

# SOLAR: Fixed orthogonal label anchors for single-cell reference mapping

**Minqi Chen**

West China Hospital of Sichuan University

**Guanjie Zou**

The University of Tokyo

**Minrui Chen**

Kyushu University

**Hiroto Saigo**

Kyushu University

**Koji Tsuda**

The University of Tokyo

**Lunxu Liu**

liulunxu@wchscu.cn

West China Hospital of Sichuan University

---

## Research Article

**Keywords:** Single-cell RNA sequencing, single-cell integration, supervised contrastive learning, label anchors, reference mapping, label transfer, batch effects

**Posted Date:** October 19th, 2026

**DOI:** <https://doi.org/10.21203/rs.3.rs-11248498/v1>

**License:** © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

**Additional Declarations:** No competing interests reported.

---

# SOLAR: Fixed orthogonal label anchors for single-cell reference mapping

Minqi Chen<sup>1,2†</sup>, Guanjie Zou<sup>4,5†</sup>, Minrui Chen<sup>3†</sup>,  
Hiroto Saigo<sup>3,5\*</sup>, Koji Tsuda<sup>4,5\*</sup>, Lunxu Liu<sup>1,2\*</sup>

<sup>1</sup>Department of Thoracic Surgery and Institute of Thoracic Oncology,  
West China Hospital, Sichuan University, Chengdu, Sichuan, 610041,  
China.

<sup>2</sup>Western China Collaborative Innovation Center for Early Diagnosis  
and Multidisciplinary Therapy of Lung Cancer, Sichuan University,  
Chengdu, Sichuan, 610041, China.

<sup>3</sup>Department of Information Science and Technology, Kyushu  
University, Motooka, Nishi-ku, Fukuoka, 819-0395, Japan.

<sup>4</sup>Department of Computational Biology and Medical Sciences,  
Graduate School of Frontier Sciences, The University of Tokyo, 5-1-5  
Kashiwanoha, Kashiwa, Chiba, 277-8563, Japan.

<sup>5</sup>Center for Advanced Intelligence Project (AIP), RIKEN, Tokyo,  
103-0027, Japan.

\*Corresponding author(s). E-mail(s): [saigo@inf.kyushu-u.ac.jp](mailto:saigo@inf.kyushu-u.ac.jp);  
[tsuda@k.u-tokyo.ac.jp](mailto:tsuda@k.u-tokyo.ac.jp); [liulunxu@wchscu.cn](mailto:liulunxu@wchscu.cn);

<sup>†</sup>These authors contributed equally to this work.

## Abstract

**Background:** Annotated single-cell references provide labels that can guide both query annotation and the organization of learned cell representations. We introduce Supervised Orthogonal Label-Anchor Representation (SOLAR), a supervised contrastive method that assigns each reference label a fixed orthogonal direction before training and learns how expression profiles map into that space.

**Results:** SOLAR trains the encoder using cell embeddings and label anchors in a shared contrastive objective, then maps query cells with the fitted reference preprocessing and trained encoder without further optimization. Across nine held-out batches from human immune, lung, and pancreas datasets, SOLAR achieved a mean query macro-F1 of 0.811, compared with 0.774 for matched principal component analysis (Matched PCA), 0.815 for the supervised contrastive comparator SCLSC, and 0.838 for the annotation-aware variational model scANVI, at a common output dimension of 30. Label average silhouette width increased from 0.574 for Matched PCA to 0.873 for SOLAR. At 128 dimensions, anchor controls showed that repetition mainly increased label separation and local batch diversity, with little change in query-label recovery and lower local-structure scores.

**Conclusions:** SOLAR provides a reference-mapping method in which fixed label anchors define the training references while the encoder learns cell positions from expression profiles. This construction makes the label relationships explicit in the output space and produces a reusable encoder for mapping new cells.

**Keywords:** Single-cell RNA sequencing, single-cell integration, supervised contrastive learning, label anchors, reference mapping, label transfer, batch effects

## 1 Background

Annotated single-cell atlases pair expression profiles with curated cell-type labels and can therefore support the mapping and annotation of new experiments. Reference labels can guide how a representation is learned as well as how query cells are annotated. When training uses one representative per class, those representatives may be obtained from expression profiles or placed directly in the output space; the two constructions give labels different roles in the objective.

Supervised contrastive learning on cells and their types (SCLSC), for example, constructs each type representative from a mean expression profile and passes both cells and representatives through a trainable encoder [1]. Their positions in the learned space therefore change during training. We introduce Supervised Orthogonal Label-Anchor Representation (SOLAR), a supervised contrastive method that instead assigns each reference label a fixed orthogonal direction before training.

The fixed anchors and normalized cell embeddings share one contrastive pool, so the anchors retain their prescribed relationships while the encoder learns where cells lie in that space. Once trained, the encoder maps query cells using the preprocessing fitted to the reference. We evaluate query-label prediction, label separation, batch diversity, and local structure, complementary aspects used in single-cell integration evaluation [2–5]. Cross-method comparisons at a common output dimension evaluate the complete mapping method, whereas repeated, unique, and absent-anchor controls at 128 dimensions examine how anchor presence and multiplicity affect the learned representation.

## 2 Related work

### *Reference projection and reusable encoders.*

Annotation-transfer methods such as scmap, Garnett, and scPred use an existing reference to assign cell identities [6–8]. Alignment methods instead identify shared structure across datasets [9–11]. Symphony provides a reference-building and query-mapping framework [12], while the single-cell variational inference model scVI and its annotation-aware extension scANVI learn probabilistic representations of single-cell expression [13, 14]. Transfer-learning frameworks such as scArches allow query adaptation to a learned reference [15]. These approaches differ in their use of labels, expression likelihoods, and query data. Here, scVI and scANVI are evaluated as frozen-reference encoders, with reference-trained parameters fixed during query encoding.

### *Supervised contrastive learning and class representatives.*

Supervised contrastive learning organizes examples through same-label and different-label relationships [16]; related single-cell approaches use contrastive representations for annotation and integration [17, 18]. Supervised contrastive learning on cells and their types (SCLSC) is the closest comparator because it explicitly contrasts cells with

139 their types [1]. It constructs each type representative as a mean expression profile  
 140 and passes both cells and these data-dependent representatives through a trainable  
 141 encoder, so their positions in the learned space change during optimization. SOLAR  
 142 instead prescribes expression-independent orthogonal label directions directly in the  
 143 output space and keeps them fixed while learning cell embeddings. Both methods  
 144 use reference labels and supervised contrastive training; their input and training  
 145 differences are specified in Methods and Additional file 1.

### 151 *Proxy and orthogonality-based metric learning.*

152 Proxy Anchor loss replaces many sample-pair comparisons with learnable class proxies  
 153 [19], and OrCo uses orthogonality and contrast in few-shot class-incremental learn-  
 154 ing [20]. OrCo adds fixed, data-independent pseudo-targets and their perturbations as  
 155 positives in supervised contrastive learning. SOLAR applies fixed data-independent  
 156 label directions directly to supplied single-cell labels, without OrCo’s image pre-  
 157 training, incremental class alignment, target perturbations, or reservation of future  
 158 targets. The experiments compare repeated, unique, and absent anchors to character-  
 159 ize how anchor presence and multiplicity affect the learned representation in single-cell  
 160 reference mapping.

## 161 **3 Methods**

162 Figure 1 summarizes the two computational stages: reference-only preprocessing and  
 163 encoder training, followed by direct query mapping with the fitted transformations  
 164 and frozen encoder. Its joint-space panel illustrates the intended label organization; it  
 165 is a schematic, not an additional query-training step.

### 166 **3.1 Study design and reference–query protocols**

167 We represent a dataset as  $\mathcal{D} = \{(x_i, y_i, b_i)\}_{i=1}^{n_{\mathcal{D}}}$ , where  $x_i$  is the expression profile,  $y_i$   
 168 the curated cell-type label, and  $b_i$  the experimental batch. Each split defines a labelled

reference set  $\mathcal{R}$  and a query set  $\mathcal{Q}$ . SOLAR learns an encoder  $f_\theta$  from the reference and normalizes its output as follows.

$$z_i = \frac{f_\theta(x_i)}{\|f_\theta(x_i)\|_2}. \quad (1)$$

The normalized embedding in Eq. (1) is the learned representation used for mapping and evaluation. Biological labels guide reference training; batch identities enter reference-only preprocessing and evaluation, but not the SOLAR loss.

The study uses five processed single-cell integration benchmarking (scIB) datasets [4, 21]. Human immune cells, lung atlas, and pancreas form the primary held-out comparison; two simulations provide supplementary checks (Table 1). In the within-batch protocol, batch-label-stratified partitions assign 80% of each stratum to the reference and 20% to the query, using split seeds 40–44. A one-cell stratum remains in the reference. In the held-out-batch protocol, one complete batch forms the query and all remaining batches form the reference. Batch identity, not its biological labels, determines this split. Three batches per real dataset were selected to span query sizes, giving nine primary query batches.

All evaluated query labels occur in the corresponding reference; the study therefore tests closed-label mapping. The held-out batch is the generalization unit. Seeds 40–44 describe repeated stochastic fits on the same split, not independent biological replicates. After the reference and query roles are assigned, query labels are excluded from feature selection, principal-component fitting, training, internal validation, checkpoint selection, and mapping. They are used only for evaluation. Reference labels remain supervised inputs.

**Table 1** Datasets and evaluation roles. The primary held-out comparison uses three real datasets with three query batches each; the simulations provide supplementary checks. Table S1 in Additional file 1 gives complete batch composition.

| Dataset            | Cells  | Labels | Held-out batches | Evaluation role                     |
|--------------------|--------|--------|------------------|-------------------------------------|
| Human immune cells | 33,506 | 16     | 3                | Primary mapping and local structure |
| Lung atlas         | 32,472 | 17     | 3                | Primary mapping and local structure |
| Human pancreas     | 16,382 | 14     | 3                | Primary mapping                     |
| Simulation 1       | 12,097 | 7      | 3                | Supplementary check                 |
| Simulation 2       | 19,318 | 4      | 3                | Supplementary check                 |

The primary held-out-batch comparison comprises nine real batches. Complete batch composition is provided in Additional file 1, Table S1. All tables and figures prefixed S below refer to Additional file 1.

### 3.2 Reference-only preprocessing

For each split, the Cell Ranger-style procedure in Scanpy selects 2,000 highly variable genes from reference cells, using reference batch as the batch key [22]. Principal-component analysis (PCA) is then fitted to the reference, and the resulting transform is applied unchanged to query cells. SOLAR receives 40 principal components in both evaluated configurations. The five-method comparison uses a 30-dimensional output, whereas the anchor controls use a 128-dimensional output. The baseline without a learned encoder is termed Matched PCA: it applies the same reference-only gene-selection procedure and a reference-fitted PCA transform, using an independently fitted 30-component transform in the common-dimensional comparison. A 40-component Matched PCA representation is reported in the method-specific supplementary summaries.

### 3.3 Expression encoder and fixed label anchors

The encoder is a three-layer multilayer perceptron with two 512-unit hidden layers, batch normalization, rectified linear units, dropout of 0.1, a linear output projection, layer normalization, and final  $\ell_2$  normalization [23, 24]. The output layer is 30-dimensional in the five-method comparison and 128-dimensional in the anchor analysis. The encoder has no reconstruction decoder, batch-adversarial component, or classification head.

Let  $K$  denote the number of reference labels and  $d \geq K$  the output dimension.  
A Gaussian matrix  $G \in \mathbb{R}^{d \times K}$  is sampled once and a reduced orthogonal-triangular  
(QR) factorization is computed, written here as follows.

$$G = UT, \quad U^\top U = I_K. \quad (2)$$

In Eq. (2), the columns of  $U \in \mathbb{R}^{d \times K}$  are orthonormal, and  $T \in \mathbb{R}^{K \times K}$  is upper  
triangular. Column  $u_k$  is the unit-length anchor for label  $k$ . Thus, for  $k \neq l$ ,  $u_k^\top u_l = 0$   
and  $\|u_k - u_l\|_2 = \sqrt{2}$ : all distinct label directions have the same pairwise angle and  
Euclidean distance, specified without expression means. Anchors and the normalized  
embeddings in Eq. (1) share the unit sphere, but cell coordinates remain learned. The  
anchor bank is fixed during training and is not used during query mapping.

### 3.4 Anchor-augmented supervised contrastive objective

SOLAR extends supervised contrastive learning [16] by placing expression embeddings  
and label anchors in one indexed pool. For a mini-batch of  $B$  reference cells, the default  
repeated-anchor pool is defined as follows.

$$q_a = \begin{cases} z_a, & 1 \leq a \leq B, \\ u_{y_{a-B}}, & B < a \leq 2B, \end{cases} \quad \tilde{y}_a = \begin{cases} y_a, & 1 \leq a \leq B, \\ y_{a-B}, & B < a \leq 2B. \end{cases} \quad (3)$$

In Eq. (3), each cell contributes an expression view and one indexed copy of its class  
anchor, giving  $V = 2B$  views. The unique-anchor control contributes one anchor for  
each of the  $K_B$  labels represented in the mini-batch, giving  $V = B + K_B$ . The no-  
anchor control uses only the  $B$  expression views, without a second augmented view.  
In all three cases,  $\tilde{y}_a$  denotes the label of view  $a$ .

For a focal view  $a$ , define the non-self candidate indices as  $C(a) = \{1, \dots, V\} \setminus \{a\}$  and the positive indices as  $\text{Pos}(a) = \{j \in C(a) : \tilde{y}_j = \tilde{y}_a\}$ . Self-exclusion is by index, not by vector value: identical copies of a repeated anchor remain distinct views. The logit for  $b \in C(a)$  is

$$\ell_{ab} = \frac{q_a^\top q_b - m \mathbf{1}[\tilde{y}_a = \tilde{y}_b]}{\tau}. \quad (4)$$

where  $\mathbf{1}[\cdot]$  equals 1 when its condition is true and 0 otherwise. All views have unit length, so the inner product in Eq. (4) is cosine similarity. Temperature  $\tau > 0$  controls softmax sharpness: smaller values amplify logit differences. Margin  $m$  subtracts an additive cosine offset from same-label logits before temperature scaling. It affects those logits in both the numerator and denominator of the loss; it is not an angular margin of the ArcFace form [25]. The settings are  $\tau = 0.07$  and  $m = 0.2$  (Section 3.5).

$$\mathcal{L}_{\text{SOLAR}} = \frac{1}{V} \sum_{a=1}^V h_a, \quad h_a = \begin{cases} -\frac{1}{|\text{Pos}(a)|} \sum_{j \in \text{Pos}(a)} \log \frac{\exp(\ell_{aj})}{\sum_{b \in C(a)} \exp(\ell_{ab})}, & |\text{Pos}(a)| > 0, \\ 0, & |\text{Pos}(a)| = 0. \end{cases} \quad (5)$$

Equation (5) first averages the negative log probabilities over same-label positive indices for each focal view, then averages the focal terms over all  $V$  views. Each denominator contains every non-self positive and negative view. The objective therefore favours same-label relationships relative to the other views in the pool.

Repetition changes indexed multiplicities rather than adding label directions. If label  $y$  occurs  $r$  times, a focal cell  $z_a$  of that label receives the anchor contribution  $r \exp[(z_a^\top u_y - m)/\tau]$  in its denominator, and all  $r$  copies are positives. The copies also become focal views, changing the candidate and positive sets and the all-view average rather than multiplying the loss by a constant. Because  $r$  is the mini-batch label count, repetition is not class balancing or rare-label compensation.

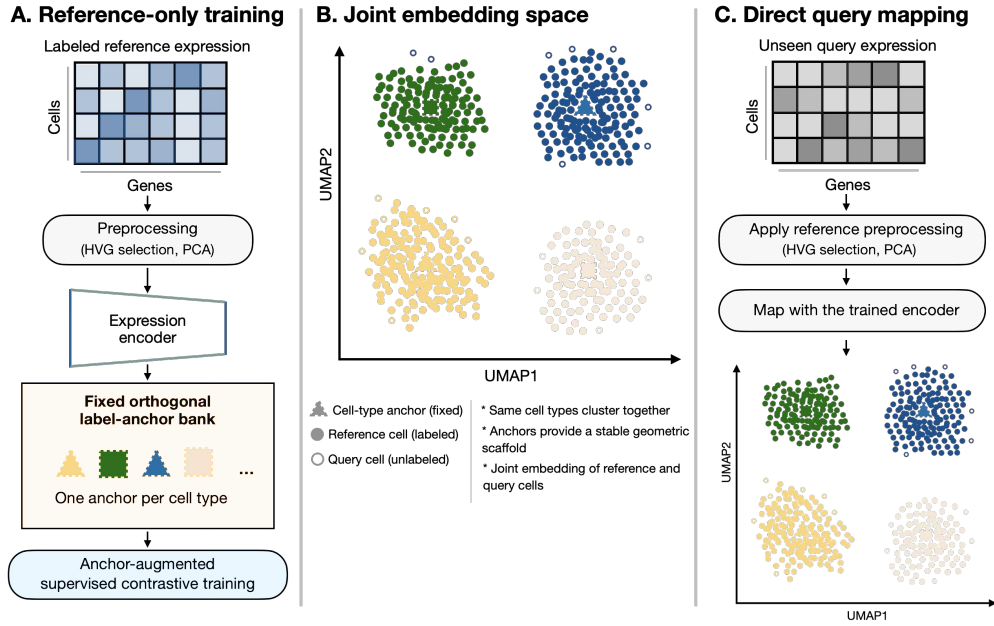


**Fig. 1 SOLAR reference training and direct query mapping.** (A) Reference-only gene selection and PCA supply a decoder-free expression encoder. The fixed anchor bank contains one orthogonal unit vector per reference label; the default training pool contains one indexed copy of the corresponding anchor for each cell, together with its expression embedding. (B) The diagram illustrates annotation-oriented organization in the shared embedding space; it is a schematic, not a quantitative result. (C) Query cells use the reference-selected genes, fitted PCA transform, and trained encoder without parameter updates. The anchor bank is used during training, not as a query-time classifier.

Repeated and unique anchors give every expression view, including a singleton label, a same-label positive. Without anchors, a singleton focal term is zero, but the cell can remain a negative in other terms and therefore need not have zero total gradient. Anchors have no trainable parameters; anchor-cell logits still depend on the encoder output, whereas anchor-anchor logits contain no trainable vector but enter the shared normalizations. Additional file 1 gives pseudocode and documents entirely empty-positive batches.

### 3.5 Training and direct query mapping

SOLAR divides the reference into 90% optimization and 10% internal validation subsets. AdamW [26] uses learning rate  $10^{-4}$ , batch size 512, temperature 0.07, margin 0.2, gradient-norm limit 1.0, and zero weight decay. Training stops after 10 epochs without a validation-loss improvement of at least  $10^{-4}$ , or after 50 epochs. Stochastic fits use model seeds 40–44.

The best validation checkpoint is restored and applied to the complete reference. Query cells are encoded in evaluation mode using the reference-selected genes, reference-fitted PCA transform, and trained encoder. No preprocessing or encoder parameter is updated with query data.

### 3.6 Comparative methods and evaluation

The primary comparison includes SOLAR, Matched PCA, frozen-reference scVI, frozen-reference scANVI, and SCLSC. Both scVI and scANVI [13, 14] are trained on reference cells and evaluated using their frozen encoders; we use the shorter names scVI and scANVI below. SCLSC [1] uses its official architecture, instance-type loss, reference-training class representatives, and reference-only stratified 90:10 internal split. Training stops according to validation loss and retains the final epoch, whereas SOLAR restores the best validation checkpoint. Because consistent raw counts are unavailable across these processed inputs, SCLSC receives the shared normalized expression matrix with reference-only selection of 2,000 genes, adapting the publication’s raw-count preprocessing to the available inputs. Additional file 1 details the training procedure, native readouts, Symphony [12], and scmap-cell [6].

The primary comparison uses 30-dimensional outputs for all five methods. SOLAR and SCLSC each contribute five fits per held-out batch, Matched PCA contributes one deterministic fit, and scVI and scANVI use their 30-dimensional frozen-reference

encoders. All five methods use a distance-weighted 15-nearest-neighbour classifier fitted to the complete reference embedding and labels. Using a common output width and readout aligns these two aspects of the evaluation; inputs, architectures, objectives, internal validation, and checkpoint selection remain method-specific. SOLAR receives 40 reference-fitted principal components as input. Additional file 1 reports the method-specific output configurations separately. Query macro-F1 and balanced accuracy evaluate label recovery, while joint reference–query representations are assessed separately for annotation structure and batch mixing [2–4].

The 128-dimensional anchor analysis compares one anchor copy per cell, one anchor per represented label, and no anchors. Additional file 1 separately reports the output-width sensitivity analysis. Uniform manifold approximation and projection (UMAP) is used only for descriptive visualization [27].

### 3.7 Evaluation metrics

#### *Query label recovery.*

Let  $\mathcal{C}_Q$  be the set of true labels observed in the query. The macro-F1 score is the unweighted mean over  $c \in \mathcal{C}_Q$  of  $2\text{TP}_c / (2\text{TP}_c + \text{FP}_c + \text{FN}_c)$ , with an undefined class F1 set to zero. Balanced accuracy is the mean class recall,  $|\mathcal{C}_Q|^{-1} \sum_{c \in \mathcal{C}_Q} \text{TP}_c / (\text{TP}_c + \text{FN}_c)$ . Both scores lie in  $[0, 1]$ , with higher values indicating better recovery of query annotations. Predictions of reference-only labels remain errors for the true query class; they are neither removed nor reassigned.

#### *Annotation separation and batch diversity.*

Representation metrics use the joint reference–query population unless otherwise stated. Label average silhouette width (label ASW) measures Euclidean separation of the supplied cell-type labels. For a cell, the raw silhouette is  $(b - a) / \max(a, b)$ , where  $a$  is its mean within-label distance and  $b$  is the smallest mean distance to another label. We report the scIB transformation  $(\bar{s} + 1) / 2 \in [0, 1]$ , averaging over cells; higher

values indicate stronger separation of the supplied annotations. The metric does not assess whether those annotations are biologically complete or correct. Batch ASW is a different score: it averages  $1 - |s|$  for batch-label silhouettes within each cell type and then averages the type means. It lies in  $[0, 1]$ , with higher values indicating less within-type batch segregation.

The integration local inverse Simpson’s index (iLISI) measures local batch diversity on the joint 15-neighbour graph. The scIB implementation uses shortest-path neighbourhoods of 90 cells, entropy-matched distance weights, and 50% subsampling. If  $p_{ib}$  is the resulting probability mass for batch  $b$  around cell  $i$ , the reported value is  $\text{median}_i(1/\sum_b p_{ib}^2) - 1$  over finite local values. It is not normalized to  $[0, 1]$ ; its nominal range for valid neighbourhoods with  $n_b$  batches is  $[0, n_b - 1]$ . Higher values indicate greater local batch diversity. Additional file 1 documents boundary cases and stochastic subsampling.

#### ***Local structure and pseudotime.***

Trustworthiness measures whether neighbours in a target representation are also neighbours in a specified source space [28]. We use a 50-dimensional PCA representation of the processed scIB expression data to define source-space neighbourhoods for this evaluation. This evaluation reference differs from the reference-fitted Matched PCA (40D) and Matched PCA (30D) targets. With Euclidean distance and  $k = 15$ , trustworthiness penalizes target neighbours absent from the source neighbourhood according to their source-distance ranks. It lies in  $[0, 1]$ ; higher values indicate stronger preservation of this measured neighbour relation. We score reference, query, and joint populations separately in immune and lung data, using the same sample of at most 6,000 cells per population for every method and dimension. Additional file 1 reports the sampling procedure and the separate within-batch analysis.

Diffusion pseudotime orders cells along diffusion-based progression from a root [29, 30]. Our immune-cell smoothness analysis uses the precomputed `dpt_pseudotime`

distributed with the processed scIB dataset; the available metadata do not record the original root used to construct this annotation. Within each reference, query, or joint population, the procedure retains cells with available pseudotime and correlates their values with the mean pseudotime of 15 neighbouring cells in the target representation using Spearman’s  $\rho$ . At least 16 annotated cells and nonconstant values are required. Smoothness lies in  $[-1, 1]$ ; higher values indicate stronger local agreement with the supplied pseudotime. Additional file 1 gives the exact neighbour handling.

Additional file 1 reports two related endpoints. Trajectory conservation compares source and target diffusion-pseudotime orderings and ranges from 0 to 1 when defined. Raw within-label silhouettes quantify association with study, sample, donor, or patient group on  $[-1, 1]$ ; their interpretation depends on the covariate and differs from batch ASW.

### ***Aggregation.***

Stochastic fits are first averaged within each fixed batch; batch means and then dataset means receive equal weight. With three evaluated batches per dataset, complete coverage gives the mean of nine batch means. The main tables report descriptive sample standard deviations (SDs) across those nine means. Per-batch supplementary tables instead report fit-to-fit SDs from five seeds. Deterministic PCA has no fit-to-fit SD, although its performance can vary across batches. Pairwise contrasts include batches with complete results for both methods; missing values are not replaced with zero. No significance or equivalence tests were performed.

### ***Use of artificial intelligence-assisted tools.***

Artificial intelligence (AI)-assisted tools supported English drafting and editing, public-source retrieval, source-code inspection, reconstruction checks for existing tables and figures, and consistency and document verification. They were not used to generate new experimental data or retrain models.

## 4 Results

### 4.1 SOLAR improved query label recovery over Matched PCA

The cross-method comparison used 30-dimensional outputs and the same reference-trained nearest-neighbour readout across nine held-out batches. SOLAR achieved mean query macro-F1 of 0.811 and balanced accuracy of 0.837, compared with 0.774 and 0.803 for Matched PCA (Table 2). Its mean macro-F1 advantage over Matched PCA was positive within each of the three datasets (Figure 2A). SCLSC produced a similar mean macro-F1 of 0.815, whereas scANVI produced higher macro-F1 (0.838) and balanced accuracy (0.882) than SOLAR (Figure 2B). The full batch-level distributions and fitting-seed variability are reported in Figure S4 and Table S3 in Additional file 1.

The immune-cell Villani batch remained difficult despite complete reference-label coverage, with errors concentrated in CD16+ monocytes (Figure S2 in Additional file 1).

**Table 2** Held-out query annotation with 30-dimensional outputs and the shared weighted 15-nearest-neighbour readout. Values are means  $\pm$  descriptive SD across nine batch means; stochastic batch means average five fits and Matched PCA is deterministic. Table S3 in Additional file 1 gives batch-level values.

| Method      | Macro-F1          | Balanced accuracy |
|-------------|-------------------|-------------------|
| SOLAR       | 0.811 $\pm$ 0.114 | 0.837 $\pm$ 0.082 |
| Matched PCA | 0.774 $\pm$ 0.127 | 0.803 $\pm$ 0.096 |
| scVI        | 0.807 $\pm$ 0.091 | 0.855 $\pm$ 0.053 |
| scANVI      | 0.838 $\pm$ 0.087 | 0.882 $\pm$ 0.061 |
| SCLSC       | 0.815 $\pm$ 0.078 | 0.845 $\pm$ 0.054 |

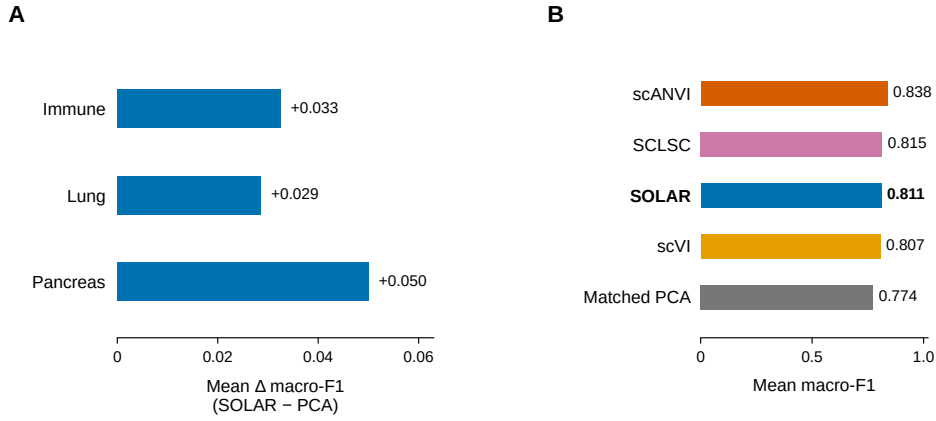


**Fig. 2 Query annotation recovery at a common output dimension.** (A) Mean query macro-F1 differences between SOLAR and Matched PCA within the immune, lung, and pancreas datasets. Each bar averages the three paired held-out-batch differences for one dataset. (B) Mean query macro-F1 for all five methods, averaging batches within datasets and datasets equally. All methods have 30-dimensional outputs and use the same reference-fitted distance-weighted 15-nearest-neighbour readout. Stochastic results first average five fits per batch; Matched PCA is deterministic. Figure S4 and Table S3 in Additional file 1 retain batch-level results and fitting variability.

## 4.2 SOLAR organized joint representations around the supplied labels

In the joint reference–query populations, SOLAR reached a label ASW of 0.873, above Matched PCA (0.574), scVI (0.578), and scANVI (0.657), but below SCLSC (0.948; Table 3).

The two batch-diversity metrics ranked the methods differently because batch ASW measures within-label batch segregation, whereas iLISI measures local batch diversity. SOLAR had a higher batch ASW than SCLSC (0.836 versus 0.743), but a lower iLISI (1.740 versus 2.241); scVI and scANVI also had higher iLISI than SOLAR.

**Table 3** Joint-population annotation separation and batch diversity with 30-dimensional outputs. Values are means across nine equally weighted batch means; stochastic batch means average five fits. Higher values indicate more of the named property, and the endpoints are reported separately.

| Method      | Label ASW | Batch ASW | iLISI |
|-------------|-----------|-----------|-------|
| SOLAR       | 0.873     | 0.836     | 1.740 |
| Matched PCA | 0.574     | 0.831     | 0.304 |
| scVI        | 0.578     | 0.878     | 2.125 |
| scANVI      | 0.657     | 0.867     | 1.902 |
| SCLSC       | 0.948     | 0.743     | 2.241 |

### 4.3 Repeated anchors increased label separation and local batch diversity

The 128-dimensional controls examined how anchor presence and multiplicity in the contrastive pool affected the learned representation. Repeated anchors increased held-out label ASW and iLISI relative to both controls (Figure 3), whereas query recovery changed little: the macro-F1 difference was about 0.007 relative to no anchors and below 0.001 relative to unique anchors, and balanced accuracy was slightly lower than with unique anchors. Trustworthiness and immune-cell pseudotime smoothness were also lower than with either control. Table S9 and Figure S3 in Additional file 1 report the complete signed contrasts, including the local-structure endpoints.

### 4.4 SOLAR retained more local structure than SCLSC on two evaluated endpoints

Across six immune and lung held-out batches, mean query trustworthiness was 0.936 for SOLAR and 0.903 for SCLSC.

Every batch mean favoured SOLAR (Tables S5 and S7 in Additional file 1). Across the three immune batches, joint reference–query pseudotime smoothness was 0.861 for SOLAR and 0.818 for SCLSC, again with the same direction in all three batch

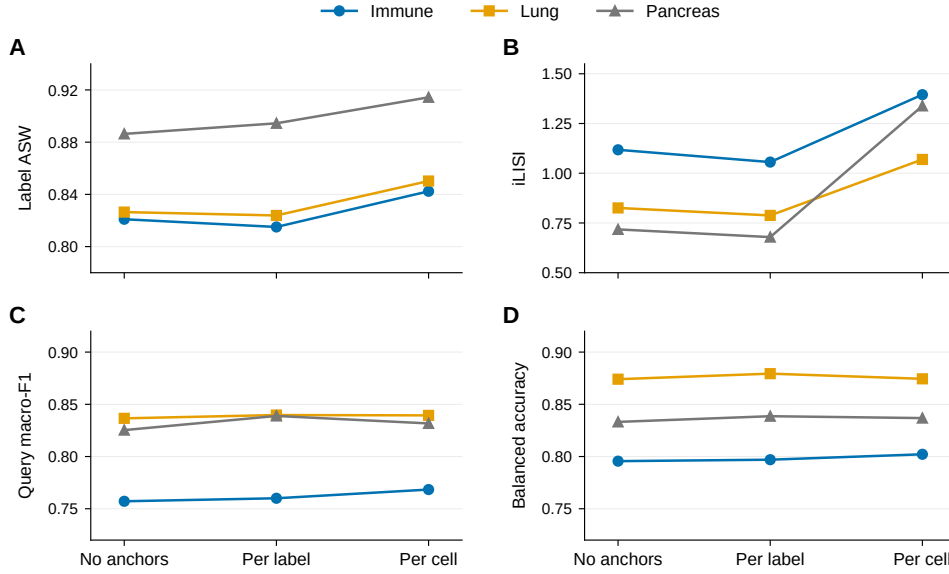


**Fig. 3 Anchor-dependent annotation separation, batch diversity, and query recovery.** (A) Joint label ASW, (B) joint iLISI, (C) query macro-F1, and (D) query balanced accuracy compare three SOLAR training conditions at 128 dimensions. “No anchors”, “Per label”, and “Per cell” denote no-anchor, unique-anchor (one anchor per label), and repeated-anchor (one copy of the corresponding label anchor per cell) training, respectively. Each point averages three held-out batch means for one dataset, after averaging five fits within each batch and condition. Lines connect the three condition means for the same dataset. Table S9 in Additional file 1 reports the full signed contrasts, including local-structure endpoints not shown here; Figure S3 shows signed contrasts for 10 metrics.

means. SCLSC had higher label separation, whereas SOLAR retained more query-neighbourhood structure and stronger local agreement with the supplied pseudotime on these evaluated populations.

Matched PCA, scVI, and scANVI retained higher query trustworthiness (0.969, 0.946, and 0.949, respectively) and joint immune pseudotime smoothness (0.894, 0.880, and 0.886) than SOLAR. Query-only pseudotime smoothness was evaluable in just one immune batch, *Oetjen\_A*; the three-batch comparison above uses the joint population.

Figure 4 places these observations beside label separation. Panel A uses the same six immune/lung batches for label ASW and query trustworthiness; panel B uses the same three immune batches for label ASW and joint pseudotime smoothness. Results

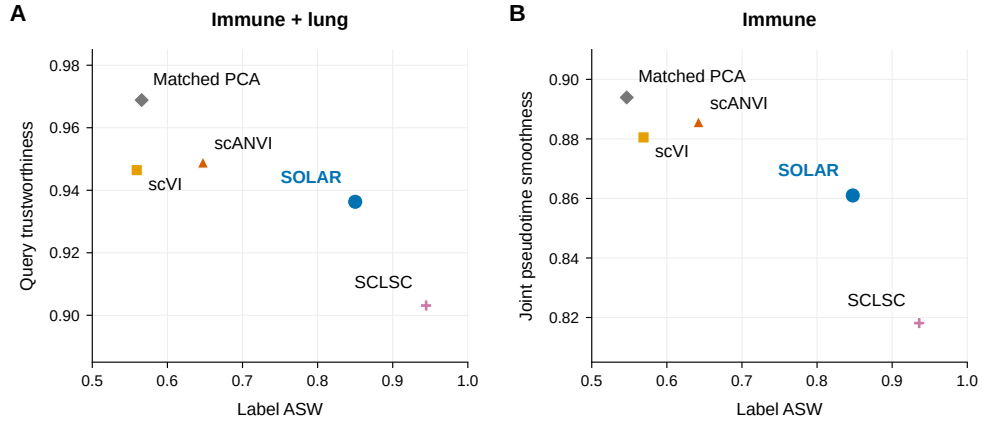


**Fig. 4 Annotation separation and local-structure retention at a common output dimension.** Moving right indicates stronger label separation; moving up indicates greater measured local-structure preservation. Each point represents one method with 30-dimensional outputs. Joint reference–query label ASW is plotted against (A) query trustworthiness across six immune and lung batches and (B) joint pseudotime neighbourhood smoothness across three immune batches. Within each panel, both axes use the same batches, averaging fits within batches, batches within datasets, and datasets equally; Matched PCA is deterministic. Trustworthiness is evaluated against the 50-dimensional PCA representation of the processed expression data, which is distinct from the reference-fitted Matched PCA targets; smoothness measures local agreement with supplied pseudotime. Tables S5 and S7 in Additional file 1 report the detailed scores and population coverage.

for the method-specific output configurations are provided in Figure S5 and Table S5 in Additional file 1.

## 5 Discussion

By fixing label-anchor relationships before training, SOLAR gives the reference-mapping encoder explicit label directions while learning data-driven cell positions through a shared contrastive objective. In the 30-dimensional comparison, the complete method improved query macro-F1 and label ASW over Matched PCA; scANVI had higher macro-F1, and SCLSC had higher label ASW.

The cross-method comparison evaluates the complete mapping method, whereas the 128-dimensional controls examine how anchor presence and repetition affect the learned representation. Repeated and unique conditions used the same directions, but

repetition added indexed copies and the third condition omitted anchors. Repetition increased label separation and local batch diversity, while query-label recovery changed little. It also reduced the evaluated local-structure scores. Query-label prediction alone therefore did not capture these changes in the learned representation.

The experiments cover closed-label mapping across nine held-out batches from three datasets; unknown cell types and incorrect reference annotations were not evaluated. The local measures quantify agreement with a 50-dimensional PCA representation and with supplied pseudotime. The anchor controls vary the presence and multiplicity of one fixed bank, not fixed versus learned anchors or orthogonal versus alternative fixed geometries. SOLAR and SCLSC also differ in inputs, objectives, and checkpoint selection, so their differences do not isolate anchor geometry.

## 6 Conclusions

We introduced SOLAR, a supervised contrastive reference-mapping method that assigns labels fixed orthogonal directions and trains an expression encoder against those anchors. Across the evaluated held-out batches, the complete method improved query-label recovery and label separation over Matched PCA. In the 128-dimensional anchor controls, anchor repetition increased label separation and local batch diversity, while query-label recovery changed little. SOLAR therefore provides a concrete way to prescribe label relationships before training while learning where cells map in the resulting space.

## Additional file 1

*Supplementary methods, figures and tables.* Additional file 1 provides reference–query protocols, comparator details, metric definitions, per-batch results, complete cross-method and anchor/dimension summaries, coverage records, and supplementary figures. Tables and figures prefixed “S” in the main text refer to this file.

875 **Abbreviations**

876  
877 *ASW* Average silhouette width  
878  
879 *iLISI* Integration local inverse Simpson’s index  
880  
881 *PCA* Principal component analysis  
882  
883 *scANVI* Single-cell annotation using variational inference  
884  
885 *scIB* Single-cell integration benchmarking  
886  
887 *SCLSC* Supervised contrastive learning on cells and their types  
888  
889 *SOLAR* Supervised Orthogonal Label-Anchor Representation  
890  
891 *UMAP* Uniform manifold approximation and projection  
892

893 **Declarations**

894 **Use of artificial intelligence-assisted tools**

895  
896  
897  
898  
899 Use of artificial intelligence (AI)-assisted tools is described in Methods. No AI sys-  
900 tem is listed as an author. The authors remain responsible for scientific accuracy,  
901 interpretation, and final approval before submission.  
902  
903  
904

905 **Ethics approval and consent to participate**

906  
907 This study used only publicly available, de-identified processed benchmark datasets.  
908  
909 The original studies from which these data were derived were responsible for par-  
910 ticipant consent and ethics approvals, as reported in their source publications. No  
911 participants were recruited, no new human tissue or identifiable information was  
912 accessed, and no animal experiments were performed for this secondary computational  
913 analysis; additional institutional approval and consent to participate were therefore  
914 not required.  
915  
916  
917  
918  
919  
920

|                                                                                                                                                   |     |
|---------------------------------------------------------------------------------------------------------------------------------------------------|-----|
| <b>Consent for publication</b>                                                                                                                    | 921 |
|                                                                                                                                                   | 922 |
| Not applicable. This manuscript contains no identifiable individual person’s data.                                                                | 923 |
|                                                                                                                                                   | 924 |
|                                                                                                                                                   | 925 |
| <b>Availability of data and materials</b>                                                                                                         | 926 |
|                                                                                                                                                   | 927 |
| The processed benchmark datasets are distributed by the scIB benchmark project                                                                    | 928 |
| through Figshare, version 8 [21]: <a href="https://doi.org/10.6084/m9.figshare.12420968.v8">https://doi.org/10.6084/m9.figshare.12420968.v8</a> . | 929 |
|                                                                                                                                                   | 930 |
| The SOLAR source, configurations, machine-readable results, and provenance records                                                                | 931 |
| are available from the <a href="#">SOLAR repository</a> . The software is released under the MIT                                                  | 932 |
| License.                                                                                                                                          | 933 |
|                                                                                                                                                   | 934 |
|                                                                                                                                                   | 935 |
|                                                                                                                                                   | 936 |
|                                                                                                                                                   | 937 |
| <b>Competing interests</b>                                                                                                                        | 938 |
|                                                                                                                                                   | 939 |
| The authors declare that they have no competing interests.                                                                                        | 940 |
|                                                                                                                                                   | 941 |
|                                                                                                                                                   | 942 |
| <b>Funding</b>                                                                                                                                    | 943 |
|                                                                                                                                                   | 944 |
| This work was supported by the Japan Society for the Promotion of Science (JSPS)                                                                  | 945 |
| through JSPS KAKENHI Grant Number JP23H03356 and by the National Key R&D                                                                          | 946 |
| Program of China (Grant No. 2022YFC3400400).                                                                                                      | 947 |
|                                                                                                                                                   | 948 |
|                                                                                                                                                   | 949 |
|                                                                                                                                                   | 950 |
| <b>Authors’ contributions</b>                                                                                                                     | 951 |
|                                                                                                                                                   | 952 |
| GZ and MQC implemented the software and conducted the computational experi-                                                                       | 953 |
| ments. MRC conceived the label-anchor strategy and contributed to the algorithmic                                                                 | 954 |
| design and mathematical formulation. GZ and MRC led manuscript preparation and                                                                    | 955 |
| revision. KT, HS and LL supervised the study and provided scientific guidance.                                                                    | 956 |
|                                                                                                                                                   | 957 |
|                                                                                                                                                   | 958 |
|                                                                                                                                                   | 959 |
|                                                                                                                                                   | 960 |
| <b>Acknowledgements</b>                                                                                                                           | 961 |
|                                                                                                                                                   | 962 |
| Minrui Chen gratefully acknowledges scholarship support from the Otsuka Toshimi                                                                   | 963 |
| Scholarship Foundation. The authors thank the developers and contributors of the scIB                                                             | 964 |
|                                                                                                                                                   | 965 |
|                                                                                                                                                   | 966 |

967 project for making the benchmark datasets, evaluation metrics and reproducibility  
968 resources publicly available.  
969

970

971

## 972 References

973

974 [1] Heryanto YD, Zhang Yz, Imoto S. Predicting cell types with supervised con-  
975

976 trastive learning on cells and their types. Scientific Reports. 2024;14:430. <https://doi.org/10.1038/s41598-023-50185-2>.  
977

978  
979 [2] Büttner M, Miao Z, Wolf FA, Teichmann SA, Theis FJ. A test metric for assessing  
980 single-cell RNA-seq batch correction. Nature methods. 2019;16(1):43–49.

981  
982 [3] Tran HTN, Ang KS, Chevrier M, Zhang X, Lee NYS, Goh M, et al. A benchmark  
983 of batch-effect correction methods for single-cell RNA sequencing data. Genome  
984 biology. 2020;21(1):12.  
985

986  
987 [4] Luecken MD, Büttner M, Chaichoompu K, Danese A, Interlandi M, Müller MF,  
988 et al. Benchmarking atlas-level data integration in single-cell genomics. Nature  
989 methods. 2022;19(1):41–50.  
990

991  
992 [5] Heumos L, Schaar AC, Lance C, Litinetskaya A, Drost F, Zappia L, et al. Best  
993 practices for single-cell analysis across modalities. Nature Reviews Genetics.  
994 2023;24(8):550–572.  
995

996  
997 [6] Kiselev VY, Yiu A, Hemberg M. scmap: projection of single-cell RNA-seq data  
998 across data sets. Nature methods. 2018;15(5):359–362. [https://doi.org/10.1038/](https://doi.org/10.1038/nmeth.4644)  
1000 [nmeth.4644](https://doi.org/10.1038/nmeth.4644).  
1001

1002 [7] Pliner HA, Shendure J, Trapnell C. Supervised classification enables rapid  
1003 annotation of cell atlases. Nature methods. 2019;16(10):983–986.  
1004

1005  
1006 [8] Alquicira-Hernandez J, Sathe A, Ji HP, Nguyen Q, Powell JE. scPred: accu-  
1007 rate supervised method for cell-type classification from single-cell RNA-seq data.  
1008 Genome biology. 2019;20(1):264.  
1009

1010

1011

1012

- [9] Haghverdi L, Lun AT, Morgan MD, Marioni JC. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nature biotechnology*. 2018;36(5):421–427. 1013–1017
- [10] Hie B, Bryson B, Berger B. Efficient integration of heterogeneous single-cell transcriptomes using Scanorama. *Nature biotechnology*. 2019;37(6):685–691. 1018–1020
- [11] Korsunsky I, Millard N, Fan J, Slowikowski K, Zhang F, Wei K, et al. Fast, sensitive and accurate integration of single-cell data with Harmony. *Nature methods*. 2019;16(12):1289–1296. 1021–1025
- [12] Kang JB, Nathan A, Weinand K, Zhang F, Millard N, Rumker L, et al. Efficient and precise single-cell reference atlas mapping with Symphony. *Nature communications*. 2021;12(1):5890. <https://doi.org/10.1038/s41467-021-25957-x>. 1026–1030
- [13] Lopez R, Regier J, Cole MB, Jordan MI, Yosef N. Deep generative modeling for single-cell transcriptomics. *Nature methods*. 2018;15(12):1053–1058. 1031–1033
- [14] Xu C, Lopez R, Mehlman E, Regier J, Jordan MI, Yosef N. Probabilistic harmonization and annotation of single-cell transcriptomics data with deep generative models. *Molecular systems biology*. 2021;17(1):MSB20209620. 1034–1038
- [15] Lotfollahi M, Naghipourfar M, Luecken MD, Khajavi M, Büttner M, Wagenstetter M, et al. Mapping single-cell data to reference atlases by transfer learning. *Nature biotechnology*. 2022;40(1):121–130. 1039–1043
- [16] Khosla P, Teterwak P, Wang C, Sarna A, Tian Y, Isola P, et al. Supervised Contrastive Learning. In: *Advances in Neural Information Processing Systems*. vol. 33. Curran Associates, Inc.; 2020. p. 18661–18673. Available from: <https://proceedings.neurips.cc/paper/2020/hash/d89a66c7c80a29b1bdbab0f2a1a94af8-Abstract.html>. 1044–1051
- [17] Yang M, Yang Y, Xie C, Ni M, Liu J, Yang H, et al. Contrastive learning enables rapid mapping to multimodal single-cell atlas of multimillion scale. *Nature Machine Intelligence*. 2022;4(8):696–709. <https://doi.org/10.1038/s42256-022-0> 1052–1058

1059 0518-z.  
1060  
1061 [18] De Donno C, Hedyeh-Zadeh S, Moinfar AA, Wagenstetter M, Zappia L, Lot-  
1062 follahi M, et al. Population-level integration of single-cell datasets enables  
1063 multi-scale analysis across samples. *Nature methods*. 2023;20(11):1683–1692.  
1064  
1065 [19] Kim S, Kim D, Cho M, Kwak S. Proxy Anchor Loss for Deep Metric Learning.  
1066 In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern*  
1067 *Recognition*; 2020. p. 3238–3247.  
1068  
1069 [20] Ahmed N, Kukleva A, Schiele B. OrCo: Towards Better Generalization via  
1070 Orthogonality and Contrast for Few-Shot Class-Incremental Learning. In:  
1071 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern*  
1072 *Recognition*; 2024. p. 28762–28771. Available from: [https://openaccess.thecvf.co](https://openaccess.thecvf.com/content/CVPR2024/html/Ahmed.OrCo_Towards_Better_Generalization_via_Orthogonality_and_Contrast_for_Few-Shot_CVPR_2024_paper.html)  
1073 [m/content/CVPR2024/html/Ahmed.OrCo\\_Towards\\_Better\\_Generalization\\_via](https://openaccess.thecvf.com/content/CVPR2024/html/Ahmed.OrCo_Towards_Better_Generalization_via_Orthogonality_and_Contrast_for_Few-Shot_CVPR_2024_paper.html)  
1074 [\\_Orthogonality\\_and\\_Contrast\\_for\\_Few-Shot\\_CVPR\\_2024\\_paper.html](https://openaccess.thecvf.com/content/CVPR2024/html/Ahmed.OrCo_Towards_Better_Generalization_via_Orthogonality_and_Contrast_for_Few-Shot_CVPR_2024_paper.html).  
1075  
1076 [21] Luecken MD, Büttner M, Danese A, Interlandi M, Müller M, Strobl D, et al.  
1077 Benchmarking atlas-level data integration in single-cell genomics: integration task  
1078 datasets. *figshare*; 2020. Version 8. Dataset. Available from: [https://doi.org/10](https://doi.org/10.6084/m9.figshare.12420968.v8)  
1079 [.6084/m9.figshare.12420968.v8](https://doi.org/10.6084/m9.figshare.12420968.v8).  
1080  
1081 [22] Wolf FA, Angerer P, Theis FJ. SCANPY: large-scale single-cell gene expression  
1082 data analysis. *Genome Biology*. 2018;19:15. [https://doi.org/10.1186/s13059-017](https://doi.org/10.1186/s13059-017-1382-0)  
1083 [-1382-0](https://doi.org/10.1186/s13059-017-1382-0).  
1084  
1085 [23] Ioffe S, Szegedy C. Batch normalization: accelerating deep network training by  
1086 reducing internal covariate shift. In: *Proceedings of the 32nd International Confer-*  
1087 *ence on Machine Learning*. vol. 37 of *Proceedings of Machine Learning Research*;  
1088 2015. p. 448–456. Available from: [https://proceedings.mlr.press/v37/ioffe15.htm](https://proceedings.mlr.press/v37/ioffe15.html)  
1089 [l](https://proceedings.mlr.press/v37/ioffe15.html).  
1090  
1091 [24] Ba JL, Kiros JR, Hinton GE. Layer normalization. *arXiv*. 2016;[arXiv:1607.06450](https://arxiv.org/abs/1607.06450).  
1092  
1093  
1094  
1095  
1096  
1097  
1098  
1099  
1100  
1101  
1102  
1103  
1104

- [25] Deng J, Guo J, Xue N, Zafeiriou S. ArcFace: Additive Angular Margin Loss for Deep Face Recognition. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2019. p. 4685–4694.
- [26] Loshchilov I, Hutter F. Decoupled weight decay regularization. International Conference on Learning Representations. 2019;2019:1–19.
- [27] McInnes L, Healy J, Saul N, Grossberger L. UMAP: Uniform Manifold Approximation and Projection. Journal of Open Source Software. 2018;3(29):861. <https://doi.org/10.21105/joss.00861>.
- [28] Venna J, Kaski S. Neighborhood preservation in nonlinear projection methods: an experimental study. In: Artificial Neural Networks—ICANN 2001. vol. 2130 of Lecture Notes in Computer Science; 2001. p. 485–491.
- [29] Coifman RR, Lafon S. Diffusion maps. Applied and Computational Harmonic Analysis. 2006;21(1):5–30. <https://doi.org/10.1016/j.acha.2006.04.006>.
- [30] Haghverdi L, Büttner M, Wolf FA, Büttner F, Theis FJ. Diffusion pseudotime robustly reconstructs lineage branching. Nature Methods. 2016;13:845–848. <https://doi.org/10.1038/nmeth.3971>.

## Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [SISOLARFixedorthogonallabelanchorsforsinglecellreferencemapping.pdf](#)