scDeepFusion: Auditable Gene-Module Representation Learning with Typed-Expert Routing for Single-Cell RNA-seq Integration

Anonymous authors
Paper under double-blind review

Abstract

Single-cell RNA sequencing presents a challenging representation learning problem: extreme sparsity, batch effects, and limited interpretability hinder accurate characterization of cellular states, particularly for rare or transcriptionally similar subtypes. We present **scDeepFusion**, a gene-set-level representation learning framework that represents cells by pathway-informed features rather than isolated gene measurements. Three components define the framework: (1) a module-activity representation that aggregates gene fluctuations over curated gene sets, suppressing dropout noise while preserving pathway-level biology; (2) a typed-expert routing architecture in which a per-cell sparse router assigns cells to knowledge-source experts (curated pathways, perturbation signatures, and a gene-line expert), yielding embeddings whose knowledge provenance is directly auditable from the exported routing weights; and (3) a teacher-aligned training scheme in which a coarse alignment loss anchors the representation to a reference embedding while auxiliary reconstruction and denoising objectives act as regularizers. Across five datasets evaluated under the standard scIB protocol with multi-seed controls on the flagship dataset and pathway-scoring baselines, scDeepFusion matches or exceeds strong integration baselines on aggregate metrics, attaining the highest Overall score among integration baselines on all four established benchmarks, with particular strengths on rare and subtype-level populations in label-weak settings. The framework further scales to 400k-cell training runs on the $>$1M-cell Tabula Sapiens atlas. This work establishes a principled and auditable framework for using curated biological knowledge in single-cell representation learning.

1 Introduction

Challenges in single-cell data. Single-cell RNA sequencing (scRNA-seq) enables high-resolution gene expression profiling at the individual cell level (Macosko et al., 2015; Regev et al., 2017), yet inherently suffers from three major challenges that impede accurate cellular characterization:

1. Extreme sparsity. scRNA-seq data exhibits extreme dropout (60–90% zeros) (Zheng et al., 2017; Ziegenhain et al., 2017; Zilionis et al., 2018; Zyla et al., 2023), creating a severe "dropout" problem where technical zeros obscure genuine biological signals (Huang et al., 2018; Van Dijk et al., 2018).

2. Limited interpretability. Gene-level profiles are unreliable under extreme sparsity: dropout events randomly obscure key genes, causing false negatives that confound cell-type assignment (Kharchenko et al., 2014; Luecken & Theis, 2019), and even the loss of a few marker genes can merge biologically distinct populations in subtle subtype discrimination. Biological identity, however, is often manifested at the pathway level rather than through individual genes (Subramanian et al., 2005; Liberzon et al., 2015); distinct cell types can share similar average expression yet differ systematically in pathway activity, making gene-only representations biologically uninformative.

3. Batch effects. Technical variations across experiments and platforms introduce systematic batch biases that can dominate biological signals (Haghverdi et al., 2018; Büttner et al., 2019); the harder cases are donor-level and technology-level batches, where over-correction erases exactly the biological variation of interest.

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.

Limitations of existing methods. Classical aligners such as Seurat (Stuart et al., 2019), Harmony (Korsunsky et al., 2019), BBKNN (Polanski & Adamusiak, 2020), and Scanorama (Hie et al., 2019) correct batches in low-dimensional space but model neither gene–gene relationships nor prior knowledge. Pathway-aware scorers (Fan et al., 2016; Liang et al., 2023; Kunes et al., 2024; Wang & Thakar, 2024; Singh et al., 2022) quantify gene-set activity yet treat pathways as static constraints rather than an interactive modality, while deep models (Lopez et al., 2018; Xu et al., 2021; Cui et al., 2024; Hao et al., 2024; Theodoris et al., 2023; Rosen et al., 2024) and reference-mapping frameworks (Lotfollahi et al., 2022) provide interpretable gene-program latents but target query-to-reference mapping rather than auditable integration. Collectively, these methods address only one challenge at a time and fail to jointly resolve sparsity, interpretability, and integration.

The dual frontier: atlas-scale and fine-grained discrimination. Consortium efforts such as the Human Cell Atlas (Regev et al., 2017) and Tabula Sapiens (Tabula Sapiens Consortium et al., 2022) advance breadth-first, cross-tissue mapping; yet an equally critical frontier is fine-grained discrimination within a shared tissue context, where distinguishing transcriptionally similar states (Trapnell et al., 2014; Semrau et al., 2017; Wolf et al., 2020) demands a resolution that atlas approaches often lack—a gap this work targets.

Key Contributions

We propose **scDeepFusion**, a gene-set-level representation learning framework whose embeddings pair competitive integration performance with per-cell auditable knowledge provenance. Our contributions are:

1. Auditable knowledge provenance through typed-expert routing. A per-cell sparse router assigns each cell to knowledge-source experts (curated pathways, perturbation signatures, and a gene-line expert); the exported routing weights state which knowledge source informed every cell's representation. The attribution is causally selective under expert masking and recovers a myeloid–lymphoid dichotomy replicated on a second dataset. Routing matches simpler fusion on aggregate metrics under identical training; its contribution is auditability, not score inflation.

2. A gene-set-level representation for single-cell analysis. Cells are represented by module activities over curated gene sets rather than isolated gene measurements, suppressing single-gene dropout noise while preserving pathway-level biology; this representation alone carries the aggregate performance (pathway-MLP control), and the architecture layers provenance on top.

3. A rigorous evaluation protocol with matched controls. All results use the standard scIB protocol (v1.1.5) in native embedding space with multi-seed replication, and a pathway-MLP control isolates the contribution of the module representation from that of the architecture. The framework shows particular strengths on rare and subtype-level populations in label-weak settings.

2 Method

2.1 Problem Formulation

Given a dataset with n cells and g genes, let $\mathbf{X} \in \mathbb{R}^{n \times g}$ denote the gene expression matrix (log-normalized counts). Each cell i belongs to a batch $b_i \in \mathcal{B}$ and has a cell type label $y_i \in \mathcal{Y}$ (used only for evaluation). Our goal is to learn an embedding function:

$$f : \mathbf{x}_i \mapsto \mathbf{h}_i \in \mathbb{R}^d \tag{1}$$

where $d \ll g$, such that: (1) $\mathbf{h}_i$ captures biological variation while removing batch effects; (2) similar cell types are close in embedding space; and (3) the embedding enables biological interpretation.

2.2 Overview

Given preprocessed scRNA-seq data, scDeepFusion consists of four core components (Figure 1):

(1) Reference Embedding Generator: a Harmony-corrected PCA teacher (Korsunsky et al., 2019) providing the alignment target (Section 2.5); the teacher is a plug-in component, replaceable by any batch-corrected reference.

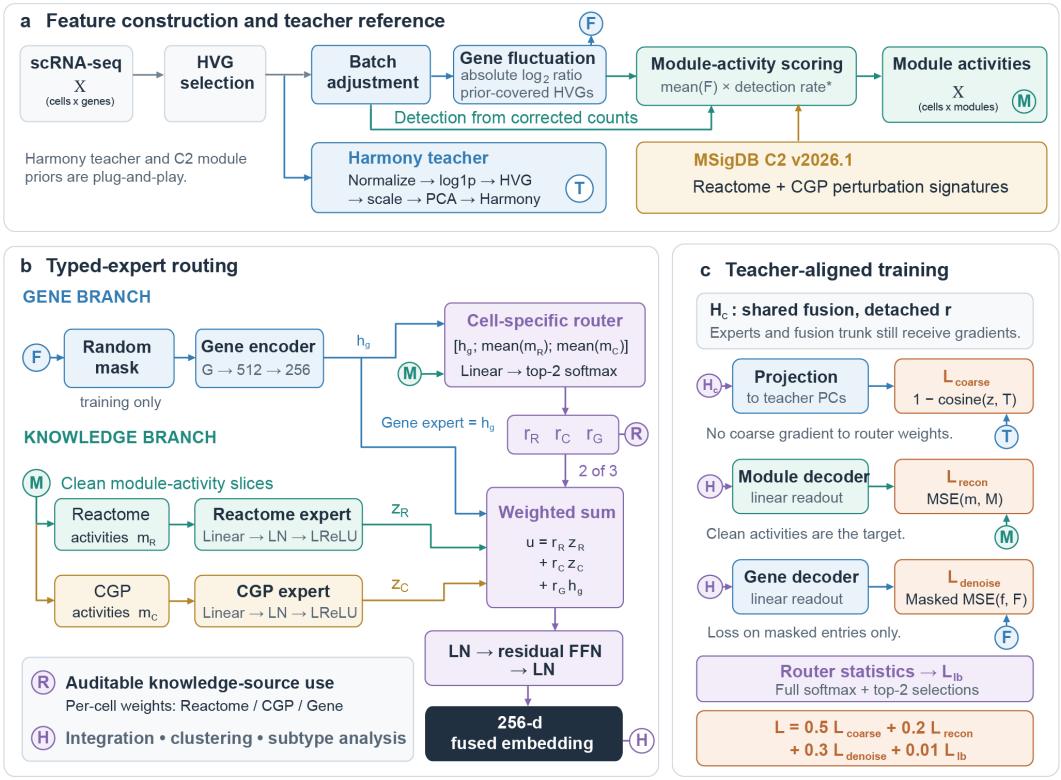


Figure 1: Architecture of scDeepFusion. **(a)** Feature construction and teacher reference: the teacher branch produces Harmony-corrected PCs; the feature branch derives gene-expression fluctuations and aggregates them over curated gene sets (MSigDB C2: Reactome pathways and CGP perturbation signatures) into module activities. **(b)** Typed-expert routing: each knowledge-source group's activity slice enters its typed expert; a per-cell router selects two of the three experts and forms the router-weighted sum with the gene embedding; per-cell weights provide knowledge-source attribution (Section 3.3). **(c)** Teacher-aligned training: coarse alignment plus auxiliary reconstruction, denoising, and load-balancing objectives (Section 2.5).

(2) Module Activity Calculator: aggregates gene fluctuations into a module-activity matrix (Section 2.3).

(3) BioNet Architecture: a dual-branch encoder fused by typed-expert routing over knowledge-source experts (Section 2.4).

(4) Training Objectives: coarse alignment to the teacher plus auxiliary reconstruction, denoising, and load-balancing objectives (Section 2.5).

Prior to these steps, standard preprocessing (quality control, library-size normalization, log-transformation, and a lightweight batch mean–variance alignment) and optional HVG selection are applied.

2.3 Module Activity Calculator

Motivation: As argued in the Introduction, biological identity is often manifested at the pathway level; averaging gene fluctuations over a co-regulated **module** therefore suppresses single-gene dropout noise while preserving pathway-level differences.

2.3.1 Gene Module Annotations

We use the MSigDB C2 collection (Curated Gene Sets; v2026.1) (Subramanian et al., 2005; Liberzon et al., 2015), consisting of 6,752 gene sets covering canonical pathways (CP, including Re-

actome) and chemical/genetic perturbations (CGP). This collection is a convenience default, not a requirement: it provides a structured, curated source for exactly the two knowledge types our typed experts consume—pathway programs and perturbation signatures—but users can substitute their own gene sets (KEGG, WikiPathways, Gene Ontology, or domain-specific sets). We additionally remove redundant parent sets: a gene set with $\geq$50 genes is dropped if it covers $\geq$90% of the genes in at least one smaller set. Let $\mathcal{G}_k$ denote the set of genes in module k.

2.3.2 Enhanced Module Activity Score

For each module k and cell i, we compute an activity score combining four metrics: mean fluctuation, detection rate, specificity, and variability (coefficient of variation). The module activity score is computed from raw expression after a lightweight batch mean–variance alignment (a simplified ComBat-style correction (Johnson et al., 2007) without empirical-Bayes shrinkage, implemented in our pipeline), as:

$$m_{ik} = g(\mu_{ik}, \rho_{ik}, \text{Spec}_{ik}, \text{CV}_{ik}, \bar{c}_k; |\mathcal{G}_k|) \tag{2}$$

where the function $g(\cdot)$ incorporates a size correction factor based on module cardinality $|\mathcal{G}_k|$. This yields the module activity matrix $\mathbf{M} \in \mathbb{R}^{n \times p}$. Notably, the module activity is computed from gene expression fluctuations (deviations from median) rather than absolute expression values, providing a directionless measure that captures the magnitude of module perturbation regardless of whether genes are up- or down-regulated. See Appendix A.2 for complete formulas.

2.4 BioNet Architecture

BioNet fuses gene fluctuations $\mathbf{F}$ and module activities $\mathbf{M}$ via typed-expert routing.

2.4.1 Gene Fluctuation Encoder

The fluctuation matrix $\mathbf{F}$ captures cell-specific expression deviations relative to gene medians. A two-layer MLP encoder with batch/layer normalization maps $\mathbf{f}_i$ to gene embedding $\mathbf{h}_i^{\text{gene}} \in \mathbb{R}^d$.

2.4.2 Module Branch

The module branch carries the module-activity matrix itself. Modules are partitioned into knowledge-source groups (Reactome, CGP); each group's activity slice feeds its own typed expert directly—there is no shared module encoder in the fusion path.

2.4.3 Typed-Expert Fusion

Motivation. Gene sets are not a homogeneous modality: curated pathways (e.g., Reactome) and perturbation signatures (e.g., CGP) differ in semantics, coverage, and reliability, and their relevance varies across cell types. Uniform fusion ignores this heterogeneity and, critically, discards the *provenance* of the knowledge being used. We therefore treat knowledge sources as typed experts and learn which sources each cell actually draws on.

Mechanism. Each knowledge-source group, together with the gene branch, defines one expert E_e (a single Linear→LayerNorm→LeakyReLU map over its activity slice; the gene expert is the gene embedding itself). For cell i, the router input is the concatenation of the gene embedding and the per-group mean activities, producing a distribution over experts; the two largest weights are kept and renormalized (k=2), following sparsely-gated mixture-of-experts (Shazeer et al., 2017):

$$\mathbf{w}_i = \text{softmax}\big(\text{Top}_2(R([\mathbf{h}_i^{\text{gene}};\ \bar{m}_{i,\text{Reactome}};\ \bar{m}_{i,\text{CGP}}]))\big), \qquad \mathbf{u}_i = \sum_{e \in \text{Top}_2} w_{ie}\, E_e(\mathbf{m}_{i,e}) + w_{iG}\, \mathbf{h}_i^{\text{gene}}, \tag{3}$$

and the fused representation follows as $\mathbf{h}_i^{\text{fuse}} = \text{LN}(\mathbf{u}_i + \text{FFN}(\text{LN}(\mathbf{u}_i)))$ with a residual feed-forward block. The gene contribution enters only through its routing weight w_{iG} (no unconditional gene residual). A Switch-style load-balancing loss (Fedus et al., 2022) (weight 0.01) prevents expert collapse, the coarse-alignment gradient is detached from the routing weights, and $\mathbf{w}_i$ is exported per cell as the knowledge-source attribution map (Section 3.3).

Design note. The purpose of typed routing is auditability of knowledge provenance, not aggregate-score gains (Appendix E).

2.4.4 PREDICTION HEADS

Two prediction heads are attached to the fused embedding: (1) a projection head producing the distilled readout $\hat{\mathbf{z}}_i \in \mathbb{R}^k$, whose dimensionality matches the teacher's; and (2) a module reconstruction head outputting $\tilde{\mathbf{m}}_i$. The fused representation $\mathbf{h}_i^{\text{fuse}}$ remains a free (unconstrained) representation, while $\hat{\mathbf{z}}_i$ is a distilled readout reported in benchmarks—analogous to the representation/projection-head separation in contrastive learning.

2.5 TRAINING OBJECTIVES

BioNet is trained with a coarse alignment loss plus auxiliary objectives:

$$\mathcal{L} = 0.5\,\mathcal{L}_{\text{coarse}} + 0.2\,\mathcal{L}_{\text{recon}} + 0.3\,\mathcal{L}_{\text{denoise}} + 0.01\,\mathcal{L}_{\text{lb}}, \tag{4}$$

where $\mathcal{L}_{\text{coarse}}$ is the cosine distance between the predicted readout $\hat{\mathbf{z}}_i$ and the Harmony teacher $\mathbf{z}_i$, $\mathcal{L}_{\text{recon}}$ is the MSE between reconstructed and computed module activities, $\mathcal{L}_{\text{denoise}}$ is a masked-gene reconstruction objective over the fluctuation input, and $\mathcal{L}_{\text{lb}}$ is the router load-balancing term (Appendix E isolates each term's contribution).

The Harmony reference is training supervision, not an architectural constraint: it operates on a PCA-compressed subspace $\mathbf{z}_i = \mathbf{W}_{\text{PCA}}\mathbf{x}_i \in \mathbb{R}^k$ ($k \ll g$), while BioNet learns from the full gene–module space, so alignment does not confine the model to Harmony's linear regime. The teacher is a plug-in component. We adopt Harmony-corrected PCs because PCA retains the coarse-grained population structure the alignment objective distills, but any batch-corrected reference embedding can serve; at inference, BioNet operates without it.

2.6 EVALUATION METRICS

Supervised Evaluation (scIB Metrics). Following the scIB benchmark (Luecken et al., 2022) (v1.1.5), we use six supervised metrics: NMI, ARI, ASW_label, Isolated labels F1, ASW_batch, and Graph connectivity, all computed in the native embedding space (for Seurat, its integrated CCA embedding; BBKNN outputs only a graph and is evaluated on its UMAP; UMAP is otherwise used for visualization only; NMI/ARI use Leiden clustering at the scIB-optimal resolution). We compute **Bio Score** (mean of the first four), **Batch Score** (mean of the last two), and **Overall Score** (60% Bio + 40% Batch).

Unsupervised Evaluation. We additionally compute a label-free, marker-anchored score from native-space KMeans averaged over cluster numbers and seeds: **MDF** (Marker Discrimination F-score over an a priori cell-type marker panel); panel construction and protocols are detailed in Appendix C. MDF is a metric we introduce in this work and is not yet standardized; to reduce clustering stochasticity, every reported value is averaged over five cluster numbers and three KMeans seeds.

3 EXPERIMENTS

3.1 EXPERIMENTAL SETUP

3.1.1 DATASETS

Cell subtype-level datasets. ifnb (Kang et al., 2018), GSE179994 (Liu et al., 2022), Immune Human, and Immune (Human+Mouse) (Luecken et al., 2022).

Tissue-level dataset. Pancreas (Luecken et al., 2022).

Atlas-scale exploratory dataset. Tabula Sapiens (Tabula Sapiens Consortium et al., 2022).

Rationale for dataset selection. We use immune datasets to test fine-grained subtype discrimination, a challenging task as immune cells are highly transcriptionally similar. Datasets progress in complexity: fine-grained discrimination (ifnb → GSE179994), multi-batch handling (GSE179994

→ Immune Human), and cross-species alignment (Immune Human → Human+Mouse). The Pancreas dataset validates that module-guided embedding generalizes beyond immune contexts to organ-level tissue analysis. Tabula Sapiens serves as an exploratory case study to probe the method's boundaries under cross-tissue heterogeneity (Table 5).

3.1.2 Baselines and Experimental Setup

We compare against representative methods from two categories:

Traditional methods: Seurat (Stuart et al., 2019), Harmony (Korsunsky et al., 2019), BBKNN (Polanski & Adamusiak, 2020), Scanorama (Hie et al., 2019).

Deep learning methods: scVI (Lopez et al., 2018; Xu et al., 2021), scGPT (Cui et al., 2024).

In addition, we benchmark three established pathway-scoring methods—AUCell, GSVA, and UCell—as module-level representation baselines under the same evaluation protocol, and a pathway-MLP control (an MLP trained on our module-activity matrix with the same teacher) that isolates the contribution of the representation from that of the architecture.

Training details and full hyperparameter configurations are provided in Appendix A. An anonymized implementation is available at `https://anonymous.4open.science/r/scdeepfusion-2026/`.

3.2 Main Results

scDeepFusion on ifnb.

The ifnb dataset is a well-established benchmark in the single-cell analysis community, with cell type annotations derived from Seurat clustering followed by manual curation and validation. These annotations are commonly used as reference labels for method evaluation.

Configuration. Immune datasets use HVG=3,000, the cutoff with the highest mean Overall in our grid scan (within seed noise of HVG=2,000; Appendix D, Table 4) and a common choice in recent integration and foundation-model studies (Cui et al., 2024; Fang et al., 2024), and a 40-PC Harmony teacher, which suffices as the supervision signal on the flat PCs axis; tissue-level systems benefit from substantially larger gene sets (validated on Pancreas below). All ifnb results use this fixed configuration without hyperparameter search, and all trainable variants are run with three random seeds.

Comparison with baseline methods. Table 1 presents the comparison under the standard scIB protocol. scDeepFusion attains the highest Overall Score (0.770 ± 0.003 over three seeds), exceeding the strongest deep baseline scVI (0.737) and the strongest classical baseline Seurat (0.732). The advantage concentrates on rare-type fidelity: scDeepFusion attains the highest isolated-label F1 (0.766 vs. 0.708 for scVI), indicating better recovery of small populations. A pathway-MLP control, an MLP readout trained on the module-activity matrix with the same teacher, matches scDeepFusion on the aggregate score (0.769 ± 0.003), confirming that the module representation itself carries the aggregate performance; established pathway scorers (AUCell, GSVA, UCell) behave identically after the same distilled readout (Appendix Table 8), and the full model additionally provides typed routing attribution and slightly higher isolated-label F1 (0.766 vs. 0.750 for the pathway-MLP control, Table 1).

Independent Validation via Marker Genes. To provide validation independent of the labeling source, we evaluate all methods with a label-free, marker-anchored metric over an a priori immune marker panel. scDeepFusion attains the highest MDF among integration baselines (0.410 vs. at most 0.399; Table 9), indicating the best-resolved marker-defined cell-type structure. UMAP visualizations (McInnes et al., 2018) (Figure 2) corroborate this, showing scDeepFusion achieves clearer separation of cell populations as well as several additional small cell clusters compared to baseline methods. A more detailed analysis is provided in Section 3.3.

Configuration Analysis and Practical Considerations. The fixed defaults are further supported by a two-round loss-weight sensitivity analysis (Table 7): Overall spans only 0.007 across the $w_{\text{coarse}} \times w_{\text{denoise}}$ grid with the default configuration at the top, so all reported results use one fixed default configuration per dataset class, with multi-seed replication for trainable variants.

Table 1: Performance comparison on ifnb (three-seed mean for scDeepFusion and pathway-MLP; best in **bold**, second underlined, by unrounded values). Seurat is evaluated on its native integrated CCA embedding (Section 2).

Method	Biological Conservation				Batch Correction		Composite Scores		
	NMI	ARI	ASW label	Iso F1	ASW batch	Graph Conn	Bio	Batch	Overall
BBKNN	0.637	0.348	0.492	0.613	0.792	0.599	0.523	0.695	0.592
Harmony	0.661	0.470	0.543	0.644	0.934	0.921	0.580	0.927	0.719
Seurat(V5)	0.655	0.455	0.558	0.741	0.890	0.964	0.602	0.927	0.732
Scanorama	0.672	0.486	0.536	0.654	**0.946**	0.930	0.587	0.938	0.727
scVI	0.668	0.470	**0.576**	0.708	0.913	0.957	0.606	0.935	0.737
scGPT (fine-tuned)	0.588	0.424	0.517	0.471	0.827	0.824	0.500	0.826	0.630
pathway-MLP (control)	**0.724**	**0.532**	0.571	0.750	0.925	0.988	**0.644**	0.957	0.769
scDeepFusion (Ours)	0.720	0.522	0.567	**0.766**	0.925	**0.993**	0.644	**0.959**	**0.770**

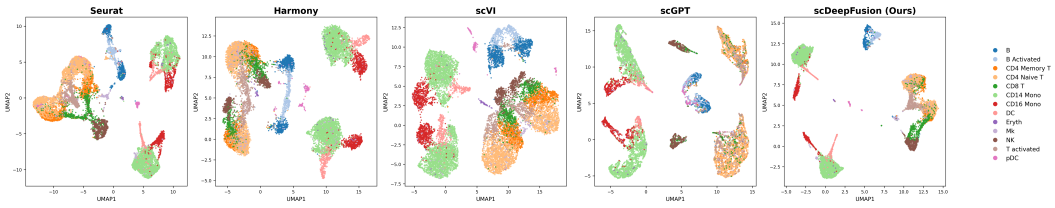


Figure 2: Qualitative comparison of integration methods on the ifnb dataset (Seurat, Harmony, scVI, fine-tuned scGPT, and scDeepFusion), colored by cell-type annotation. scDeepFusion (rightmost) achieves clearer cell population separation and reveals smaller distinct clusters compared to baselines, suggesting enhanced sensitivity in detecting rare subtypes.

Fine-grained Subtype Discrimination. GSE179994 is not a standard integration benchmark: its labels were assigned by the original study's own analysis pipeline rather than curated, and $CD4^+$ and $CD8^+$ T cells are transcriptionally too similar for reliable fine-grained annotation. The two-class labels also limit the supervised protocol: all biological-conservation metrics are keyed to the same coarse partition and isolated-label F1 saturates once the split is captured (0.84–0.95 for four of the eight methods), so supervised scores (Table 14; Harmony first, Overall 0.830, with scDeepFusion second at 0.776) are treated as reference only.

Label-free subtype-marker analysis is therefore the primary evidence. On an a priori 28-gene panel (24 genes available in the HVG subset) spanning CD4/CD8 lineages and functional T-cell states, scDeepFusion attains the highest MDF among all methods (0.301 ± 0.023 vs. 0.276 ± 0.014 for the strongest baseline, mean $\pm$ std over 15 cluster-number/seed runs; tied with the pathway-MLP control at 0.300; Table 15), and marker-expression maps corroborate this: a cytotoxic effector $CD8^+$ subpopulation marked by high *PRF1* and *GZMB* expression is clearly visible only in scDeepFusion (Figure 3).

Multi-batch and Cross-species Generalization. Table 16 presents results on the **Immune Human** atlas (32,484 cells from bone marrow and PBMCs across 9 technical batches). scDeepFusion attains the highest Overall Score (0.820), exceeding the strongest deep baseline scVI (0.792) and the strongest classical baseline Harmony (0.802), with the highest NMI (0.802) among all methods. The pathway-MLP control again matches scDeepFusion within noise (0.816). UMAP visualizations (Figure 4) corroborate the cleaner population separation.

Table 17 shows results on the cross-species **Immune Human+Mouse** dataset (96,839 cells, 22 batches; mouse cells mapped to human genes). scDeepFusion ties the pathway-MLP control (Overall 0.637 vs. 0.644) and both exceed all integration baselines (scGPT 0.623, scVI 0.622). Module-guided features thus remain competitive under cross-species variation, though without a uniform advantage over the simplest module readout.

Tissue-Level Validation. Tissue-level systems exhibit substantially higher transcriptional heterogeneity than immune datasets. Table 18 scans HVG cutoffs and teacher PCs on Pancreas (Fang et al., 2024): on iso_f1-free metrics the 3,000→6,000 HVG gain is real while further gains are dominated by the isolated-label F1 term, and the PCs curve plateaus at PCs≈50–150 and declines beyond PCs≈200. HVG filtering itself helps: unfiltered input trails HVG=12,000 by 0.04–0.07 on iso_f1-free metrics at matched PCs (its apparently competitive Overall at PCs=60 is an artifact of the degenerate isolated-label F1). scDeepFusion attains the highest Overall in the scan at PCs=120 (0.824 vs. 0.758 for Harmony and 0.754 for scVI); the nominally higher Overall at PCs=100–120 is driven by the degenerate isolated-label F1 and does not persist on iso_f1-free metrics, where the curve is flat (0.870–0.879 vs. 0.864 for Harmony). On this dataset the isolated-label F1 hinges on a single micro-population of seven annotated T cells (*t_cell*, 0.04% of all cells, spread over four batches): scDeepFusion is the only method that isolates it (0.46 vs. at most 0.21 for all baselines), demonstrating rare-population sensitivity; the metric itself is fragile at this size and should not be read as a reliable ranking signal.

Cross-dataset Summary. Table 2 consolidates scDeepFusion's performance across the five datasets evaluated under the standard protocol. scDeepFusion attains the highest Overall score on the four established benchmarks (ifnb, Immune Human, Immune H+M, and Pancreas); the exception, GSE179994, is not a standard benchmark and is instead analyzed by label-free metrics (Table 15). Detailed per-method comparisons for each dataset are provided in Appendix (Tables 14–18). Beyond the standard benchmarks, we assess scalability on the >1M-cell Tabula Sapiens atlas by full training runs at increasing scales (Table 19): training time grows near-linearly (2,393 s at 100k to 8,174 s at 400k), quality on the reduced protocol leads scVI by +0.12–0.14, and scDeepFusion remains stable at 400k where scVI diverges in all three seeds (Table 19, Figure 7); at 800k cells, preprocessing exceeds available memory, so we claim support up to the demonstrated 400k scale.

Table 2: scDeepFusion performance summary across datasets. For each dataset, the best baseline and its scores are shown alongside scDeepFusion (bold where scDeepFusion exceeds the baseline). The ifnb row reports the three-seed mean; all other rows report single-run results. *Coarse two-class labels; subtype analysis in Table 15. †Best Overall 0.824 at PCs=120; isolated-label F1 hinges on a single micro-population; iso_f1-free Overall 0.873 vs. 0.864 (Harmony). ‡Reduced protocol: no isolated-label F1, ASW on a fixed 5,000-cell subsample (Table 19).

Dataset				*Best Baseline*				*scDeepFusion*		
Name	**Cells**	**Types**	**Batch**	**Method**	**Bio**	**Batch**	**Overall**	**Bio**	**Batch**	**Overall**
Benchmark datasets										
ifnb	13,654	13	8	scVI	0.606	0.935	0.737	**0.644**	**0.959**	**0.770**
Immune Human	32,484	16	9	Harmony	0.710	0.940	0.802	**0.742**	0.937	**0.820**
Immune (H+M)	96,839	19	22	scGPT	0.458	0.871	0.623	0.447	**0.922**	**0.637**
Pancreas	16,382	14	9	Harmony	0.638	0.937	0.758	**0.740**	0.933	**0.824**†
Non-benchmark datasets										
GSE179994	150,849	2	36	Harmony	0.754	0.944	0.830	0.672	0.931	0.776*
Atlas-scale datasets (reduced protocol‡)										
Tabula Sapiens (200k)	200,000	182	46	scVI	0.492	0.618	0.543	**0.614**	**0.780**	**0.680**

Ablation Studies. Appendix E reports detailed ablations on ifnb (Table 6). The coarse alignment objective is the driver of the representation: removing it collapses Overall to 0.543 (reconstruction-only: 0.504), whereas a variant with only the alignment term matches the full model (0.767 vs. 0.767 at seed 42). The reconstruction and denoising terms act as regularizers. The teacher embedding itself attains 0.767 Overall under the same protocol (Table 1), and the coarse-only variant reproduces it almost exactly: the alignment term transfers the teacher's geometry into the module space, and the full model's gain concentrates on rare populations (isolated-label F1 0.766 vs. 0.750). Typed-expert routing matches plain concatenation (0.766) and dropping the module branch entirely matches as well (0.766) on this dataset's aggregate scores, so we do not claim an aggregate-score advantage of the routing mechanism; its contribution is the per-cell knowledge-source attribution analyzed in Section 3.3.

3.3 Interpretability Analysis

Routing Attribution. The exported per-cell routing weights make knowledge provenance directly auditable (Figure 6). On ifnb, the experts specialize by lineage: antigen-presenting myeloid cells route primarily to the CGP and gene-line experts (mean weights of 0.54/0.41 for DC), whereas lymphocytes route almost exclusively to Reactome (0.80–0.97). This dichotomy replicates on the Immune Human atlas, which has no stimulation condition (monocyte-derived DC and $CD16^+$ monocytes again route to CGP and the gene-line expert at 0.55/0.42 and 0.54/0.40; Table 12). The weights are also causally selective. Masking CGP displaces DC, pDC, and monocytes (cosine shift 0.10–0.13) while masking Reactome displaces $CD4^+$ T cells (0.46–0.48) and B cells (0.34), leaving DC unchanged in both cases (0.002; Table 11; protocol: the masked expert's routing weight is zeroed, the remaining weights are renormalized per cell, and the fused combination $\mathbf{u}_i$ is recomputed before the readout; shifts are mean cosine distances per cell type). Such attribution offers a provenance check on every population and an actionable error signal, and it is exact for the fused combination $\mathbf{u}_i$ rather than a post-hoc approximation (best read as relative preferences, since the alignment gradient is detached from the router). The megakaryocyte and erythroid populations below show a third profile (Reactome-dominated, 0.73–0.75), indicating that their consolidation reflects a shared knowledge source rather than a teacher artifact. **Case Study: Rare-population Consolidation.** On ifnb, scDeepFusion consolidates rare populations into single dominant clusters (Figure 5). For the megakaryocyte population (Mk, 224 cells), it forms the purest dominant region of all methods (15-nearest-neighbor label purity 0.70 vs. at most 0.59, concentrated in its Leiden cluster 18; Table 10), whereas Seurat's native CCA space attains a slightly higher isolated-label F1 (0.70 vs. 0.66), consistent with a more inclusive but less pure grouping. For the erythroid precursors (Eryth, 54 cells), scDeepFusion matches the best isolated-label F1 (0.94; scVI 0.94, Seurat 0.90), while Seurat shows the cleanest immediate neighborhoods (purity 0.95 vs. 0.88); a Leiden cluster computed on our embedding (cluster 21, 49 cells) recovers 48 of the 54 cells at 98% purity. The groups' identities are corroborated by marker expression: the 54 cells with the highest mean *HBA1*/*HBA2*/*HBB* expression are exactly the 54 annotated Eryth cells, and *PPBP*/*PF4*/*GP9*/*ITGA2B* expression co-localizes with the consolidated Mk region.

4 Limitations and Future Work

Current limitations include: (1) reliance on curated gene sets with potential redundancy and incomplete coverage for non-model organisms; (2) cross-tissue atlas integration requires larger gene sets and dedicated configuration; and (3) scope is currently limited to scRNA-seq, with no empirical validation on multi-omics or spatial transcriptomics data (Moses & Pachter, 2022).

Future work will extend scDeepFusion to spatial and multi-omics settings and to full-atlas training without subsampling; methodological directions include routing-aware objectives that convert attribution into aggregate-score gains, alternative teacher references, organism-specific module libraries, sparse-attention variants (Child et al., 2019; Katharopoulos et al., 2020), and learnable gene-set selection (Liu et al., 2019).

5 Conclusion

We introduced scDeepFusion, a gene-set-level representation learning framework that advances single-cell analysis from isolated gene measurements to pathway-informed cellular states. Under the standard scIB protocol, scDeepFusion attains the highest Overall score among integration baselines on all four established benchmarks while offering two properties integration baselines lack: auditable knowledge provenance (routing weights decompose each cell's representation across curated pathways, perturbation signatures, and the gene line, recovering a myeloid–lymphoid dichotomy without lineage supervision) and resolution of rare and subtype-level populations that competing methods fragment (highest isolated-label F1 on ifnb; first-ranking subtype-marker discrimination on the unreliably labeled GSE179994; the only isolation of the 7-cell *t_cell* micro-population on Pancreas; marker-verified consolidation of erythroid and megakaryocyte populations into single compact, high-purity groups). Together, these results establish a principled and auditable framework for leveraging curated biological knowledge in single-cell representation learning.

REFERENCES

Maren Büttner, Zhichao Miao, F Alexander Wolf, Sarah A Teichmann, and Fabian J Theis. A test metric for assessing single-cell RNA-seq batch correction. *Nature Methods*, 16(1):43–49, 2019.

Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. Generating long sequences with sparse transformers. *arXiv:1904.10509*, 2019. URL https://arxiv.org/abs/1904.10509.

Haotian Cui, Chloe Wang, Hassaan Maan, Kuan Pang, Fengning Luo, Nan Duan, and Bo Wang. scGPT: toward building a foundation model for single-cell multi-omics using generative AI. *Nature Methods*, 21(8):1470–1480, 2024. doi: 10.1038/s41592-024-02201-0.

Jean Fan, Neeraj Salathia, Rui Liu, Gwendolyn E Kaeser, Yun C Yung, Joseph L Herman, Fiona Kaper, Jian-Bing Fan, Kun Zhang, Jerold Chun, et al. Characterizing transcriptional heterogeneity through pathway and gene set overdispersion analysis. *Nature Methods*, 13(3):241–244, 2016.

Yi Fang, Junjie Chen, He Wang, Shousen Wang, Mengqi Chang, Qingcai Chen, Qinglei Shi, Liang Xian, Ming Feng, Baotian Hu, and Renzhi Wang. Integrating large-scale single-cell RNA sequencing in central nervous system disease using self-supervised contrastive learning. *Communications Biology*, 7(1):1107, 2024. doi: 10.1038/s42003-024-06813-2.

William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022.

Laleh Haghverdi, Aaron TL Lun, Maria Morgan, and John C Marioni. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nature Biotechnology*, 36(5):421–427, 2018.

Minsheng Hao, Jing Gong, Xin Zeng, Chiming Liu, Yucheng Guo, Xingyi Cheng, Taifeng Wang, Jianzhu Ma, Xuegong Zhang, and Le Song. Large-scale foundation model on single-cell transcriptomics. *Nature Methods*, 21(8):1481–1491, 2024. doi: 10.1038/s41592-024-02305-7.

Brian Hie, Bryan Bryson, and Bonnie Berger. Efficient integration of heterogeneous single-cell transcriptomes using Scanorama. *Nature Biotechnology*, 37(6):685–691, 2019.

Mo Huang, Jingshu Wang, Eduardo Torre, Hannah Dueck, Sydney Shaffer, Roberto Bonasio, John I Murray, Arjun Raj, Mingyao Li, and Nancy R Zhang. SAVER: gene expression recovery for single-cell RNA sequencing. *Nature Methods*, 15(7):539–542, 2018.

W Evan Johnson, Cheng Li, and Ariel Rabinovic. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*, 8(1):118–127, 2007.

Hyun Min Kang, Meena Subramaniam, Sasha Targ, Michelle Nguyen, Lenka Maliskova, Elizabeth McCarthy, Eunice Wan, Simon Wong, Lauren Byrnes, Cristina M Lanata, et al. Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. *Nature Biotechnology*, 36 (1):89–94, 2018.

Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are RNNs: Fast autoregressive transformers with linear attention. In *International Conference on Machine Learning*, pp. 5156–5165. PMLR, 2020.

Peter V Kharchenko, Lev Silberstein, and David T Scadden. Bayesian approach to single-cell differential expression analysis. *Nature Methods*, 11(7):740–742, 2014.

Ilya Korsunsky, Nghia Millard, Jean Fan, Kamil Slowikowski, Fan Zhang, Kevin Wei, Yury Baglaenko, Michael Brenner, Po-ru Loh, and Soumya Raychaudhuri. Fast, sensitive and accurate integration of single-cell data with harmony. *Nature Methods*, 16(12):1289–1296, 2019.

Russell Z Kunes, Thomas Walle, Max Land, Tal Nawy, and Dana Pe'er. Supervised discovery of interpretable gene programs from single-cell data. *Nature Biotechnology*, 42(7):1084–1095, 2024. doi: 10.1038/s41587-023-01940-3.

Qingnan Liang, Yuefan Huang, Shan He, and Ken Chen. Pathway centric analysis for single-cell RNA-seq and spatial transcriptomics data with GSDensity. *Nature Communications*, 14(1):8416, 2023. doi: 10.1038/s41467-023-44206-x.

Arthur Liberzon, Chet Birger, Helga Thorvaldsdöttir, Mahmoud Ghandi, Jill P Mesirov, and Pablo Tamayo. The molecular signatures database hallmark gene set collection. *Cell Systems*, 1(6): 417–425, 2015.

Baolin Liu, Xueda Hu, Kaichao Feng, Ranran Gao, Zhiqiang Xue, Sujie Zhang, Yuanyuan Zhang, Emily Corse, Han Yi, Lei Zhang, et al. Temporal single-cell tracing reveals clonal revival and expansion of precursor exhausted T cells during anti-PD-1 therapy in lung cancer. *Nature Cancer*, 3(1):108–121, 2022.

Hanxiao Liu, Karen Simonyan, and Yiming Yang. DARTS: Differentiable architecture search. In *International Conference on Learning Representations*, 2019.

Romain Lopez, Jeffrey Regier, Michael B Cole, Michael I Jordan, and Nir Yosef. Deep generative modeling for single-cell transcriptomics. *Nature Methods*, 15(12):1053–1058, 2018.

Mohammad Lotfollahi, Mohsen Naghipourfar, Malte D Luecken, Masoud Khajavi, Maren B"uttner, Tim Wuestefeld, Žiga Avsec, Adam Gayoso, Nadia Jouy, Alexander S Misharin, et al. Mapping single-cell data to reference atlases by transfer learning. *Nature Biotechnology*, 40(1):121–130, 2022.

Malte D Luecken and Fabian J Theis. Current best practices in single-cell RNA-seq analysis: a tutorial. *Molecular Systems Biology*, 15(6):e8746, 2019.

Malte D Luecken, Maren Büttner, Kridsadakorn Chaichoompu, Andrea Danese, Marta Interlandi, Michaela F Müller, Daniel C Strobl, Luke Zappia, Martin Dugas, Maria Colomé-Tatché, et al. Benchmarking atlas-level data integration in single-cell genomics. *Nature Methods*, 19(1):41–50, 2022.

Evan Z Macosko, Anindita Basu, Rahul Satija, James Nemesh, Karthik Shekhar, Melissa Goldman, Itay Tirosh, Allison R Bialas, Nolan Kamitaki, Emily M Martersteck, et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*, 161(5):1202–1214, 2015.

Leland McInnes, John Healy, and James Melville. UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

Lambda Moses and Lior Pachter. Museum of spatial transcriptomics. *Nature Methods*, 19(5):534–546, 2022.

Krzysztof Polanski and Tomasz Adamusiak. BBKNN: fast batch alignment of single cell transcriptomes. *Bioinformatics*, 36(3):964–965, 2020.

Aviv Regev, Sarah A Teichmann, Eric S Lander, Ido Amit, Christophe Benoist, Ewan Birney, Bernd Bodenmiller, Peter Campbell, Piero Carninci, Menna Clatworthy, et al. The Human Cell Atlas. *eLife*, 6:e27041, 2017.

Yanay Rosen, Maria Brbić, Yusuf Roohani, Kyle Swanson, Ziang Li, and Jure Leskovec. Toward universal cell embeddings: integrating single-cell RNA-seq datasets across species with SATURN. *Nature Methods*, 21(8):1492–1500, 2024. doi: 10.1038/s41592-024-02191-z.

Rahul Satija, Jeffrey A Farrell, David Gennert, Alexander F Schier, and Aviv Regev. Spatial reconstruction of single-cell gene expression data. *Nature Methods*, 13(6):495–498, 2016.

Stefan Semrau, Johanna E Goldmann, Magali Soumillon, Tarjei S Mikkelsen, Rudolf Jaenisch, and Alexander van Oudenaarden. Dynamics of lineage commitment revealed by single-cell transcriptomics of differentiating embryonic stem cells. *Nature Communications*, 8(1):1096, 2017.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc V. Le, Geoffrey E. Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *International Conference on Learning Representations (ICLR)*, 2017.

Anand Pratap Singh, Shefali Seth Verma, Idoia Ajuria, Vipin Narang, Mohit Kumar Jolly, and Kausalya Subramanian. Unipath: a uniform approach for pathway and gene-set based analysis of heterogeneity in single-cell epigenome and transcriptome profiles. *Nucleic Acids Research*, 50(3):e14, 2022.

Tim Stuart, Andrew Butler, Paul Hoffman, Christoph Hafemeister, Efthymia Papalexi, William M Mauck, Yuhan Hao, Marlon Stoeckius, Peter Smibert, and Rahul Satija. Comprehensive integration of single-cell data. *Cell*, 177(7):1888–1902, 2019.

Aravind Subramanian, Pablo Tamayo, Vamsi K Mootha, Sayan Mukherjee, Benjamin L Ebert, Michael A Gillette, Amanda Paulovich, Scott L Pomeroy, Todd R Golub, Eric S Lander, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43):15545–15550, 2005.

Tabula Sapiens Consortium, Robert C Jones, Jimena Karkanias, Mark A Krasnow, Angela Oliveira Pisco, Stephen R Quake, Julia Salzman, Nir Yosef, Bryan Bulthaup, Phillip Brown, et al. The Tabula Sapiens: a multiple-organ, single-cell transcriptomic atlas of humans. *Science*, 376(6594): eabl4896, 2022.

Christina V Theodoris, Ling Xiao, Anant Chopra, Mark D Chaffin, Zeina R Al Sayed, Matthew C Hill, Harriet Mantineo, Elizabeth M Brydon, Zexian Zeng, X Shirley Liu, et al. Transfer learning enables predictions in network biology. *Nature*, 618(7965):616–624, 2023.

Cole Trapnell, Davide Cacchiarelli, Jonna Grimsby, Prapti Pokharel, Shuqiang Li, Michael Morse, Niall J Lennon, Kenneth J Livak, Tarjei S Mikkelsen, and John L Rinn. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nature Biotechnology*, 32(4):381–386, 2014.

David Van Dijk, Roshan Sharma, Juozas Nainys, Kristina Yim, Pooja Kathail, Ambrose J Carr, Cassandra Burdziak, Kevin R Moon, Christine L Chaffer, Diwakar Pattabiraman, et al. Recovering gene interactions from single-cell data using data diffusion. *Cell*, 174(3):716–729, 2018.

Ruoqiao H Wang and Juilee Thakar. Comparative analysis of single-cell pathway scoring methods and a novel approach. *NAR Genomics and Bioinformatics*, 6(3):lqae124, 2024. doi: 10.1093/nargab/lqae124.

Dennis Wolf, Till Gerhardt, Holger Winkels, Nora A Michel, Akshay B Pramod, Yasmin Ghosheh, Samuel Brunel, Konrad Buscher, Jacob Miller, Sara McArdle, et al. Pathogenic autoimmunity in atherosclerosis evolves from initially protective apolipoprotein B100–reactive $CD4^+$ T-regulatory cells. *Circulation*, 142(13):1279–1293, 2020.

Chenling Xu, Romain Lopez, Edouard Mehlman, Jeffrey Regier, Michael I Jordan, and Nir Yosef. Probabilistic harmonization and annotation of single-cell transcriptomics data with deep generative models. *Molecular Systems Biology*, 17(1):e9620, 2021. doi: 10.15252/msb.20209620.

Grace XY Zheng, Jessica M Terry, Phillip Belgrader, Paul Ryvkin, Zachary W Bent, Ryan Wilson, Solongo B Ziraldo, Thomas D Wheeler, Gregory P McDermott, Junjie Zhu, et al. Massively parallel digital transcriptional profiling of single cells. *Nature Communications*, 8:14049, 2017.

Christoph Ziegenhain, Beate Vieth, Swati Parekh, Björn Reinius, Amy Guillaumet-Adkins, Martha Smets, Heinrich Leonhardt, Holger Heyn, Ines Hellmann, and Wolfgang Enard. Comparative analysis of single-cell RNA sequencing methods. *Molecular Cell*, 65(4):631–643, 2017.

Ruslan Zilionis, Christoph Engblom, Christina Pfirschke, Virginia Savova, David Zemmour, Hatice D Saatcioglu, Indhu Krishnan, Geraldine Maroni, Christina V Meyerovitz, Catherine M Kerwin, et al. Single-cell transcriptomics of human and mouse lung cancers reveals conserved myeloid populations across individuals and species. *Immunity*, 48(4):922–938, 2018.

Joanna Zyla, Anna Papiez, Jun Zhao, Rihao Qu, Xiaotong Li, Yuval Kluger, Joanna Polanska, Christos Hatzis, Lajos Pusztai, and Michal Marczyk. Evaluation of zero counts to better understand the discrepancies between bulk and single-cell RNA-Seq platforms. *Computational and Structural Biotechnology Journal*, 21:4663–4674, 2023. doi: 10.1016/j.csbj.2023.09.035.

A Model Implementation Details

An anonymized implementation, including training scripts, preprocessing pipelines, and model configurations, is available at `https://anonymous.4open.science/r/scdeepfusion-2026/`.

A.1 Training Configuration and Hardware Platform

Result stability. Across three seeds (42/43/44), scDeepFusion's Overall varies by only ± 0.003 (Table 13), demonstrating robustness to random initialization.

For all experiments we use fixed hyperparameters without search: embedding dimension $d = 256$, hidden dimension $h = 512$, dropout rate 0.3, batch size 128. Training uses AdamW with learning rate 5×10^{-4} and weight decay 1×10^{-4} for up to 300 epochs with early stopping (patience=15). The loss weights are $0.5\,\mathcal{L}_{\text{coarse}} + 0.2\,\mathcal{L}_{\text{recon}} + 0.3\,\mathcal{L}_{\text{denoise}} + 0.01\,\mathcal{L}_{\text{lb}}$. All reported experiments use these fixed defaults without hyperparameter search, avoiding post-hoc selection; a two-round loss-weight sensitivity analysis confirms the defaults sit in a flat region (Table 7).

Hardware Platform. All experiments are conducted on an Apple Studio large-memory workstation, using Metal Performance Shaders (MPS) for GPU acceleration. Implemented in PyTorch, the pipeline is compatible with both Apple Silicon (MPS) and NVIDIA GPUs (CUDA). Per-step wall time and peak memory of every pipeline stage are logged automatically (Appendix B.2).

A.2 Module Activity Calculator

This subsection provides complete formulas for the module activity score computation described in Section 2.3.

For each module k with gene set $\mathcal{G}_k$ and cell i, the score is computed from two inputs: the fluctuation matrix $\mathbf{F}$, where $F_{ij} = |\log_2(\tilde{X}_{ij}+\epsilon) - \log_2(m_j+\epsilon)|$ with $\tilde{X}$ the library-size-normalized ComBat-corrected expression and m_j the per-gene median over non-zero cells (entries are zeroed where $m_j = 0$ or the count is zero), and the ComBat-corrected raw counts $\tilde{X}^{\text{raw}}_{ij}$, which are used only for detection:

1. **Mean fluctuation:** $\mu_{ik} = \frac{1}{|\mathcal{G}_k|}\sum_{j\in\mathcal{G}_k} F_{ij}$
2. **Detection rate:** $\rho_{ik} = \frac{1}{|\mathcal{G}_k|}\sum_{j\in\mathcal{G}_k} \mathbb{I}[\tilde{X}^{\text{raw}}_{ij} > 0]$
3. **Specificity:**
$$\text{Spec}_{ik} = \frac{\max_{j\in\mathcal{G}_k} F_{ij} - \min_{j\in\mathcal{G}_k} F_{ij}}{\max_{j\in\mathcal{G}_k} F_{ij} + \min_{j\in\mathcal{G}_k} F_{ij} + \epsilon} \tag{5}$$
4. **Variability:** $\text{CV}_{ik} = \sigma_{ik}/\mu_{ik}$ (set to zero when $\mu_{ik} \le 10^{-8}$), where σ_{ik} is the standard deviation of $\{F_{ij}\}_{j\in\mathcal{G}_k}$

The module activity score is computed as follows:

First, we compute the base activity and transformation factors:

$$\text{Base}_{ik} = \mu_{ik} \cdot \rho_{ik} \tag{6}$$
$$\text{SpecFactor}_{ik} = (\text{Spec}_{ik})^{0.8} \tag{7}$$
$$\text{Dynamic}_{ik} = 0.7 \cdot \text{SpecFactor}_{ik} + 0.3 \tag{8}$$
$$\text{CVFactor}_{ik} = \sigma(2 \cdot \text{CV}_{ik}) \tag{9}$$

where $\sigma(\cdot)$ is the sigmoid function. The preliminary activity is:

$$a_{ik} = \text{Base}_{ik} \cdot (1 + \text{CVFactor}_{ik} \cdot \text{Dynamic}_{ik}) \tag{10}$$

Finally, the size correction factor S_k is applied:

$$S_k = \begin{cases} 1.0 + \frac{1.0}{|\mathcal{G}_k|+1} & \text{if } |\mathcal{G}_k| < 10 \\ 1.0 - 0.15 \cdot \bar{\rho}_k \text{ (mean detection rate of module } k) & \text{if } |\mathcal{G}_k| > 30 \\ 1.0 & \text{otherwise} \end{cases} \tag{11}$$

The final module activity is winsorized per module at its 99.9th percentile (preventing a handful of extreme values from dominating the reconstruction loss) and log-transformed:

$$m_{ik} = \log\big(1 + \min(a_{ik} \cdot S_k,\ q_{99.9}(a_{\cdot k} \cdot S_k))\big). \tag{12}$$

A.3 BioNet Architecture

This subsection provides complete architectural specifications for the BioNet model described in Section 2.4.

A.3.1 Gene Fluctuation Encoder

The fluctuation matrix $\mathbf{F}$ captures cell-specific expression deviations. First, library-size normalization is applied:

$$\tilde{X}_{ig} = \frac{X_{ig}}{\sum_j X_{ij}} \times 10^4 \tag{13}$$

Then, the fluctuation is computed as the absolute log-fold change relative to the gene's median expression (computed over non-zero values):

$$F_{ig} = \left| \log_2 \left(\frac{\tilde{X}_{ig} + \epsilon}{\text{median}(\tilde{X}_{*g}[\tilde{X}_{*g} > 0]) + \epsilon} \right) \right| \tag{14}$$

The gene encoder maps $\mathbf{f}_i$ to embedding space using a two-layer MLP with batch normalization, layer normalization, and dropout:

$$\mathbf{h}_i^{(1)} = \text{LeakyReLU}(\text{BN}(\mathbf{W}_1 \mathbf{f}_i)) \tag{15}$$

$$\mathbf{h}_i^{\text{gene}} = \text{LeakyReLU}(\text{LN}(\mathbf{W}_2 \mathbf{h}_i^{(1)})) \tag{16}$$

where $\mathbf{W}_1 \in \mathbb{R}^{h \times g}$, $\mathbf{W}_2 \in \mathbb{R}^{d \times h}$, and dropout is applied between the two layers with rate δ.

A.3.2 Module Branch

Modules are partitioned into knowledge-source groups by their prefix in the curated library; in the reported configuration the groups are Reactome pathways and CGP perturbation signatures. Each group's activity slice feeds one module expert directly (no shared module encoder in the fusion path), and the gene branch defines an additional gene expert, giving $E{=}3$ experts in total.

A.3.3 Typed-Expert Fusion

For cell i, the router input is the concatenation of the gene embedding and the per-group mean activities (one scalar per group):

$$\mathbf{w}_i = \text{softmax}\big(\mathbf{W}_r\,[\mathbf{h}_i^{\text{gene}};\ \bar{m}_{i,\text{Reactome}};\ \bar{m}_{i,\text{CGP}}]\big) \in \Delta^{E-1} \tag{17}$$

Top-k sparse gating ($k{=}2$) keeps the two largest weights, zeroes the rest, and renormalizes:

$$\tilde{w}_{ie} = \frac{w_{ie} \cdot \mathbb{1}[e \in \text{Top}_2(\mathbf{w}_i)]}{\sum_{e' \in \text{Top}_2(\mathbf{w}_i)} w_{ie'}} \tag{18}$$

Each module expert is a single Linear→LayerNorm→LeakyReLU map over its group's activity slice; the gene expert is the gene embedding itself. The fused embedding is the router-weighted combination, followed by LayerNorm, a residual feed-forward block and LayerNorm:

$$\mathbf{u}_i = \sum_e \tilde{w}_{ie}\, E_e(\mathbf{m}_{i,e}) + \tilde{w}_{iG}\, \mathbf{h}_i^{\text{gene}}, \qquad \mathbf{h}_i^{\text{fuse}} = \text{LN}\big(\mathbf{u}_i + \text{FFN}(\text{LN}(\mathbf{u}_i))\big). \tag{19}$$

The gene contribution enters only through its routing weight $\tilde{w}_{iG}$; there is no unconditional gene residual.

Load balancing. Following Fedus et al. (2022), the auxiliary loss uses dense softmax probabilities and top-2 selection frequencies, $\mathcal{L}_{\text{lb}} = E \cdot \sum_e f_e \cdot P_e$ with weight 0.01, where f_e is the fraction of cells routed to expert e and P_e is the mean dense-softmax probability, preventing expert collapse.

Detach shield. The coarse-alignment loss is computed on a fused representation whose routing weights are detached from the router (stop-gradient), while the reconstruction and denoising losses train the router normally. This prevents the alignment gradient from flattening the routing distribution into near-uniform weights.

Auditability. The per-cell routing weights $\tilde{\mathbf{w}}_i$ are exported alongside the embedding, yielding a knowledge-source attribution map (Section 3.3).

A.3.4 Prediction Heads

1. **Embedding Projection:** $\hat{\mathbf{z}}_i = \tanh(\text{BN}(\mathbf{W}_z \mathbf{h}_i^{\text{fuse}}))$
2. **Module Reconstruction:** $\tilde{\mathbf{m}}_i = \mathbf{W}_d \mathbf{h}_i^{\text{fuse}}$ (linear decoder)

A.3.5 Training Objective Details

Coarse alignment loss (cosine distance to the Harmony teacher $\mathbf{z}_i$):

$$\mathcal{L}_{\text{coarse}} = 1 - \frac{\hat{\mathbf{z}}_i \cdot \mathbf{z}_i}{\|\hat{\mathbf{z}}_i\| \|\mathbf{z}_i\|} \tag{20}$$

Module reconstruction loss (MSE on log1p-normalized activities; the decoder is linear, without a sigmoid, so the target range is reachable):

$$\mathcal{L}_{\text{recon}} = \|\tilde{\mathbf{m}}_i - \mathbf{m}_i\|_2^2 \tag{21}$$

Denoise loss (MSE): a random subset of gene fluctuation entries is masked (set to zero) at the encoder input, and a light decoder reconstructs the unmasked fluctuation vector.

B Computational and Performance Analysis

B.1 Complexity Analysis (Theoretical)

We provide a detailed complexity analysis of scDeepFusion, demonstrating that despite incorporating module-level computations and typed-expert routing, the method maintains linear scalability with respect to the number of cells.

Time Complexity. Per-epoch complexity is $O(n \cdot (g \cdot h + p \cdot d + E \cdot d^2))$ with E experts; with fixed architecture, this scales linearly with n.

Detailed Component Analysis. We decompose the per-cell computational cost:

(1) Gene Encoder: Two linear transformations $\mathbf{W}_1 \in \mathbb{R}^{h \times g}$ and $\mathbf{W}_2 \in \mathbb{R}^{d \times h}$ require $O(g \cdot h)$ and $O(h \cdot d)$ operations, respectively. Normalization and activation functions are $O(h)$ and $O(d)$, dominated by the matrix multiplications.

(2) Typed-Expert Routing: The router is a linear map over the concatenated gene embedding and group-pooled module activities, costing $O(E \cdot d)$; each selected expert is a single Linear→LayerNorm→LeakyReLU map on its activity slice, costing $O(p_e \cdot d)$ with p_e the slice dimensionality. With E=3 experts and k=2 selected, the fusion cost is a constant $O(p + d^2)$ per cell.

(3) Prediction Heads: The projection head and decoder each require $O(d^2)$ and $O(d \cdot p)$ operations, respectively.

Summing these components: $O(g \cdot h + h \cdot d + p \cdot d + d^2)$ per cell. With the reported configuration ($h = 512$, $d = 256$, $g \approx 10^4$, $p \approx 6 \times 10^3$), the dominant term is $g \cdot h$. Because all architectural dimensions (h, d, E) are fixed hyperparameters, the per-cell cost is constant and total complexity scales linearly with n.

Space Complexity. The space complexity is $O(g \cdot h + h \cdot d + p \cdot d + n \cdot d)$ for model parameters and activations.

Storage Breakdown: Gene encoder parameters ($g \cdot h + h \cdot d$), router ($E \cdot d$), typed experts ($p \cdot d$ for the two module slices), and decoder ($d \cdot p$) sum to $O(g \cdot h + h \cdot d + p \cdot d)$, independent of n. The dense module-activity matrix adds $O(n \cdot p)$, which dominates memory at atlas scale (Appendix B.2). Cell embeddings require $O(n \cdot d)$ storage. With $d \ll g$, total storage scales linearly with n.

B.2 Scalability and Performance (Empirical)

The linear scaling with the number of cells (Appendix B.1) ensures asymptotic scalability. Table 3 reports the measured end-to-end runtime and peak memory of every pipeline stage on ifnb (13,654 cells), recorded automatically by the pipeline's built-in performance recorder on the same workstation. scDeepFusion trains $1.9\times$ faster than scVI and $1.5\times$ faster than fine-tuned scGPT; the module-activity preprocessing adds only 4 s (one-time, cached). The modular PyTorch implementation is compatible with both Apple MPS and standard CUDA accelerators.

Table 3: Measured runtime and peak memory on ifnb (13,654 cells; hardware platform in Appendix A). Train = wall-clock training time (scDeepFusion and pathway-MLP: mean over three seeds); Peak RSS = peak resident memory. scDeepFusion inference throughput is ≈13,300 cells/s. Timings are from the runs producing the reported embeddings.

Method	Train time (s)	Peak memory (GB)
Seurat (R: SCTransform + CCA)	427	14.1
Scanorama	46	4.0
BBKNN + Harmony	37	4.8
scVI	3,071	3.4
scGPT (fine-tune)	2,437	6.5
data preprocessing + module activity (one-time, cached)	4	2.8
pathway-MLP (control)	84	2.5
scDeepFusion (Ours)	1,599	9.6

Hardware and Implementation. All experiments are conducted on the same Apple Studio workstation (Appendix A). While our implementation leverages Apple's Metal Performance Shaders (MPS) for GPU acceleration, the method is fully compatible with standard CUDA-based PyTorch. The theoretical complexity analysis is provided in Appendix B.1.

C Evaluation Metrics Details

Supervised protocol. All supervised metrics follow scIB v1.1.5 (Luecken et al., 2022) and are computed in the native embedding space (for Seurat, the integrated CCA embedding produced by `IntegrateLayers`; for Harmony, the corrected PCs): ASW and isolated-label F1 use the embedding directly; graph connectivity uses the standard scanpy nearest-neighbor graph on the embedding; NMI/ARI use Leiden clustering at the scIB-optimal resolution (`cluster_optimal_resolution`). UMAP is used for visualization only.

MDF (Marker Discrimination F-score): Evaluates whether a priori marker genes exhibit distinct expression patterns across clusters identified by KMeans in the native embedding space. For each marker gene m and clustering with K clusters, we compute:

$$F_m = \frac{\sum_{c=1}^{K} n_c(\bar{x}_{cm} - \bar{x}_m)^2}{\sum_{c=1}^{K} \sum_{i \in C_c} (x_{im} - \bar{x}_{cm})^2 + \epsilon} \tag{22}$$

where n_c is the number of cells in cluster c, $\bar{x}_{cm}$ is the mean expression of marker m in cluster c, $\bar{x}_m$ is the overall mean, and x_{im} is the expression in cell i. The per-marker F-statistics are averaged over all markers, cluster numbers ($K \in \{8, 10, 12, 15, 20\}$), and three KMeans seeds; the mean is then normalized as $\bar{F}/(1 + \bar{F})$ (monotone, parameter-free). MDF uses the cell-type marker panel. Marker expression is read from the normalized expression layer; marker panels are fixed a priori and genes absent from the data are reported, never silently dropped.

D HIGHLY VARIABLE GENE SELECTION RESULTS

HVG selection is an optional preprocessing step that determines which genes from the raw count matrix are fed into the module activity calculation and BioNet training. HVGs were selected using the Seurat v3 method (Satija et al., 2016) on raw counts, with a batch-aware stratified fit that falls back to pooled selection when per-batch loess fits hit numerical singularities.

Convention. Table 4 scans HVG cutoffs (2,000–6,000) crossed with teacher PCs (20–100) on ifnb with the current pipeline. HVG=3,000 attains the highest mean Overall (0.7683): larger cutoffs do not help (4,000: 0.7638) and HVG=6,000 clearly degrades performance (0.7404), consistent with the view that low-information genes add noise on immune datasets (Fang et al., 2024); HVG=3,000 is also a common choice in recent integration and foundation-model studies (Cui et al., 2024). The PCs axis is flat (column means within 0.010), so a 40-PC teacher suffices as the supervision signal and there is no need for an excessively large PC count.

Table 4: Overall score of scDeepFusion on ifnb across HVG cutoffs and teacher PCs (current pipeline; single run per cell, seed 42; the main configuration is **bolded**).

HVG	PCs=20	40	60	80	100	mean
2,000	0.773	0.776	0.765	0.764	0.759	0.7676
3,000	0.769	**0.767**	0.766	0.761	0.779	**0.7683**
4,000	0.763	0.767	0.761	0.760	0.768	0.7638
6,000	0.758	0.741	0.733	0.737	0.733	0.7404

Pancreas HVG/PCs scan (current pipeline). Table 18 reports a full scan over HVG cutoffs (3,000 / 6,000 / 12,000 / no filtering) crossed with teacher PCs up to 300 (all single-run). Mean Overall improves with HVG cutoff (0.680 $\rightarrow$ 0.753 $\rightarrow$ 0.777); on iso_f1-free metrics the 3,000$\rightarrow$6,000 gain is real while returns diminish beyond 6,000, the PCs curve plateaus at PCs$\approx$50–150 and declines beyond PCs$\approx$200, and unfiltered input trails HVG-filtered input by 0.04–0.07 across the operating range; the nominally higher Overall values at PCs=100–120 are driven by the degenerate isolated-label F1 and do not persist on iso_f1-free metrics, where Overall is recomputed with Bio restricted to the mean of NMI, ARI, and ASW_label (plateau at 0.870–0.879 from PCs=30 to 150).

E Ablation Studies

We conduct systematic ablation studies on the ifnb dataset (HVG=3,000, PCs=40 teacher, standard scIB protocol) to isolate the contribution of each architectural and objective component (Table 6).

E.1 Component Analysis

We evaluate eight variants: (1) *Full model*: typed-expert routing with the complete objective; (2) *w/o typed routing (concat)*: replacing routing with concatenation followed by an MLP; (3) *w/o module branch*: gene fluctuation input only; (4) *w/o coarse alignment*: removing the teacher-alignment term; (5) *coarse-only*: the alignment term with the auxiliary terms' weights set to zero (the load-balancing term is retained); (6) *recon-only*: reconstruction term only; (7) *w/o denoise*; (8) *distill-only*: a pure distillation objective mode in which cosine similarity of the readout to the teacher is the entire training signal. All variants are single runs (seed 42); the main model and the pathway-MLP control additionally have three-seed runs in Table 1 and Table 13. A pathway-MLP control (MLP on the module-activity matrix with the same teacher) is included for reference.

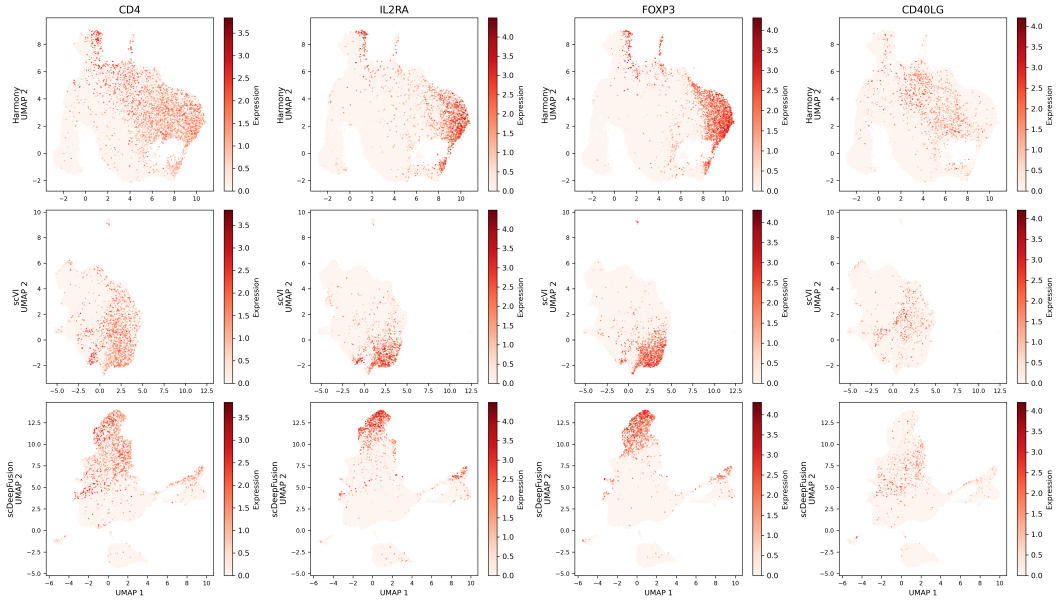
The coarse alignment objective is indispensable. Removing it collapses Overall to 0.543 (w/o coarse) and 0.504 (recon-only), while the coarse-only variant matches the full model (0.767 vs. 0.767). The reconstruction and denoising terms therefore act as regularizers rather than drivers. The distill-only variant attains the highest Overall in the table (0.774), slightly above the full model and the pathway-MLP control; we do not read this as evidence against the full objective, because the gap is within the single-run variation we observe across seeds (the full model's three-seed mean is 0.770 ± 0.003), and the auxiliary terms are what keep the routing weights and the fused (pre-readout) representation well-formed for downstream use.

Typed routing matches simpler fusion on aggregate scores. The concat variant (0.766) and the no-module variant (0.766) tie the full model (0.767), and the pathway-MLP control ties them all (0.772). We therefore do not claim an aggregate-score advantage for typed routing; its value is the per-cell knowledge-source attribution it exposes (Section 3.3). The isolated-label F1 results reinforce this reading: the full model edges concat (0.767 vs. 0.732), but the no-module variant attains the highest value in the table (0.788), so rare-type fidelity is carried by the distilled readout over the gene-fluctuation branch—dropping the module branch does not hurt it—while the module representation's contribution concentrates on the aggregate metrics rather than on rare-type isolation.

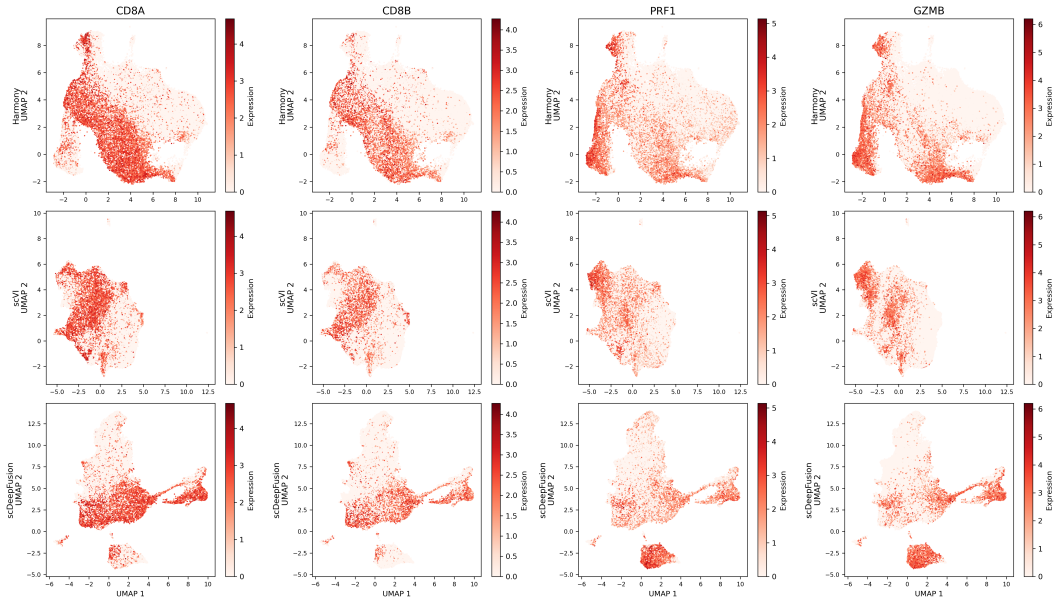
E.2 Hyperparameter Sensitivity

All reported results use fixed hyperparameters without search (Appendix A). A two-round sensitivity analysis over the loss weights (Table 7) shows the objective is insensitive to the exact weights: Overall spans only 0.007 across the $w_{\text{coarse}} \times w_{\text{denoise}}$ grid, with the default configuration at the top. The only sensitive axis is w_{recon}: shifting weight to it from w_{denoise} degrades Overall, corroborating that the reconstruction term acts as a regularizer—consistent with the term-level ablation, where removing the coarse-alignment term collapses performance while the auxiliary terms are dispensable.

F SUPPLEMENTARY FIGURES

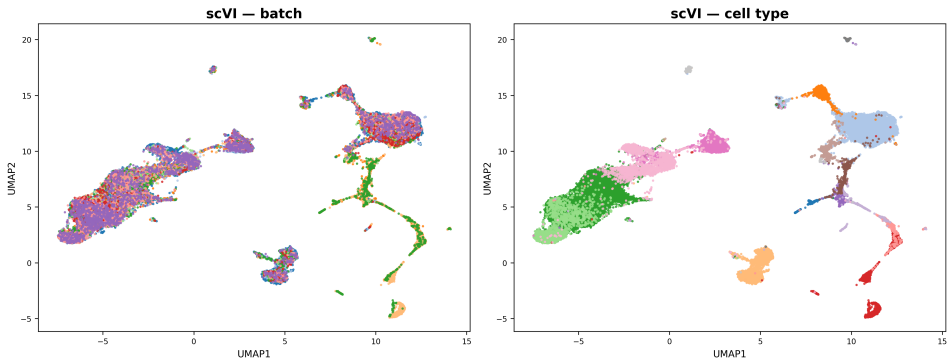


(a) Harmony / scVI / scDeepFusion (top to bottom) on CD4 markers

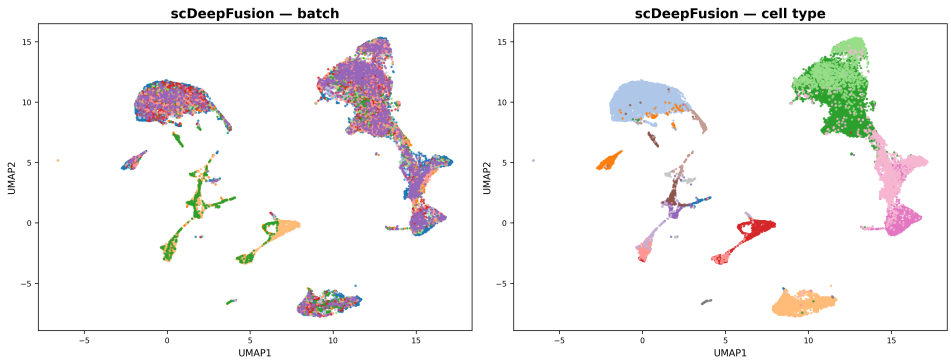


(b) Harmony / scVI / scDeepFusion (top to bottom) on CD8 markers

Figure 3: Marker gene expression patterns on the GSE179994 dataset. Each panel shows four marker genes arranged horizontally, with rows showing Harmony, scVI, and scDeepFusion (top to bottom). $CD4^+$ and $CD8^+$ marker genes exhibit more distinct, clustered expression patterns in scDeepFusion (bottom rows) than in Harmony and scVI (top and middle rows); in particular, the cytotoxic effector $CD8^+$ subpopulation (high *PRF1*/*GZMB*) is distinctly visible in scDeepFusion (b, bottom row) but remains diffuse in the other two (b, top and middle rows).

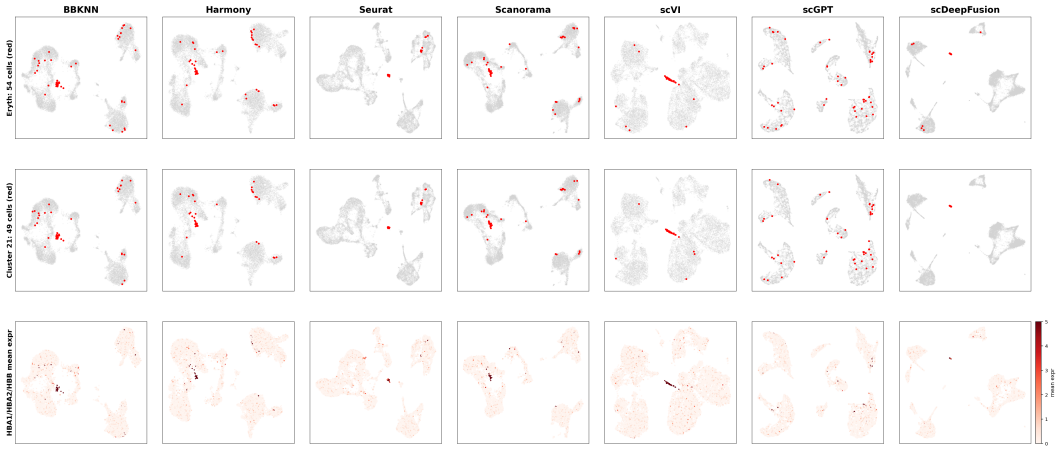


(a) scVI embeddings



(b) scDeepFusion (PCs=40) embeddings

Figure 4: UMAP visualizations of the Immune Human dataset. Comparison of (a) scVI and (b) scDeepFusion embeddings colored by batch labels (left panels) and cell type annotations (right panels, same color per cell type across both methods). scDeepFusion achieves clearer separation of immune cell populations, revealing finer subpopulation structures.



(a) Eryth (54 cells)



(b) Mk (224 cells)

Figure 5: Rare-population consolidation on ifnb. Columns show UMAPs of six baselines and scDeepFusion (rightmost). (a) Erythroid precursors (Eryth, 54 cells): reference annotation (top), scDeepFusion's Leiden cluster 21 (middle), and mean expression of *HBA1*/*HBA2*/*HBB* (bottom). (b) Megakaryocytes (Mk, 224 cells): reference annotation (top), scDeepFusion's Leiden cluster 18 (middle), and mean expression of *PPBP*/*PF4*/*GP9*/*ITGA2B* (bottom). Panels show UMAP visualizations; quantitative metrics are computed in each method's native embedding space. See Section 3.3 for the quantitative analysis.

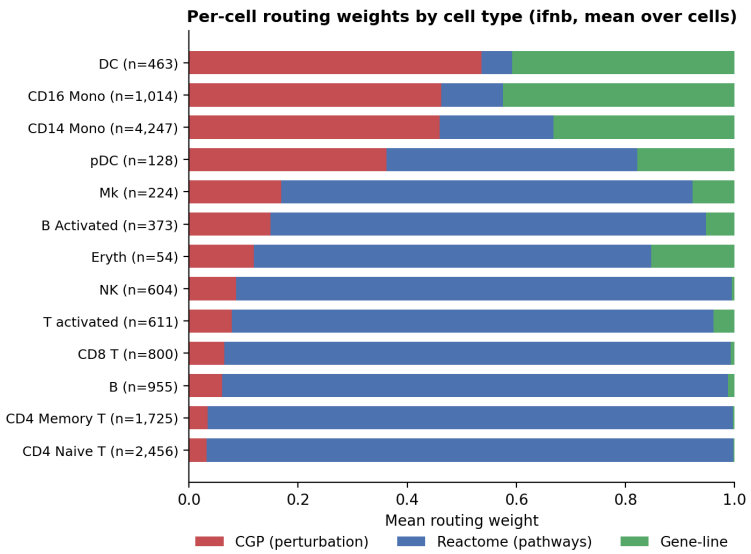


Figure 6: Routing-weight attribution on ifnb. Mean per-cell routing weights of the three typed experts (CGP perturbation signatures, Reactome canonical pathways, and the data-driven gene-line expert) by cell type. Antigen-presenting myeloid cells route primarily to the CGP and gene-line experts, whereas lymphocytes route almost exclusively to the Reactome expert—a myeloid–lymphoid dichotomy that emerges without any lineage supervision in the router input.

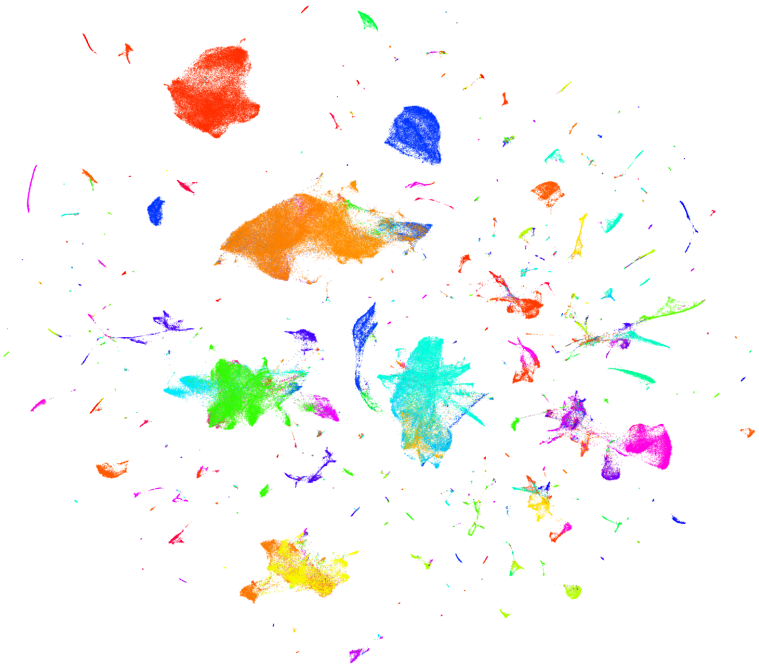


Figure 7: UMAP of the scDeepFusion embedding trained on 400,000 Tabula Sapiens cells (all points shown), colored by cell-ontology class (179 classes present in the 400k subset, of 182 in the full dataset; rare classes are drawn on top). The 400k-cell model yields a coherent atlas embedding with resolved tissue compartments and numerous small, separated rare populations.

G SUPPLEMENTARY TABLES

Table 5: Dataset statistics.

Dataset	Cells	Genes	Batches	Types	Description
ifnb	13,654	14,053	8	13	Interferon-stimulated PBMCs. Flagship dataset for ablations with manageable batch complexity.
GSE179994	150,849	17,277	36	2	$CD4^+$/$CD8^+$ T cell subsets (coarse labels). Tests fine-grained subtype discrimination beyond label reliability.
Immune Human	32,484	12,303	9	16	Human immune atlas (bone marrow + PBMCs). Tests multi-batch generalization across broad immune categories.
Immune (H+M)	96,839	8,135	22	19	Cross-species immune data (mouse mapped to human genes). Tests conserved signatures across species.
Pancreas	16,382	19,093	9	14	Human pancreas (scIB). Validates generalization beyond immune cells and tests organ-level cell-type discrimination.
Tabula Sapiens	1,136,218	61,806	46	182	Multi-tissue atlas. Exploratory analysis of method boundaries under cross-tissue heterogeneity.

Table 6: Ablation study on the ifnb dataset (HVG=3,000, PCs=40 teacher, standard scIB protocol, native embedding space). All variants share the same data pipeline, teacher, and training hyperparameters. Best scores in **bold**, second best underlined (marks by unrounded values).

	Biological Conservation				Batch Correction		Composite Scores		
Variant	NMI	ARI	ASW label	Iso F1	ASW batch	Graph Conn	Bio	Batch	Overall
Full model (typed routing)	0.717	0.516	0.567	0.767	0.919	0.993	0.641	0.956	0.767
w/o typed routing (concat)	0.720	0.522	0.571	0.732	0.926	**0.995**	0.636	0.960	0.766
w/o module branch	0.716	0.516	**0.583**	**0.788**	0.900	0.978	**0.651**	0.939	0.766
w/o coarse alignment	0.310	0.117	0.498	0.370	0.874	0.871	0.323	0.872	0.543
coarse-only (w/o recon+denoise)	0.720	0.518	0.569	0.756	0.922	0.990	0.641	0.956	0.767
recon-only (w/o coarse+denoise)	0.280	0.097	0.497	0.277	0.850	0.805	0.288	0.827	0.504
w/o denoise	0.696	0.510	0.574	0.744	0.902	0.953	0.631	0.927	0.750
distill-only objective	0.725	0.533	0.571	0.774	0.923	0.993	0.651	0.958	**0.774**
pathway-MLP (control)	**0.728**	**0.534**	0.569	0.755	**0.926**	0.995	0.647	**0.961**	0.772

Table 7: Loss-weight sensitivity on ifnb (current architecture; three-seed mean Overall, scIB v1.1.5 protocol, 30 training runs over two rounds). Grid: $w_{\text{coarse}} \times w_{\text{denoise}}$ with w_{recon}=0.2 fixed; the default cell (0.770 ± 0.003) coincides with the main configuration. Bottom: shifting weight from w_{denoise} to w_{recon} (w_{coarse}=0.5).

	w_{denoise}		
w_{coarse}	0.2	0.3	0.4
0.4	0.763	0.770	0.767
0.5	0.769	0.770	0.767
0.6	0.764	0.767	0.767

Shift weight from $w_{denoise}$ to w_{recon} (w_{coarse}=0.5):
w_{recon} 0.3 → 0.762; w_{recon} 0.4 → 0.763

Table 8: Pathway-scoring baselines on ifnb (standard scIB protocol, native embedding space). Bare = the pathway-score matrix used directly as the embedding; distilled = an MLP readout trained on the scores with the same teacher as scDeepFusion. Bare rows and AUCell/GSVA/UCell distilled rows are single runs; the module-activity distilled row reports the three-seed mean (as in Table 1). After the distilled readout, all module-level representations tie (Overall 0.768–0.770), indicating that the choice of module-scoring formula is not the source of scDeepFusion's performance. Note that pathway scores are batch-invariant by construction, inflating ASW_batch.

		Biological Conservation				**Batch Correction**		**Composite Scores**		
Method		NMI	ARI	ASW label	Iso F1	ASW batch	Graph Conn	Bio	Batch	Overall
AUCell	bare	0.568	0.407	0.512	0.526	0.970	0.899	0.503	0.934	0.676
GSVA	bare	0.485	0.281	0.507	0.421	0.974	0.846	0.423	0.910	0.618
UCell	bare	0.476	0.325	0.504	0.479	0.974	0.827	0.446	0.900	0.628
module activity (ours)	bare	0.270	0.078	0.497	0.245	0.877	0.671	0.272	0.774	0.473
AUCell	distilled	0.714	0.524	0.573	0.750	0.926	0.996	0.640	0.961	0.768
GSVA	distilled	0.716	0.513	0.572	0.752	0.928	0.996	0.638	0.962	0.768
UCell	distilled	0.712	0.523	0.572	0.767	0.927	0.995	0.643	0.961	0.770
module activity (ours)	distilled	0.724	0.532	0.571	0.750	0.925	0.988	0.644	0.957	0.769

Table 9: Label-free marker-anchored evaluation on ifnb (MDF, native-space KMeans, 5 cluster numbers $\times$ 3 seeds; a priori 43-gene cell-type marker panel; three-seed mean for scDeepFusion). The pathway-MLP control scores 0.411 on this metric, comparable to scDeepFusion. Best in **bold**.

Method	**MDF**
BBKNN	0.374
Harmony	0.377
Seurat(V5)	0.399
Scanorama	0.375
scVI	0.385
scGPT (fine-tuned)	0.388
scDeepFusion (Ours)	**0.410**

Table 10: Per-population isolation metrics for the two case-study populations on ifnb (per-label isolated-label F1 under the scIB protocol, and 15-nearest-neighbor label purity in each native embedding). Best per column in **bold**.

	isolated-label F1		**15-NN purity**	
Method	**Eryth**	**Mk**	**Eryth**	**Mk**
BBKNN	0.500	0.075	0.389	0.027
Harmony	0.659	0.079	0.519	0.152
Seurat(V5)	0.904	**0.704**	**0.951**	0.589
Scanorama	0.699	0.065	0.489	0.087
scVI	0.941	0.390	0.785	0.263
scGPT (fine-tuned)	0.057	0.086	0.037	0.059
scDeepFusion (Ours)	**0.942**	0.663	0.877	**0.701**

Table 11: Expert-masking selectivity on ifnb: mean cosine shift of the fused representation $\mathbf{u}_i$ when one expert's weight is set to zero and the remaining weights are renormalized (recomputed from the trained checkpoint; per-cell values averaged within each cell type).

Cell type	mask CGP	mask Reactome
DC	0.129	0.002
pDC	0.129	0.042
CD16 Mono	0.105	0.005
CD14 Mono	0.103	0.007
B Activated	0.079	0.184
NK	0.060	0.200
Mk	0.056	0.213
CD8 T	0.043	0.326
B	0.043	0.342
T activated	0.036	0.274
Eryth	0.031	0.082
CD4 Memory T	0.022	0.460
CD4 Naive T	0.020	0.476

Table 12: Mean per-cell routing weights by cell type on the Immune Human atlas (no stimulation condition), showing the same myeloid–lymphoid dichotomy as on ifnb (Figure 6).

Cell type	CGP	Reactome	Gene-line
Monocyte-derived DC	0.551	0.033	0.416
$CD16^+$ Monocytes	0.537	0.066	0.397
Erythroid progenitors	0.512	0.122	0.366
Megakaryocyte progenitors	0.476	0.158	0.366
HSPCs	0.439	0.314	0.247
$CD14^+$ Monocytes	0.419	0.321	0.261
$CD20^+$ B cells	0.154	0.615	0.232
NK cells	0.154	0.619	0.227
NKT cells	0.098	0.666	0.236
$CD8^+$ T cells	0.073	0.716	0.210
$CD4^+$ T cells	0.063	0.726	0.211

Table 13: Multi-seed stability validation on the ifnb dataset (HVG=3,000, PCs=40 teacher). Three independent runs with different random seeds; mean±std reported.

Seed	Biological Conservation				Batch Correction		Composite Scores		
	NMI	ARI	ASW label	Iso F1	ASW batch	Graph Conn	Bio	Batch	Overall
42	0.717	0.516	0.567	0.767	0.919	0.993	0.641	0.956	0.767
43	0.723	0.519	0.566	0.761	0.927	0.992	0.642	0.959	0.769
44	0.720	0.531	0.567	0.771	0.930	0.993	0.647	0.961	0.773
Mean±Std	0.720 ±0.003	0.522 ±0.008	0.567 ±0.001	0.766 ±0.005	0.925 ±0.006	0.993 ±0.001	**0.644 ±0.003**	**0.959 ±0.003**	**0.770 ±0.003**

Table 14: Performance comparison on the GSE179994 dataset (150,849 cells, 36 donor batches; standard scIB v1.1.5 protocol). Supervised metrics use the coarse two-class labels (CD4/CD8; cells without labels excluded), which cannot test subtype resolution—treat as reference only. scDeepFusion uses HVG=3,000, PCs=40 teacher. Best scores in **bold**, second best underlined (marks by unrounded values).

Method	Biological Conservation				Batch Correction		Composite Scores		
	NMI	ARI	ASW label	Iso F1	ASW batch	Graph Conn	Bio	Batch	Overall
BBKNN	0.252	0.073	**0.636**	0.216	0.648	0.952	0.294	0.800	0.497
Harmony	**0.726**	**0.796**	0.549	**0.946**	0.889	0.999	**0.754**	0.944	**0.830**
Seurat(V5)	0.329	0.306	0.555	0.687	0.866	0.997	0.469	0.931	0.654
Scanorama	0.458	0.527	0.540	0.838	0.926	**0.999**	0.591	0.963	0.740
scVI	0.479	0.548	0.518	0.844	**0.951**	0.999	0.597	**0.975**	0.749
scGPT (fine-tuned)	0.268	0.080	0.528	0.562	0.926	0.999	0.360	0.962	0.601
pathway-MLP (control)	0.546	0.621	0.562	0.884	0.885	0.998	0.653	0.941	0.769
scDeepFusion (Ours)	0.575	0.662	0.555	0.896	0.863	0.998	0.672	0.931	0.776

Table 15: Unsupervised subtype-marker evaluation on GSE179994 (native-space KMeans, 5 cluster numbers $\times$ 3 seeds; values are means over the 15 runs. The a priori panel spans CD4/CD8 lineages and functional T-cell states (naive, memory, Treg, cytotoxic effector, exhaustion, proliferation; 24 of 28 panel genes available in the HVG subset, reported not dropped). Run-level spread for the top two methods: scDeepFusion 0.301 ± 0.023, Seurat 0.276 ± 0.014 (mean $\pm$ std). Best MDF in **bold**.

Method	MDF
BBKNN	0.276
Harmony	0.270
Seurat(V5)	0.276
Scanorama	0.257
scVI	0.270
scGPT (fine-tuned)	0.215
pathway-MLP (control)	0.300
scDeepFusion (Ours)	**0.301**

Table 16: Performance comparison on the Immune Human dataset (32,484 cells, 9 batches, 16 cell types; standard scIB v1.1.5 protocol, native embedding space). scDeepFusion uses HVG=3,000, PCs=40 teacher. Best scores in **bold**, second best underlined (marks by unrounded values).

Method	Biological Conservation				Batch Correction		Composite Scores		
	NMI	ARI	ASW label	Iso F1	ASW batch	Graph Conn	Bio	Batch	Overall
BBKNN	0.640	0.248	0.501	0.799	0.639	0.706	0.547	0.672	0.597
Harmony	0.793	**0.761**	0.572	0.713	0.893	**0.988**	0.710	0.940	0.802
Seurat(V5)	0.705	0.578	0.537	0.433	0.842	0.919	0.563	0.881	0.690
Scanorama	0.646	0.330	0.554	0.843	0.855	0.930	0.593	0.892	0.713
scVI	0.766	0.625	0.558	0.815	**0.899**	0.988	0.691	**0.943**	0.792
scGPT (fine-tuned)	0.755	0.608	0.598	**0.858**	0.814	0.984	0.704	0.899	0.782
pathway-MLP (control)	0.797	0.750	**0.603**	0.838	0.867	0.974	**0.747**	0.920	0.816
scDeepFusion (Ours)	**0.802**	0.758	0.596	0.814	0.888	0.986	0.742	0.937	**0.820**

Table 17: Performance comparison on the Immune Human+Mouse dataset (96,839 cells, 22 batches, 19 cell types; standard scIB v1.1.5 protocol, native embedding space). scDeepFusion uses HVG=3,000, PCs=40 teacher. Best scores in **bold**, second best underlined (marks by unrounded values).

Method	Biological Conservation				Batch Correction		Composite Scores		
	NMI	ARI	ASW label	Iso F1	ASW batch	Graph Conn	Bio	Batch	Overall
BBKNN	0.569	0.176	0.374	**0.408**	0.670	0.545	0.382	0.607	0.472
Harmony	0.610	0.398	0.469	0.140	0.901	0.937	0.404	0.919	0.610
Seurat(V5)	0.274	0.142	0.426	0.023	0.843	0.812	0.216	0.828	0.461
Scanorama	0.616	0.360	0.488	0.253	0.881	0.831	0.429	0.856	0.600
scVI	0.632	0.446	0.487	0.053	**0.937**	**0.960**	0.404	**0.949**	0.622
scGPT (fine-tuned)	0.616	0.364	**0.521**	0.328	0.858	0.883	0.458	0.871	0.623
pathway-MLP (control)	**0.639**	**0.527**	0.448	0.243	0.880	0.949	**0.464**	0.915	**0.644**
scDeepFusion (Ours)	0.636	0.451	0.441	0.259	0.896	0.949	0.447	0.922	0.637

Table 18: Performance comparison on the Pancreas dataset (16,382 cells, 19,093 genes, 9 batches, 14 cell types; standard scIB v1.1.5 protocol, native embedding space). Baselines are re-evaluated under their default configurations. scDeepFusion rows scan HVG cutoffs and teacher PCs; all rows report single-run results for direct comparability. scGPT retains its pretrained architecture (1,200 genes, 512-dim). Best scores in **bold**.

	Biological Conservation				**Batch Correction**		**Composite Scores**		
Method	NMI	ARI	ASW_label	Iso_F1	ASW_batch	Graph_Conn	Bio	Batch	Overall
Baseline methods (default configuration, standard protocol)									
BBKNN	0.625	0.290	0.478	0.125	0.575	0.620	0.379	0.598	0.467
Harmony	0.897	0.917	0.630	0.107	0.905	0.970	0.638	0.937	0.758
Scanorama	0.622	0.238	0.558	0.111	0.821	0.838	0.382	0.830	0.561
Seurat	0.837	0.872	0.636	0.083	0.855	0.978	0.607	0.916	0.731
scGPT (fine-tuned)	0.816	0.700	0.708	0.212	0.842	0.964	0.609	0.903	0.727
scVI	0.912	0.946	0.617	0.101	0.859	0.980	0.644	0.920	0.754
scDeepFusion, HVG=3,000									
PCs=10	0.730	0.726	0.618	0.024	0.815	0.832	0.525	0.824	0.644
PCs=15	0.718	0.697	0.583	0.033	0.846	0.842	0.507	0.844	0.642
PCs=20	0.763	0.749	0.622	0.160	0.845	0.924	0.573	0.885	0.698
PCs=25	0.768	0.751	0.617	0.133	0.846	0.897	0.567	0.872	0.689
PCs=30	0.768	0.750	0.607	0.140	0.850	0.914	0.566	0.882	0.693
PCs=35	0.790	0.771	0.604	0.138	0.862	0.912	0.576	0.887	0.700
PCs=40	0.789	0.769	0.609	0.103	0.867	0.915	0.568	0.891	0.697
scDeepFusion, HVG=6,000									
PCs=20	0.904	0.938	0.731	0.093	0.809	0.941	0.667	0.875	0.750
PCs=25	0.903	0.937	0.716	0.079	0.836	0.944	0.659	0.890	0.751
PCs=30	0.899	0.938	0.686	0.079	0.857	0.961	0.651	0.909	0.754
PCs=35	0.906	0.940	0.685	0.091	0.827	0.986	0.656	0.906	0.756
PCs=40	0.844	0.826	0.649	0.092	0.847	0.972	0.603	0.910	0.725
PCs=45	0.903	0.937	0.684	0.192	0.859	0.989	0.679	0.924	0.777
PCs=50	0.907	0.940	0.661	0.092	0.849	0.988	0.650	0.919	0.757
scDeepFusion, HVG=12,000									
PCs=30	0.905	0.941	0.709	0.091	0.840	0.968	0.662	0.904	0.758
PCs=40	0.907	0.944	0.686	0.090	0.855	0.982	0.657	0.919	0.762
PCs=50	0.909	0.943	0.690	0.092	0.869	0.986	0.658	0.927	0.766
PCs=60	0.901	0.937	0.679	0.127	0.879	0.985	0.661	0.932	0.770
PCs=70	0.905	0.939	0.668	0.124	0.882	0.986	0.659	0.934	0.769
PCs=80	0.897	0.936	0.664	0.127	0.894	0.986	0.656	0.940	0.770
PCs=90	0.906	0.937	0.657	0.462	0.898	0.968	0.740	0.933	0.817
PCs=100	0.905	0.941	0.650	0.500	0.874	0.981	0.749	0.927	0.820
PCs=120	0.905	0.940	0.637	0.519	0.881	0.987	**0.750**	**0.934**	**0.824**
PCs=150	0.903	0.941	0.646	0.130	0.891	0.987	0.655	0.939	0.769
PCs=200	0.877	0.902	0.625	0.101	0.894	0.985	0.626	0.940	0.752
PCs=250	0.900	0.937	0.626	0.131	0.887	0.984	0.648	0.936	0.763
PCs=300	0.895	0.931	0.629	0.194	0.877	0.985	0.662	0.931	0.770
scDeepFusion, no HVG filter (all genes)									
PCs=30	0.838	0.812	0.649	0.237	0.855	0.959	0.634	0.907	0.743
PCs=40	0.824	0.789	0.619	0.237	0.824	0.964	0.617	0.894	0.728
PCs=60	0.852	0.819	0.621	0.538	0.889	0.973	0.708	0.931	0.797
PCs=80	0.845	0.816	0.628	0.538	0.865	0.980	0.707	0.922	0.793
PCs=100	0.855	0.820	0.617	0.024	0.890	0.985	0.579	0.938	0.723
PCs=120	0.857	0.821	0.626	0.169	0.884	0.968	0.618	0.926	0.741
PCs=150	0.853	0.896	0.591	0.200	0.871	0.974	0.635	0.923	0.750
PCs=200	0.896	0.932	0.629	0.200	0.915	0.985	0.664	0.950	0.779
PCs=250	0.872	0.897	0.638	0.200	0.898	0.980	0.652	0.939	0.767
PCs=300	0.883	0.900	0.621	0.128	0.890	0.981	0.633	0.935	0.754

Table 19: Scalability evaluation on Tabula Sapiens (>1.1M cells, 182 cell-ontology types; HVG=12,000 as on the tissue-level Pancreas dataset; non-protein-coding genes removed, 30,564 retained). Full scIB evaluation is computationally infeasible at atlas scale: silhouette-based metrics require pairwise distances (quadratic in cell count) and isolated-label F1 requires per-label neighbor graphs and re-clustering. Quality is therefore reported under a reduced protocol (exact kNN graph; NMI, ARI, graph connectivity, and ASW on a fixed 5,000-cell subsample with seed 42; no isolated-label F1; Overall = 0.6 Bio + 0.4 Batch). Training time and peak memory are measured end-to-end at each scale for both methods. scVI's converged runs are faster and lighter per run, but under its library-default configuration it succeeds in only one of three seeds at 100k (its 100k quality value is from that single converged run) and its failed attempts at 400k consume 8,439 s without producing a model (latent-posterior NaN); per-method stabilization such as gradient clipping or learning-rate reduction was not attempted, to keep all method configurations symmetric. scDeepFusion completes every run, and its higher memory is dominated by the dense module-activity matrix (Appendix B.1), which sparse or streaming computation should address; preprocessing peak memory likewise grows from 48 GB at 100k to 292 GB at 400k and exceeds available memory at 800k, so we claim support up to the demonstrated 400k scale.

	scDeepFusion cost		**scVI cost**		**Quality (reduced)**	
Cells	**Time (s)**	**Peak (GB)**	**Time (s)**	**Peak (GB)**	**scDeepFusion**	**scVI**
100k	2,393	75	1,253	23	**0.676**	0.554
200k	3,975	224	2,779	43	**0.680**	0.543
400k	8,174	411	8,439	81	**0.668**	diverged
800k	preprocessing exceeds available memory (training not attempted)					

REPRODUCIBILITY STATEMENT

All reported results are produced by an end-to-end pipeline with fixed configurations and recorded random seeds; multi-seed runs (three seeds) are reported for trainable variants on the flagship ifnb dataset, and all cross-dataset results use single runs for direct comparability with the re-run baselines. Module libraries are derived from MSigDB C2 v2026.1 by a deterministic conversion script included in the release. All baselines were re-run locally under identical data splits and evaluation protocols (scIB v1.1.5; fine-tuned scGPT), and per-step runtime and peak-memory logs are recorded for every pipeline step. An anonymized implementation, including all training and evaluation scripts, is available at `https://anonymous.4open.science/r/scdeepfusion-2026/`.

AI USE STATEMENT

Generative AI tools were used to assist with code debugging, language polishing, and formatting of this manuscript. All AI-assisted content was independently verified by the authors, who take full responsibility for the entire submission. No experimental results, figures, or citations were generated by AI without independent verification.