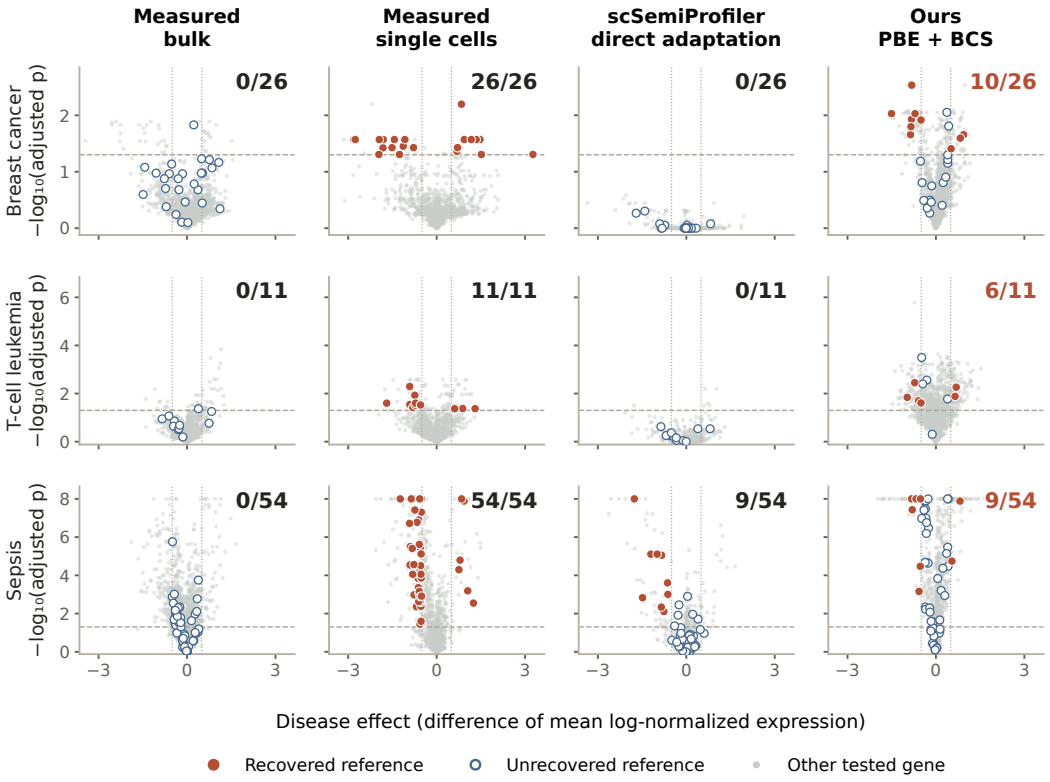


Figure 2: **Recovery of cell-type disease effects.** Filled clay points mark important reference genes recovered with the measured direction; open blue points mark unrecovered references. Grey points are other tested genes. Axes are shared within each cohort, with thresholds at $q = 0.05$ and absolute effect 0.5. Labels give recovered/reference counts. The vertical scale is capped at 8. Measured cells define the reference, so bulk recovery is zero by construction.

The result persists under two complementary reference checks. Without literature filtering, recovery is 10/28, 9/39, and 16/106 on the full bulk-missed sets. Restricting the reference to literature-verified genes gives 10/26, 2/7, and 9/54, or 21/87 overall versus 9/87 for scSemiProfiler-direct. Thus the Breast and T-ALL gains remain when unresolved priors are removed. Ours also exceeds the bulk-residual mean predictor in observed important-gene recovery: 10 versus four, six versus one, and nine versus eight (Appendix C). These are observed improvements in cohort-level disease-signal recovery, not a claim that stochastic sampling is necessary or that every count difference is statistically significant.

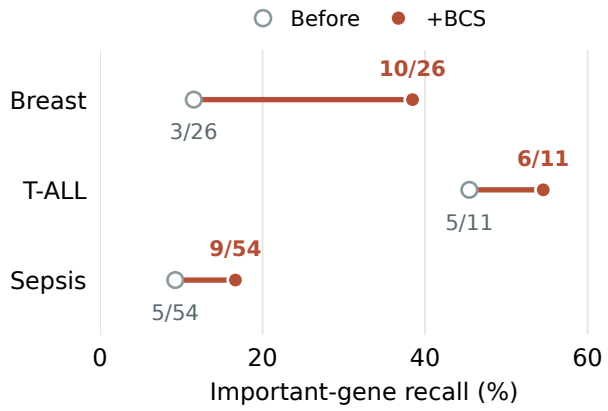


Figure 3: **BCS ablation.** Important-gene recall before and after BCS, with the same parent generator. Labels show recovered/reference counts.

4.3 Contribution of Bulk-Conditioned Shift

To isolate BCS, we apply it to the same parent generator outputs within each cohort. Important-gene recovery increases from three to 10 in Breast, five to six in T-ALL, and five to nine in Sepsis (Figure 3). Recovery on the full bulk-missed reference also increases in every cohort. This matched comparison supports the contribution of the post-generation shift to the recovery endpoint in these cohorts; it does not isolate the effect of PBE.

The continuous effect metrics give a more qualified picture (Table 3). Rank correlation with measured effects improves in all three cohorts, but mean squared error decreases only in Breast. BCS therefore improves thresholded recovery without uniformly reducing disease-effect estimation error. The comparison with official CFGen is complementary: it measures performance against an external generator, whereas this ablation holds the parent model fixed. These metrics answer different questions: recovery measures whether a disease-relevant effect crosses the specified evidence threshold with the correct direction, while mean squared error averages discrepancies over the full tested family. Improved recovery can therefore coexist with unchanged or slightly higher effect-estimation error across the tested genes.

Table 3: **Additional BCS metrics (before → after BCS).** Error and correlation use the fixed tested families; bulk-missed hits include genes outside the important reference.

Cohort	Effect MSE ↓	Spearman ↑	Bulk-missed hits
Breast	$0.1976 \rightarrow 0.1730$	$0.554 \rightarrow 0.585$	$3/28 \rightarrow 10/28$
T-ALL	$0.0983 \rightarrow 0.0999$	$0.472 \rightarrow 0.488$	$8/39 \rightarrow 9/39$
Sepsis	$0.0756 \rightarrow 0.0786$	$0.444 \rightarrow 0.445$	$8/106 \rightarrow 16/106$

4.4 Contribution of the Agentic Workflow

We compare legacy retrieval (W0) with enhanced retrieval (W1), which adds contrast-specific supplement extraction, normalized directions, source locations, and gene-identity and source-family checks (Figure 5). Both lists are scored against the same full-axis measured-cell reference, independently of modeled-panel recovery.

W1 increases matched disease priors from 10 to 36 in Breast, seven to 60 in T-ALL, and three to 13 in Sepsis (Figure 4). Equally sized, expression-matched random lists average 8.72, 9.95, and 1.59 matches across 10,000 draws. These results support targeted retrieval beyond list-size effects. The workflow establishes evaluable cohort questions and traceable gene evidence; generating more cells cannot repair incorrect patient pairings or missing biological references.

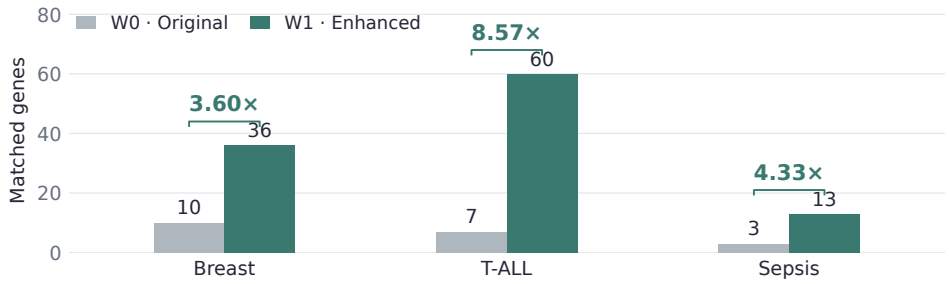


Figure 4: **Disease-prior retrieval.** W0 is the original workflow and W1 is the enhanced workflow. Bars count disease priors matching the same fixed measured-cell reference. Labels above each pair show the W1/W0 ratio of matches. These are retrieval gains, not changes in statistical testing power.

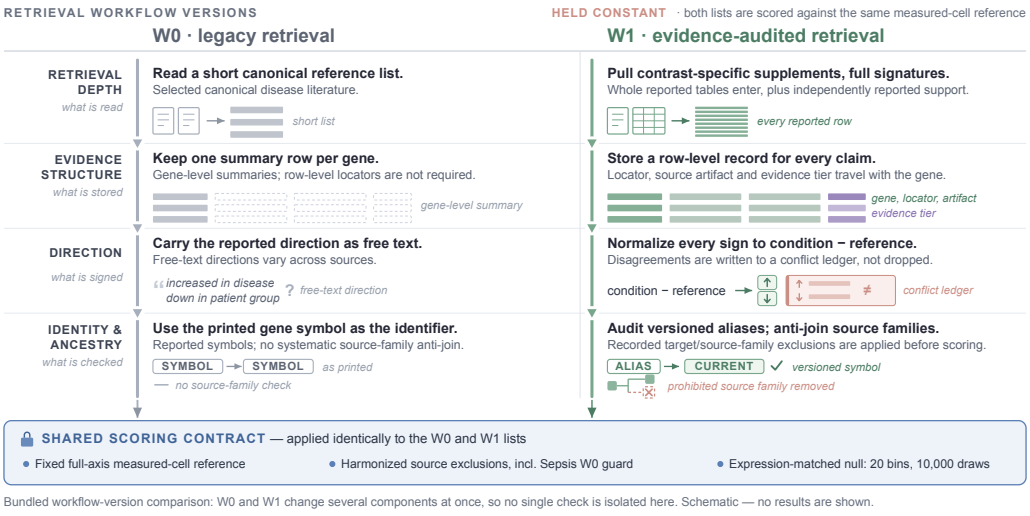


Figure 5: **Retrieval workflow comparison.** W1 expands extraction and adds direction, identity, and ancestry checks. Both lists share the measured-cell reference and source-exclusion rules. Several components change together, so individual causal contributions are not isolated.

This version-level comparison measures retrieval yield, not statistical power or the causal contribution of individual agents: W1 changes several steps and list sizes together. The package supports offline evaluation replay; independent retrieval and evaluation on new cohorts remain untested.

5 Conclusion

We connect bulk-to-single-cell generation to disease analysis through two contributions: an agentic benchmark with traceable evidence and participant-level scoring, and a patient-conditioned generator combining bulk context with a source-fitted expression shift. Across three paired cohorts, Ours has higher full-cohort recovery scores than the executed generative baselines in Breast and T-ALL and ties scSemiProfiler-direct in Sepsis. Matched ablations support BCS; a separate comparison supports enhanced evidence retrieval.

These results support hypothesis generation from existing bulk cohorts, with measured cells as the evaluation reference. Bulk does not uniquely determine the original cells: aggregate recovery establishes neither faithful cell-to-cell variation, calibrated uncertainty, nor the necessity of stochastic generation. The retrospective study is limited by cohort size, reference coverage, and baseline adaptations. Broader validation requires new paired cohorts and independently operated retrieval.

REPRODUCIBILITY STATEMENT

Section 3.2 and Algorithm 1 specify reference construction and patient-level scoring. Appendices A–D document model settings, source and target accessions, participant separation, comparator adaptations, and evidence eligibility. They also distinguish deterministic evaluation replay from live retrieval and identify missing historical execution records.

AI USE STATEMENT

Generative AI assistants supported research planning, method and evaluation-code implementation, experiment orchestration, result interpretation, literature and supplement retrieval, manuscript editing, and figure construction. The agentic benchmark uses AI-assisted cohort and evidence judgments, followed by deterministic checks of participant/gene axes, eligibility records, and recovery counts. Reported metrics and scientific plots derive from saved numerical outputs; evidence records retain source locators and distinguish verified from unresolved support. These checks do not independently validate every biological interpretation. Complete historical prompts, tool/model versions, and human interventions were not recorded. The authors retain responsibility for the manuscript and its claims.

ETHICS STATEMENT

This study uses existing transcriptomic research data. Generated profiles are computational predictions, not patient observations or diagnostic evidence. Any data release must comply with the source cohorts' access, consent, and ethics requirements.

REFERENCES

B. Y. Besecker et al. A comparison of zinc metabolism, inflammation, and disease severity in critically ill infected and noninfected adults early after intensive care unit admission. *American Journal of Clinical Nutrition*, 2011. doi: 10.3945/ajcn.110.008417. URL https://pubmed.ncbi.nlm.nih.gov/21525204/.

Tinyi Chu, Zhong Wang, Dana Pe'er, and Charles G. Danko. Cell type and gene expression deconvolution with BayesPrism enables bayesian integrative analysis across bulk and single-cell RNA sequencing in oncology. *Nature Cancer*, 3:505–517, 2022. URL https://doi.org/10.1038/s43018-022-00356-3.

Natalie R. Davidson, Fan Zhang, and Casey S. Greene. BuDDI: Bulk deconvolution with domain invariance to predict cell-type-specific perturbations from bulk. *PLOS Computational Biology*, 21 (1):e1012742, 2025. URL https://doi.org/10.1371/journal.pcbi.1012742.

Allegra G. Hawkins, Joshua A. Shapiro, Stephanie J. Spielman, et al. The Single-Cell Pediatric Cancer Atlas: Data portal and open-source tools for single-cell transcriptomics of pediatric tumors. *Cell Genomics*, 6(7):101283, 2026. doi: 10.1016/j.xgen.2026.101283. URL https://doi.org/10.1016/j.xgen.2026.101283.

Katherine A. Hoadley, Christina Yau, Toshinori Hinoue, et al. Cell-of-origin patterns dominate the molecular classification of 10,000 tumors from 33 types of cancer. *Cell*, 173(2):291–304.e6, 2018. doi: 10.1016/j.cell.2018.03.022. URL https://doi.org/10.1016/j.cell.2018.03.022.

Alexander Lachmann, Denis Torre, Alexandra B. Keenan, et al. Massive mining of publicly available RNA-seq data from human and mouse. *Nature Communications*, 9:1366, 2018. doi: 10.1038/s41467-018-03751-6. URL https://doi.org/10.1038/s41467-018-03751-6.

Jie Liao, Jingyang Qian, Yin Fang, et al. De novo analysis of bulk RNA-seq data at spatially resolved single-cell resolution. *Nature Communications*, 13:6498, 2022. URL https://doi.org/10.1038/s41467-022-34271-z.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2210.02747.

Erpai Luo, Minsheng Hao, Lei Wei, and Xuegong Zhang. scDiffusion: conditional generation of high-quality single-cell data using diffusion model. *Bioinformatics*, 40(9):btae518, 2024. URL https://doi.org/10.1093/bioinformatics/btae518.

Aaron M. Newman, Chloé B. Steen, Chih Long Liu, et al. Determining cell type abundance and expression from bulk tissues with digital cytometry. *Nature Biotechnology*, 37: 773–782, 2019. doi: 10.1038/s41587-019-0114-2. URL https://doi.org/10.1038/s41587-019-0114-2.

Alessandro Palma, Till Richter, Hanyi Zhang, Manuel Lubetzki, Alexander Tong, Andrea Dittadi, and Fabian J. Theis. Multi-modal and multi-attribute generation of single cells with CFGen. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=3MnMGLctKb.

Andrea Rubbi, Andrea Giuseppe Di Francesco, Mohammad Lotfollahi, and Pietro Liò. A standardized framework for evaluating gene expression generative models. *arXiv preprint arXiv:2603.11244*, 2026. URL https://arxiv.org/abs/2603.11244.

Single-Cell Pediatric Cancer Atlas. Single cell profiling of early T cell precursor acute lymphoblastic leukemia: SCPCP000003. ScPCA Portal, project dataset, n.d. URL https://scpca.alexslemonade.org/projects/SCPCP000003. Accessed September 19, 2026.

Jordan W. Squair, Matthieu Gautier, Claudia Kathe, et al. Confronting false discoveries in single-cell differential expression. *Nature Communications*, 12:5692, 2021. URL https://doi.org/10.1038/s41467-021-25960-2.

Deepak Verma et al. Decoding the genetic symphony: Profiling protein-coding and long noncoding rna expression in t-acute lymphoblastic leukemia for clinical insights. *PNAS Nexus*, 3(2): pgae011, 2024. doi: 10.1093/pnasnexus/pgae011. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC10847906/.

Jiebiao Wang, Kathryn Roeder, and Bernie Devlin. Bayesian estimation of cell type-specific gene expression with prior derived from single-cell data. *Genome Research*, 31:1807–1818, 2021. doi: 10.1101/gr.268722.120. URL https://doi.org/10.1101/gr.268722.120.

Jingtao Wang, Gregory J. Fonseca, and Jun Ding. scSemiProfiler: Advancing large-scale single-cell studies through semi-profiling with deep generative models and active learning. *Nature Communications*, 15:5989, 2024. URL https://doi.org/10.1038/s41467-024-50150-1.

Christopher Wilks, Shijie C. Zheng, Feng Yong Chen, et al. recount3: summaries and queries for large-scale RNA-seq expression and splicing. *Genome Biology*, 22:323, 2021. doi: 10.1186/s13059-021-02533-6. URL https://doi.org/10.1186/s13059-021-02533-6.

Sunny Z. Wu, Ghamdan Al-Eryani, Daniel L. Roden, et al. A single-cell and spatially resolved atlas of human breast cancers. *Nature Genetics*, 53:1334–1347, 2021. doi: 10.1038/s41588-021-00911-1. URL https://doi.org/10.1038/s41588-021-00911-1.

Jiedan Xiao and Andreas Raue. bulk2scdiff: A pseudobulk-conditioned diffusion model for bulk-to-single-cell rna-seq generation. *bioRxiv*, 2026. doi: 10.64898/2026.08.20.745960. URL https://www.biorxiv.org/content/10.64898/2026.08.20.745960v1.

Q. Ye, F. Xu, and J. Cao. Multiomics analyses unveil immune landscape in pediatric and adult sepsis [bulk RNA-seq]. NCBI Gene Expression Omnibus, GSE279448, 2025a. URL https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE279448. Dataset; released September 28, 2025.

Q. Ye, F. Xu, and J. Cao. Multiomics analyses unveil immune landscape in pediatric and adult sepsis [single-cell RNA-seq]. NCBI Gene Expression Omnibus, GSE279452, 2025b. URL https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE279452. Dataset; released September 28, 2025.