TASK-AWARE BENCHMARKING OF BATCH CORRECTION METHODS FOR SINGLE-CELL RNA SEQUENCING ATLAS CONSTRUCTION IN TRIPLE-NEGATIVE BREAST CANCER

Peter Scheible
Computer Science
Old Dominion University
Norfolk, VA, United States
psche004@odu.edu

Jing He, Ph.D.
Computer Science
Old Dominion University
Norfolk, VA, United States
jhe@cs.odu.edu

Amy H. Tang, Ph.D.
Biomedical and Translational Sciences
Macon & Joan Brock Virginia Health Sciences
Old Dominion University
Norfolk, VA, United States
TangAH@odu.edu

Jiangwen Sun, Ph.D.
The Corresponding Author
Computer Science
Old Dominion University
Norfolk, VA, United States
jsun@cs.odu.edu

***Abstract*—Single-cell RNA sequencing (scRNA-seq) offers unprecedented resolution for characterizing the cellular landscape of triple-negative breast cancer (TNBC), but constructing multi-cohort atlases requires reliable batch correction. While numerous integration methods exist, no systematic evaluation has addressed how method choice affects TNBC-specific downstream tasks. We assembled a curated compendium of approximately 1.14 million cells from seven publicly available TNBC datasets and benchmarked seven batch correction methods—scVI, Harmony, Seurat v5 RPCA, scCRAFT, sysVI, BA-scVI, and fastMNN—using a three-domain evaluation framework spanning batch correction quality, biological conservation, and TNBC-specific task performance. TNBC-specific evaluation assessed rare cell type preservation and Lehmann molecular subtype recovery, metrics absent from prior general-purpose benchmarks. No single method ranked in the top three across all three domains, revealing substantial method-by-domain interactions. fastMNN achieved the highest overall score (0.477) via harmonic mean aggregation, and Harmony was the most balanced performer (0.469), ranking first for both batch correction and biological conservation. Notably, uncorrected PCA outperformed most methods on rare cell type preservation, underscoring the overcorrection risk of aggressive batch removal. These results provide TNBC-specific guidance for atlas construction and demonstrate the necessity of task-aware evaluation criteria when selecting integration methods.**

***Index Terms*—RNA sequencing, Triple-negative breast cancer, Batch correction, Benchmarking, Data integration, Single-cell genomics, Bioinformatics, Computational biology**

The authors gratefully acknowledge funding support from the Virginia Space Grant Consortium (VSGC) and the Hampton Roads Biomedical Research Consortium (HRBRC).

I. Introduction

Triple-negative breast cancer (TNBC) is defined by the absence of estrogen receptor (ER), progesterone receptor (PR), and HER2 expression, representing approximately 15–20% of all breast cancer diagnoses [1]. Unlike hormone receptor-positive or HER2-amplified subtypes, TNBC lacks actionable targeted therapies; platinum-based chemotherapy and, increasingly, immunotherapy remain the standard of care [2]. The disease carries the worst prognosis of all breast cancer subtypes, with high rates of early relapse and distant metastasis.

Transcriptomic profiling has revealed that enquoteTNBC is not a single entity. Bulk RNA-seq analysis of 587 samples identified six molecular subtypes—basal-like 1 (BL1), basal-like 2 (BL2), immunomodulatory (IM), mesenchymal (M), mesenchymal stem-like (MSL), and luminal androgen receptor (LAR)—with distinct pathway activation and differential chemotherapy response [1]. Single-cell RNA sequencing (scRNA-seq) has since resolved an even richer landscape of intratumoral heterogeneity, motivating the construction of integrated atlases that span multiple cohorts to characterize cell states, rare subpopulations, and treatment response biology [3]–[5].

Realising such an atlas requires integrating heterogeneous data across patients, sequencing platforms, library chemistries, and tissue processing protocols. Batch effects—systematic technical variation introduced by these differences—can dominate the variance in scRNA-seq data, obscuring biological signals and producing spurious clustering along technical rather than biological axes [6], [7]. Critically, batch correction is not without cost: aggressive correction can distort biologically meaningful variation, as demonstrated by Moir et al., who showed that standard correction pipelines compromise PAM50 biomarker signals across large breast cancer cohorts [8].

This work addresses the foundational methodological question of *how to choose* a batch correction method for TNBC atlas construction. Prior work on atlas integration has focused on building atlases using a single chosen integration method; here we step back and systematically evaluate seven methods before any integration decision is made. Existing benchmarks assess integration quality using generic embedding-space metrics that do not evaluate disease-specific downstream tasks [7]. A recent assessment of over 1,700 single-cell algorithms has explicitly called for living, task-aware benchmarks as a field-wide priority [9], and benchmarking 13 methods across 12 datasets spanning 5 cancer types found that optimal method choice is cancer-type and task-dependent [10]. No analogous task-aware evaluation exists specifically for TNBC atlas construction.

We address this gap with the following contributions:

1) A curated TNBC scRNA-seq compendium assembled from seven published datasets, totalling approximately 1.14 million cells spanning multiple sequencing platforms and clinical contexts (Table I).
2) A three-domain benchmarking pipeline evaluating seven batch correction methods—scVI, Harmony, Seurat v5 RPCA, scCRAFT, sysVI, BA-scVI, and fastMNN—on standard integration metrics (batch mixing and biological conservation) and TNBC-specific metrics (rare cell type preservation and molecular subtype recovery).
3) Evidence that no single method dominates across all three domains, with practical recommendations for TNBC atlas construction.

II. Literature Review

A. scRNA-seq in Breast Cancer Research

Single-cell RNA sequencing has transformed cancer research by enabling high-resolution profiling of individual cells within

tumors. Chung et al. conducted the first comprehensive scRNA-seq study of breast cancer, profiling 515 cells from 11 patients and resolving T cells, B cells, macrophages, and distinct epithelial and stromal subpopulations within individual tumors [3]. For TNBC specifically, Karaayvaz et al. applied scRNA-seq to six primary tumors to uncover subclonal heterogeneity and identify gene expression signatures linked to treatment resistance [4]. Vishnubalaji and Alajez linked cancer cell-state persistence to neoadjuvant chemotherapy resistance, distinguishing transcriptional enquoteextinction from enquotepersistence programmes [11]. At larger scale, the Breast Cancer Single-Cell Atlas catalogued 49 subpopulations across 39,214 cells from 26 primary tumors and 13 cell lines, establishing the breadth of heterogeneity a multi-cohort TNBC atlas must capture [5].

B. Batch Effects and Integration Challenges

Batch effects in scRNA-seq arise from differences in cell dissociation protocols, library preparation kits, sequencing platforms, and data processing pipelines [6]. When uncorrected, these technical effects can dominate dataset variance, producing spurious clustering along platform rather than biological axes and obscuring genuine biological signals [7]. Moir et al. demonstrated concretely that standard correction pipelines can compromise PAM50 breast cancer biomarker signals, illustrating the dual risk of both under-correction (persistent batch artifacts) and over-correction (biological signal loss) [8].

C. Batch Correction Methods

Batch correction methods span four algorithmic families. *Anchor-based* methods identify mutual nearest neighbours (MNN) across batches as cross-dataset cell-type counterparts; fastMNN pioneered this approach [12] and Seurat v5 RPCA extends it with reciprocal PCA for conservative integration [13]. *Variance decomposition* methods adjust embeddings statistically; Harmony iteratively corrects PCA space via soft k-means clustering to minimize batch separation while preserving cluster structure [14]. *Deep learning* methods capture nonlinear batch effects through neural networks: scVI uses a variational autoencoder (VAE) conditioned on batch identity [15]; BA-scVI augments this with adversarial training [16]; sysVI applies a conditional VAE with cycle-consistency regularization [17]; and scCRAFT leverages topological structure for anchor-free integration [18].

D. Existing Benchmarking Frameworks

Standard metrics for evaluating batch correction span embedding geometry and neighbourhood composition. kBET applies chi-squared tests to k-nearest-neighbour neighbourhoods to assess local batch mixing [19]. The local inverse Simpson's index (LISI) measures the effective number of batches (iLISI) or cell types (cLISI) in each cell's neighbourhood. Average silhouette width (ASW) measures global separation in embedding space. The scIB benchmark evaluated 16 methods across 13 integration tasks using these metrics, recommending scANVI, Scanorama, and scVI as top performers for general use [7]. Tran et al. benchmarked 14 methods and recommended Harmony, LIGER, and Seurat for broad applicability [6].

However, these metrics evaluate integration quality in the embedding itself and do not assess performance on downstream biological tasks. Barkmann et al. addressed this gap for cancer research by benchmarking 13 methods across 12 datasets spanning 5 cancer types, finding that Harmony, BBKNN, and fastMNN best support cancer cell-state discovery, and that the optimal choice varies by cancer type [10]. A comprehensive assessment of the single-cell field has explicitly called for task-aware benchmarks as a priority [9]. No analogous evaluation exists for TNBC, where disease-specific metrics such as molecular subtype recovery and rare cell type preservation are critical for clinical relevance.

TABLE I
TNBC SCRNA-SEQ DATASETS IN THE COMPENDIUM. CELL COUNTS ARE POST-QC, PRE-TNBC-FILTER TOTALS.

Study	Cells	Clinical context
Chen et al. [20] GSE161529	388,167	Multi-subtype; normal
Tietscher et al. [21] E-MTAB-10607	129,013	T cell exhaustion; TME
Bassez et al. [22] EGAS00001004809	10,671	Anti-PD-1; V(D)J
Wang et al. [23] GSE248288	24,330	Macrophage-immune axis
Zhang et al. [24] GSE169246	489,490	Paclitaxel ± anti-PD-L1
Liu et al. [25] GSE225600	63,561	Primary + lymph-node mets
Gao et al. [26] GSE148673	32,724	Copy-number subclone

III. METHODOLOGY

Seven batch correction methods were applied to a curated TNBC scRNA-seq compendium and evaluated using a three-domain framework. All code was implemented in Python using Scanpy and scverse libraries; Seurat v5 RPCA and fastMNN were invoked via rpy2.

A. Data Compendium

We assembled a scRNA-seq compendium from seven publicly available datasets encompassing breast cancer and normal breast tissue (Table I), totalling approximately 1.14 million cells, of which approximately 727,000 are retained after TNBC-specific filtering. Datasets were selected to cover biological contexts relevant to TNBC, including normal tissue references, tumor microenvironment profiling, immunotherapy response, and primary–metastasis comparisons. Data span $3'$ and $5'$ library chemistries, 10x Chromium v2 and v3 platforms, and multiple clinical contexts (treatment-naïve, neoadjuvant chemotherapy, anti-PD-1/PD-L1).

B. Preprocessing

Each dataset was processed independently before integration. Cells were retained if they expressed at least 200 genes; genes were retained if detected in at least 3 cells; and cells with mitochondrial read fractions exceeding 20% were removed. Raw counts were preserved prior to normalisation. Library sizes were normalised to 10^4 counts per cell followed by $\log(1 + x)$ transformation. For integration, joint highly variable gene (HVG) selection was performed: for each gene across all seven datasets, the number of datasets flagging it as highly variable was counted, and the top 3,000 genes by vote count were selected. Principal component analysis (PCA) was computed with 50 components on the joint feature set, providing a shared embedding used as input by linear and graph-based correction methods.

C. Batch Correction Methods

Seven methods spanning four algorithmic families were evaluated (Table II). BBKNN [27] was considered but excluded because it outputs a corrected neighbourhood graph rather than a coordinate embedding, making it incompatible with embedding-geometry metrics (ASW, iLISI, cLISI) and rendering fair comparison against the other methods infeasible. The batch variable for all methods was the dataset of origin. Methods operating on raw count matrices (scVI, BA-scVI, Seurat v5 RPCA) received raw counts; methods operating on log-normalised data (scCRAFT, sysVI, fastMNN) received $\log(1 + x)$-transformed values; and Harmony received the joint 50-dimensional PCA embedding.

scVI [15]: A variational autoencoder (VAE) with a zero-inflated negative binomial likelihood that learns a low-dimensional latent representation conditioned on batch identity.

Harmony [14]: Iteratively corrects PCA embeddings via soft k-means clustering, computing per-cluster, per-batch correction factors to minimize batch separation while preserving cluster structure.

Seurat v5 RPCA [13]: Projects datasets into a shared PCA subspace, identifies mutual nearest-neighbour anchors between dataset

TABLE II
SUMMARY OF THE SEVEN BATCH CORRECTION METHODS EVALUATED. DNN: USES A DEEP NEURAL NETWORK. EXPR.: CAN OUTPUT FULL CORRECTED EXPRESSION VALUES.

Method	Family	Input	DNN	Expr.
scVI	VAE + ZINB	Raw	✓	✓
Harmony	Iterative PCA	PCA	–	–
Seurat v5 RPCA	Anchor-based RPCA	Raw	–	–
scCRAFT	Topology-guided AE	Log-norm	✓	–
sysVI	cVAE + VampPrior	Log-norm	✓	✓
BA-scVI	Adversarial VAE	Raw	✓	✓
fastMNN	Mutual nearest nbrs	Log-norm	–	–

pairs, and uses anchor correspondences to align embeddings conservatively.

scCRAFT [18]: An autoencoder framework that separates cell-type signals from batch effects by leveraging partially characterized topological structure as a supervision signal, enabling anchor-free integration.

sysVI [17]: A conditional VAE employing a VampPrior and cycle-consistency regularization, designed for datasets with substantial batch effects.

BA-scVI [16]: Augments the scVI framework with an adversarial discriminator network that penalises batch predictability from the latent code.

fastMNN [12]: Identifies mutual nearest neighbours between dataset pairs in PCA space, computes batch-correction vectors from MNN pairs, and applies them in a reduced-rank approximation.

D. Cell Type Annotation

Automated cell type annotation was performed using CellTypist [28] with the pre-trained `Cells_Adult_Breast.pkl` model, which provides fine-grained labels for adult breast tissue cell types. Annotation was performed in chunks of 50,000 cells, with raw counts normalised and log-transformed before classification. Majority voting was disabled to preserve fine-grained assignments and avoid artificially homogenising rare populations. The 58 fine-grained CellTypist labels were mapped to 10 coarse categories: T cell, NK cell, B cell, myeloid, dendritic cell, mast cell, epithelial, stromal, endothelial, and lymphatic.

E. Three-Domain Evaluation Framework

Integration quality was assessed using a three-domain framework following the single-cell integration benchmarking (scIB) methodology [7]. All metrics were computed on a stratified subsample of 50,000 cells (stratified by dataset, seed $= 42$). The batch variable was the dataset of origin; the label variable was the coarse cell type annotation.

Domain 1 – Batch Correction: Three metrics assess technical variation removal. kBET [19] applies a χ^2 test to k-nearest-neighbour neighbourhoods; the reported score is $1 -$ mean rejection rate, with higher values indicating better local batch mixing. $\text{ASW}_{\text{batch}}$ [29] measures the global degree to which cells from different batches are mixed within cell type clusters. iLISI [14] measures the effective number of batches represented in each cell's local neighbourhood, scaled to $[0, 1]$. These three metrics are complementary: kBET and iLISI capture local neighbourhood-level mixing while $\text{ASW}_{\text{batch}}$ summarises global geometry.

Domain 2 – Biological Conservation: Five metrics assess preservation of biological structure. ARI and NMI compare Leiden cluster assignments to cell type labels across resolutions $[0.1\text{–}2.0]$. $\text{ASW}_{\text{celltype}}$ measures how tightly cells of the same type cluster in embedding space. cLISI measures cell-type purity within local neighbourhoods. Graph connectivity assesses whether cells of the same type form a connected component in the k-nearest-neighbour graph.

Domain 3 – TNBC-Specific: Two metrics evaluate preservation of clinically relevant structure. *Rare cell type sensitivity* identifies populations below 1% frequency and three marker-based populations ($CD44^{+}CD24^{-}$ stem-like cells, cycling tumor cells, and regulatory T cells) and measures cluster-based preservation via mean F1 across populations. *Molecular subtype recovery* scores each epithelial cell against the six Lehmann subtype signatures (BL1, BL2, M, MSL, LAR, IM) [1] and quantifies recovery as the ARI between subtype assignments and Leiden cluster assignments.

F. Overall Scoring and Aggregation

Methods were ranked using a harmonic mean across the three domain scores:

$$\text{Overall} = \frac{3}{\frac{1}{\text{Batch}} + \frac{1}{\text{Bio}} + \frac{1}{\text{TNBC}}}$$

where Batch is the mean of $\text{ASW}_{\text{batch}}$ and kBET; Bio is the mean of ARI, NMI, $\text{ASW}_{\text{celltype}}$, and graph connectivity; and TNBC is the mean of rare cell F1 and subtype cluster ARI.

The harmonic mean was chosen because batch correction and biological conservation represent competing objectives analogous to precision and recall: a method that excels at one while neglecting the other should not receive a high composite score. The harmonic mean penalises such imbalance more heavily than arithmetic or geometric means, consistent with the F-score formulation. Consequently, a method's weakest domain dominates its overall rank, rewarding balanced performance across all three evaluation dimensions.

IV. Results

A. Batch Correction Metrics

All seven methods improved upon uncorrected PCA across the batch-mixing metrics (Fig. 1). Domain-level ranking (composite of $\text{ASW}_{\text{batch}}$ and kBET) placed Harmony first (0.660), followed by scCRAFT (0.659), BA-scVI, Seurat v5 RPCA, fastMNN, scVI, and sysVI. $\text{ASW}_{\text{batch}}$ was uniformly high across all methods and the PCA baseline (0.833–0.904), indicating that the joint PCA embedding already reduces gross batch separation in silhouette terms, leaving limited room for further differentiation on this metric alone [30]. kBET proved the most discriminative metric, spanning 0.229 (PCA) to 0.436 (Harmony), and drove the majority of domain-level separation. iLISI values were low for all methods (0.007–0.086), as expected on large multi-batch datasets where the LISI scaling formula compresses scores toward zero; concordant $\text{ASW}_{\text{batch}}$ and kBET results confirm genuine batch mixing nonetheless.

B. Biological Conservation Metrics

Biological conservation scores revealed meaningful differences across methods (Fig. 1). Domain-level ranking placed Harmony first, followed by fastMNN, sysVI, scCRAFT, scVI, Seurat v5 RPCA, PCA, and BA-scVI. ARI was the most discriminative biological metric, ranging from 0.475 (BA-scVI) to 0.649 (Harmony); NMI followed a concordant pattern. cLISI was near-perfect for all methods (~ 0.99), confirming that local neighbourhoods remained cell-type-pure regardless of correction strategy.

The batch-versus-biological conservation trade-off (Fig. 2) reveals notable asymmetries. BA-scVI ranked third for batch mixing but last for biological conservation, consistent with the overcorrection phenomenon described by Moir et al., in which aggressive batch removal erodes genuine biological variation [8]. Harmony occupied the Pareto-optimal position on the batch-versus-bio trade-off: first for batch mixing and first for biological conservation simultaneously (Fig. 2).

C. TNBC-Specific Metrics

TNBC-specific scores were substantially lower than those of the other two domains,

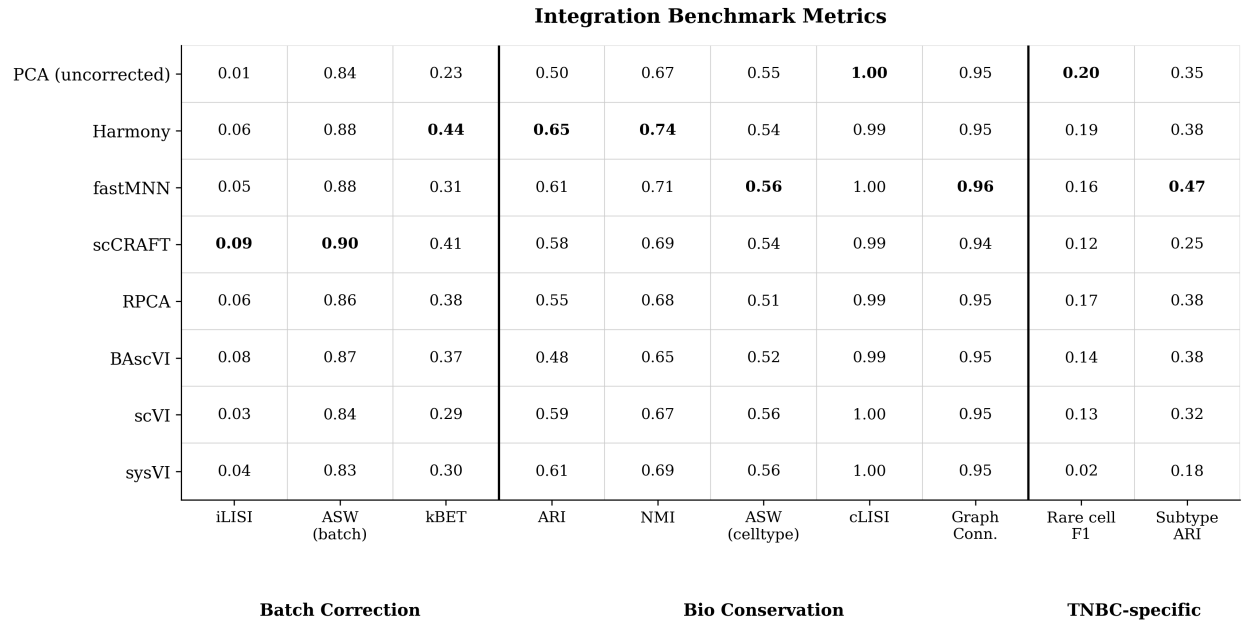

Fig. 1. Integration benchmarking metrics for all seven batch correction methods and the uncorrected PCA baseline. Batch correction metrics: iLISI, $\text{ASW}_{\text{batch}}$, kBET. Biological conservation metrics: ARI, NMI, $\text{ASW}_{\text{celltype}}$, cLISI, graph connectivity. TNBC-specific metrics: rare cell F1, subtype cluster ARI. Bold indicates best value per column.

reflecting the difficulty of preserving fine-grained tumor structure during batch correction. Domain-level ranking placed fastMNN first, followed by Harmony, PCA, Seurat v5 RPCA, BA-scVI, scVI, scCRAFT, and sysVI. Notably, uncorrected PCA ranked third, indicating that several correction methods actively degrade TNBC-specific structure.

Subtype recovery ARI varied widely, from 0.184 (sysVI) to 0.468 (fastMNN). LAR and IM subtypes were well-recovered across all methods, whereas BL1 and MSL were poorly recovered; Harmony achieved the best BL1 recovery (subtype ARI $= 0.380$). BL2 and M subtypes had insufficient representation in the filtered compendium for reliable recovery across all methods, aligning with the heterogeneous expression programmes described by Lehmann et al. [1].

Rare cell type F1 scores were low across the board (0.023–0.196), with PCA achieving the highest score (0.196) and Harmony a close second (0.193). sysVI produced a near-zero rare F1 (0.023), indicating that its latent space collapses rare populations into dominant cell types. That PCA outperforms most correction methods on rare cell preservation underscores the overcorrection risk: batch correction can systematically displace small, biologically distinct populations toward larger clusters.

D. Aggregated Rankings

Table III summarises the composite rankings across all three evaluation domains. fastMNN achieved the highest overall score (0.477), with Harmony a close second (0.469). scCRAFT, despite ranking second on batch correction, dropped to seventh overall because its weak TNBC score (0.189) was heavily penalised by the harmonic mean aggregation. sysVI ranked last (0.234), driven almost entirely by its near-zero rare cell F1 score.

The central finding is that no method ranked in the top three across all three domains. scCRAFT placed second for batch but seventh for TNBC-specific metrics; fastMNN placed first for TNBC but fifth for batch; sysVI placed third for biological conservation but last for TNBC. Harmony was the most balanced

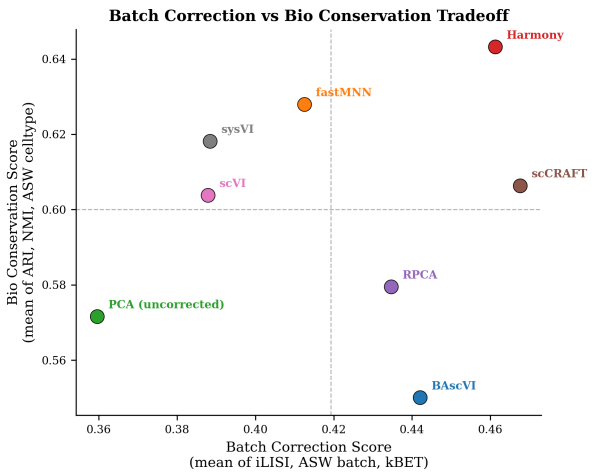

Fig. 2. Batch correction versus biological conservation trade-off. Each point represents one method; the dashed line marks the uncorrected PCA baseline. Harmony occupies the Pareto-optimal position (upper right), achieving the best performance on both axes simultaneously.

TABLE III
COMPOSITE RANKINGS OF THE SEVEN BATCH CORRECTION METHODS AND THE UNCORRECTED PCA BASELINE. COLUMNS SHOW DOMAIN MEANS; OVERALL IS THE HARMONIC MEAN OF THE THREE DOMAIN SCORES.

Method	Batch	Bio	TNBC	Overall
fastMNN	0.594	0.710	0.313	0.477
Harmony	0.660	0.719	0.286	0.469
Seurat v5 RPCA	0.622	0.672	0.271	0.442
BA-scVI	0.623	0.651	0.260	0.429
PCA (uncorrected)	0.536	0.666	0.273	0.427
scVI	0.568	0.690	0.225	0.392
scCRAFT	0.659	0.691	0.189	0.363
sysVI	0.564	0.702	0.104	0.234

method, ranking first, first, and second across batch, bio, and TNBC domains respectively, while fastMNN achieved the highest overall score by virtue of its leading TNBC performance. This domain-dependent pattern mirrors the findings of Barkmann et al., who showed that the optimal integration method varies by cancer type [10]; here we show an analogous dependence on evaluation domain within a single cancer type.

Fig. 3 illustrates the top-performing methods visually. Harmony and fastMNN produce well-mixed, cell-type-coherent embeddings relative to uncorrected PCA, shifting from dataset-driven clustering to cell-type-driven clustering while preserving the identity of distinct populations.

V. DISCUSSION

The central finding of this work is that batch correction method performance is strongly domain-dependent: no single method ranked in the top three across all three evaluation domains. This result has direct practical implications—choosing a method based solely on one criterion (e.g., batch mixing) will yield suboptimal results for TNBC-specific analyses that depend on rare cell type preservation or molecular subtype recovery.

The harmonic mean aggregation amplifies this finding. Because the overall score is dominated by a method's weakest domain, methods with extreme specialisation—such as scCRAFT (strong batch, weak TNBC) or sysVI (strong bio conservation, near-zero rare F1)—score poorly overall despite strong performance in at least one area. This mirrors the precision-recall trade-off in information retrieval: a method that achieves perfect batch mixing at the cost of biological fidelity should not be considered a strong integrator.

Harmony's balanced profile across all three domains (ranked 1st, 1st, 2nd) suggests it is well-suited as a default choice for TNBC atlas construction where preserving diverse cellular populations is important. fastMNN's leading overall score is driven primarily by its superior TNBC-specific performance—particularly subtype ARI (0.468)—which is clinically relevant for downstream analyses of molecular heterogeneity. For workflows prioritizing maximum batch mixing, such as pooling heterogeneous cohorts with very different sequencing platforms, Harmony and scCRAFT are the strongest performers on that single dimension.

The overcorrection phenomenon revealed by this benchmark warrants emphasis. Uncorrected PCA outperformed all seven methods on rare cell type preservation, achieving the

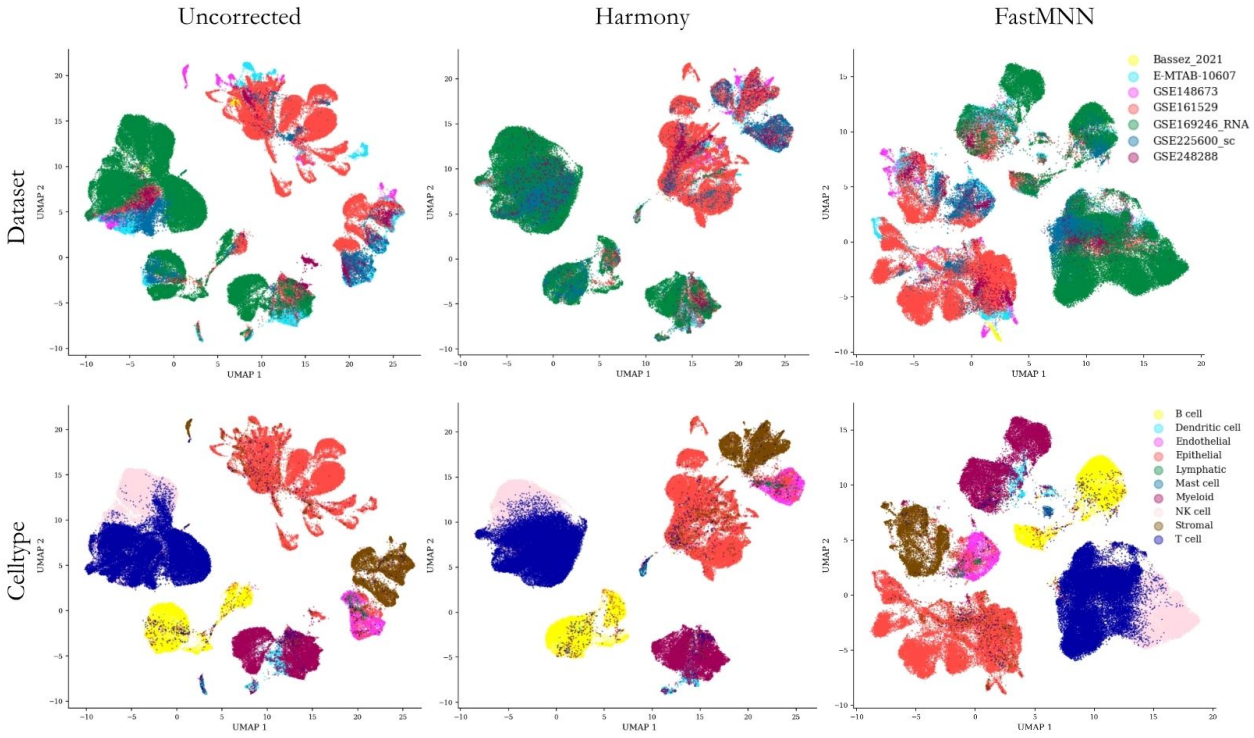

Fig. 3. UMAP projections for the uncorrected PCA baseline and the two top-ranked methods, Harmony and fastMNN. Top row: cells coloured by dataset of origin; bottom row: cells coloured by annotated cell type. Harmony and fastMNN produce substantially better interleaving of datasets while preserving cell-type structure relative to the uncorrected PCA embedding.

highest F1 (0.196) against sysVI's near-zero (0.023). This is not merely a failure of specific methods—it reflects a fundamental tension in batch correction: the same transformations that merge dataset-specific technical clusters can merge biologically distinct rare populations that happen to share local neighbourhood structure with dominant cell types. Practitioners building TNBC atlases should explicitly monitor rare population preservation, particularly for stem-like, cycling, and regulatory T cell populations that are biologically and clinically significant.

This framework parallels challenges in NASA's GeneLab program, which integrates multi-omics datasets from astronaut cohorts exposed to microgravity and radiation. Just as TNBC studies pool heterogeneous clinical samples that introduce technical variation, space biology integrates data across missions with different experimental conditions. The task-aware evaluation approach developed here—measuring method performance against specific downstream analysis goals rather than generic embedding metrics—applies directly to selecting integration strategies for NASA multi-mission datasets.

Limitations of this study include its restriction to publicly available TNBC datasets, which may not fully represent clinical cohort diversity. The molecular subtype labels from bulk-level classifiers [1] serve as proxy ground truth; single-cell subtype labels would provide a more direct evaluation target. Additionally, differential expression reliability and trajectory inference fidelity were not evaluated, and would provide complementary views of integration quality. Future work will incorporate additional TNBC datasets, extended task evaluation, and automated method selection guided

by dataset characteristics.

VI. Conclusion

We presented a task-aware benchmarking framework for evaluating batch correction methods in the context of TNBC scRNA-seq atlas construction. By evaluating seven methods not only on generic integration-quality metrics but also on TNBC-specific metrics for rare cell type preservation and molecular subtype recovery, we found that method performance is strongly domain-dependent: no single method ranked in the top three across all three evaluation domains, and overall ranking was dominated by each method's weakest domain through the harmonic mean aggregation.

For TNBC atlas construction, we recommend fastMNN as the default integration method based on its highest overall score (0.477) and leading TNBC-specific performance. Harmony (overall 0.469) is an equally strong alternative, particularly when balanced performance across all three domains is required. When maximum batch mixing is the priority, Harmony and scCRAFT offer the strongest batch correction. Practitioners should monitor rare cell type preservation regardless of method choice, as uncorrected PCA outperformed all seven methods on this metric.

This work addresses a gap in TNBC bioinformatics infrastructure by providing cancer-type-specific method guidance grounded in downstream task performance. The three-domain evaluation framework is extensible to other cancer atlases and to integration challenges in space biology, where task-aware method selection is equally critical for reliable multi-dataset analyses. Future directions include expanding the compendium to additional TNBC datasets, incorporating differential expression and trajectory inference evaluation, and automating method selection based on dataset characteristics.

References

[1] B. D. Lehmann, J. A. Bauer, X. Chen, M. E. Sanders, A. B. Chakravarthy, Y. Shyr, and J. A. Pietenpol, "Identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies," *Journal of Clinical Investigation*, vol. 121, no. 7, pp. 2750–2767, 2011.

[2] C. Denkert and S. Loibl, "Response-based molecular subtyping—emergence of the third generation of breast cancer subtypes," *Cancer Cell*, vol. 40, no. 6, pp. 592–594, 2022.

[3] W. Chung, H. H. Eum, H.-O. Lee, K.-M. Lee, H.-B. Lee, K.-T. Kim, H. S. Ryu, S. Kim, J. E. Lee, Y. H. Park, Z. Kan, W. Han, and W.-Y. Park, "Single-cell RNA-seq enables comprehensive tumour and immune cell profiling in primary breast cancer," *Nature Communications*, vol. 8, p. 15081, 2017.

[4] M. Karaayvaz, S. Cristea, S. M. Gillespie, A. P. Patel, R. Mylvaganam, C. C. Luo, M. C. Specht, B. E. Bernstein, F. Michor, and L. W. Ellisen, "Unravelling subclonal heterogeneity and aggressive disease states in TNBC through single-cell RNA-seq," *Nature Communications*, vol. 9, p. 3588, 2018.

[5] A. Dave, D. Charytonowicz, N. J. Francoeur, M. Beaumont, K. G. Beaumont, B. Vieth, N. Bhatt, N. V. Bhanu, I. Ramos, R. Sebra *et al.*, "The breast cancer single-cell atlas: defining cellular heterogeneity within model cell lines and primary tumors to inform disease subtype, stemness, and treatment options," *Cell Oncology*, vol. 46, pp. 603–628, 2023.

[6] H. T. N. Tran, K. S. Ang, M. Chevrier, X. Zhang, N. Y. S. Lee, M. Goh, and J. Chen, "A benchmark of batch-effect correction methods for single-cell RNA sequencing data," *Genome Biology*, vol. 21, no. 1, p. 12, 2020.

[7] M. D. Luecken, M. Büttner, K. Chaichoompu, A. Danese, M. Interlandi, M. F. Müller, D. C. Strobl, L. Zappia, M. Dugas, and F. J. Theis, "Benchmarking atlas-level data integration in single-cell genomics," *Nature Methods*, vol. 19, no. 1, pp. 41–50, 2022.

[8] N. Moir, D. A. Pearce, S. P. Langdon, and T. I. Simpson, "The significance of molecular heterogeneity in breast cancer batch correction and dataset integration," *Breast Cancer Research*, vol. 27, 2025.

[9] M. D. Luecken, S. Gigante, D. B. Burkhardt *et al.*, "Defining and benchmarking open problems in single-cell analysis," *Research Square*, 2024.

[10] F. Barkmann, J. Yates, P. Czyż *et al.*, "CanSig benchmarks methods for reproducible cancer cell state discovery from single-cell transcriptomic data," *Cancer Research*, vol. 86, no. 4, pp. 873–888, 2026.

[11] R. Vishnubalaji and N. M. Alajez, "Transcriptional landscape associated with TNBC resistance to neoadjuvant chemotherapy revealed by single-cell RNA-seq," *Molecular Therapy – Oncolytics*, vol. 23, pp. 151–162, 2021.

[12] L. Haghverdi, A. T. L. Lun, M. D. Morgan, and J. C. Marioni, "Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbours," *Nature Biotechnology*, vol. 36, no. 5, pp. 421–427, 2018.

[13] Y. Hao, T. Stuart, M. H. Kowalski, S. Choudhary, P. Hoffman, A. Hartman, A. Srivastava, G. Molla, S. Madad, C. Fernandez-Granda, and R. Satija, "Dictionary learning for integrative, multimodal and scal-

able single-cell analysis," *Nature Biotechnology*, vol. 42, no. 2, pp. 293–304, 2024.

[14] I. Korsunsky, N. Millard, J. Fan, K. Slowikowski, F. Zhang, K. Wei, Y. Baglaenko, M. Brenner, P.-r. Loh, and S. Raychaudhuri, "Fast, sensitive and accurate integration of single-cell data with Harmony," *Nature Methods*, vol. 16, no. 12, pp. 1289–1296, 2019.

[15] R. Lopez, J. Regier, M. B. Cole, M. I. Jordan, and N. Yosef, "Deep generative modeling for single-cell transcriptomics," *Nature Methods*, vol. 15, no. 12, pp. 1053–1058, 2018.

[16] J. J. Díaz-Mejía, E. Williams, B. Innes, O. Focsa, D. Mendonca, S. Singh, A. Nixon, R. Schuster, M. B. Buechler, B. Hinz, and S. Cooper, "Adversarial learning enables unbiased organism-wide cross-species alignment of single-cell RNA data at scale," bioRxiv, 2024, BA-scVI; preprint.

[17] K. Hrovatin, A. A. Moinfar, L. Zappia, and F. J. Theis, "Integrating single-cell RNA-seq datasets with substantial batch effects," *bioRxiv*, 2024, sysVI; preprint.

[18] C. He, P. Filippidis, S. H. Kleinstein, and L. Guan, "Partially characterized topology guides reliable anchor-free scRNA integration," *Communications Biology*, vol. 8, p. 561, 2025, scCRAFT.

[19] M. Büttner, Z. Miao, F. A. Wolf, S. A. Teichmann, and F. J. Theis, "A test metric for assessing single-cell RNA-seq batch correction," *Nature Methods*, vol. 16, no. 1, pp. 43–49, 2019.

[20] Y. Chen, B. Pal, G. J. Lindeman, J. E. Visvader, and G. K. Smyth, "R code and downstream analysis objects for the scRNA-seq atlas of normal and tumorigenic human breast tissue," *Scientific Data*, vol. 9, p. 96, 2022, GEO: GSE161529.

[21] S. Tietscher, J. Wagner, T. Anzeneder, C. Langwieder, M. Rees, B. Sobottka, N. de Souza, and B. Bodenmiller, "A comprehensive single-cell map of T cell exhaustion-associated immune environments in human breast cancer," *Nature Communications*, vol. 14, p. 98, 2023, arrayExpress: E-MTAB-10607.

[22] A. Bassez, H. Vos, L. Van Dyck, G. Floris, I. Arijs, C. Desmedt, D. Lambrechts *et al.*, "A single-cell map of intratumoral changes during anti-PD1 treatment of patients with breast cancer," *Nature Medicine*, vol. 27, no. 5, pp. 820–832, 2021.

[23] L. Wang, W. Guo, Z. Guo *et al.*, "PD-L1-expressing tumor-associated macrophages are immunostimulatory and associate with good clinical outcome in human breast cancer," *Cell Reports Medicine*, vol. 5, no. 2, p. 101420, 2024, GEO: GSE248288.

[24] Y. Zhang, H. Chen, H. Mo, X. Hu, R. Gao, Y. Zhao, B. Liu, L. Niu, X. Su, X. Chang *et al.*, "Single-cell analyses reveal key immune cell subsets associated with response to PD-L1 blockade in triple-negative breast cancer," *Cancer Cell*, vol. 39, no. 12, pp. 1578–1593, 2021, GEO: GSE169246.

[25] Y.-M. Liu, J.-Y. Ge, Y.-F. Chen, Y. Liu, S.-Q. Wu, Q. Li *et al.*, "Combined single-cell and spatial transcriptomics reveal the metabolic evolvement of breast cancer during early dissemination," *Advanced Science*, vol. 10, no. 6, p. e2205395, 2023, GEO: GSE225600.

[26] R. Gao, S. Bai, Y. C. Henderson, Y. Lin, A. Schalck, Y. Yan, T. Kumar, M. Hu, E. Sei, A. Davis *et al.*, "Delineating copy number and clonal substructure in human tumors from single-cell transcriptomes," *Nature Biotechnology*, vol. 39, no. 5, pp. 599–608, 2021, GEO: GSE148673.

[27] K. Polański, M. D. Young, Z. Miao, K. B. Meyer, S. A. Teichmann, and J.-E. Park, "BBKNN: fast batch alignment of single cell transcriptomes," *Bioinformatics*, vol. 36, no. 3, pp. 964–965, 2020.

[28] C. Domínguez Conde, C. Xu, L. B. Jarvis, D. B. Rainbow, S. B. Wells, T. Gber, R. J. Skeels, A. R. M. Kraft, L. A. Poole, A. Jones *et al.*, "Cross-tissue immune cell analysis reveals tissue-specific features in humans," *Science*, vol. 376, no. 6594, p. eabl5197, 2022, CellTypist.

[29] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53–65, 1987.

[30] P. Rautenstrauch and U. Ohler, "Shortcomings of silhouette in single-cell integration benchmarking," *Nature Biotechnology*, pp. 1–5, 2025.