

# A cell atlas foundation model for scalable search of similar human cells

<https://doi.org/10.1038/s41586-024-08411-y>

Received: 13 August 2023

Accepted: 14 November 2024

Published online: 20 November 2024

Open access

 Check for updates

Graham Heimberg<sup>1,2,7</sup>, Tony Kuo<sup>3,7</sup>, Daryle J. DePianto<sup>2</sup>, Omar Salem<sup>1</sup>, Tobias Heigl<sup>2</sup>, Nathaniel Diamant<sup>1</sup>, Gabriele Scalia<sup>1</sup>, Tommaso Biancalani<sup>1</sup>, Shannon J. Turley<sup>2,4</sup>, Jason R. Rock<sup>2,4</sup>, Héctor Corrada Bravo<sup>1</sup>, Josh Kaminker<sup>5,8</sup>, Jason A. Vander Heiden<sup>1,2,8</sup> & Aviv Regev<sup>6,8</sup>

Single-cell RNA sequencing has profiled hundreds of millions of human cells across organs, diseases, development and perturbations to date. Mining these growing atlases could reveal cell–disease associations, identify cell states in unexpected tissue contexts and relate in vivo biology to in vitro models. These require a common measure of cell similarity across the body and an efficient way to search. Here we develop SCimilarity, a metric-learning framework to learn a unified and interpretable representation that enables rapid queries of tens of millions of cell profiles from diverse studies for cells that are transcriptionally similar to an input cell profile or state. We use SCimilarity to query a 23.4-million-cell atlas of 412 single-cell RNA-sequencing studies for macrophage and fibroblast profiles from interstitial lung disease<sup>1</sup> and reveal similar cell profiles across other fibrotic diseases and tissues. The top scoring in vitro hit for the macrophage query was a 3D hydrogel system<sup>2</sup>, which we experimentally demonstrated reproduces this cell state. SCimilarity serves as a foundation model for single-cell profiles that enables researchers to query for similar cellular states across the human body, providing a powerful tool for generating biological insights from the Human Cell Atlas.

Over 100 million individual cells have been profiled using single-cell (scRNA-seq) or single-nucleus (snRNA-seq) RNA-sequencing analysis across homeostatic, disease and experimentally perturbed conditions<sup>3</sup>. By comparing cell profiles from hundreds of studies, researchers can connect cell states across different developmental stages, tissues or diseases, or between the human body and in vitro laboratory models. Despite this promise and rapid data growth, current models were not designed to search for similar cell profiles in massive corpora, and cross-dataset, pan-body, analyses are hampered by challenges in dataset curation and harmonization, difficulty in defining a common low-dimensional representation between datasets, lack of principled metrics to compare between cell profiles and no methods to search for complete cell profiles. As a result, most aggregation efforts have been limited in scope, with a few recent exceptions<sup>4–7</sup>.

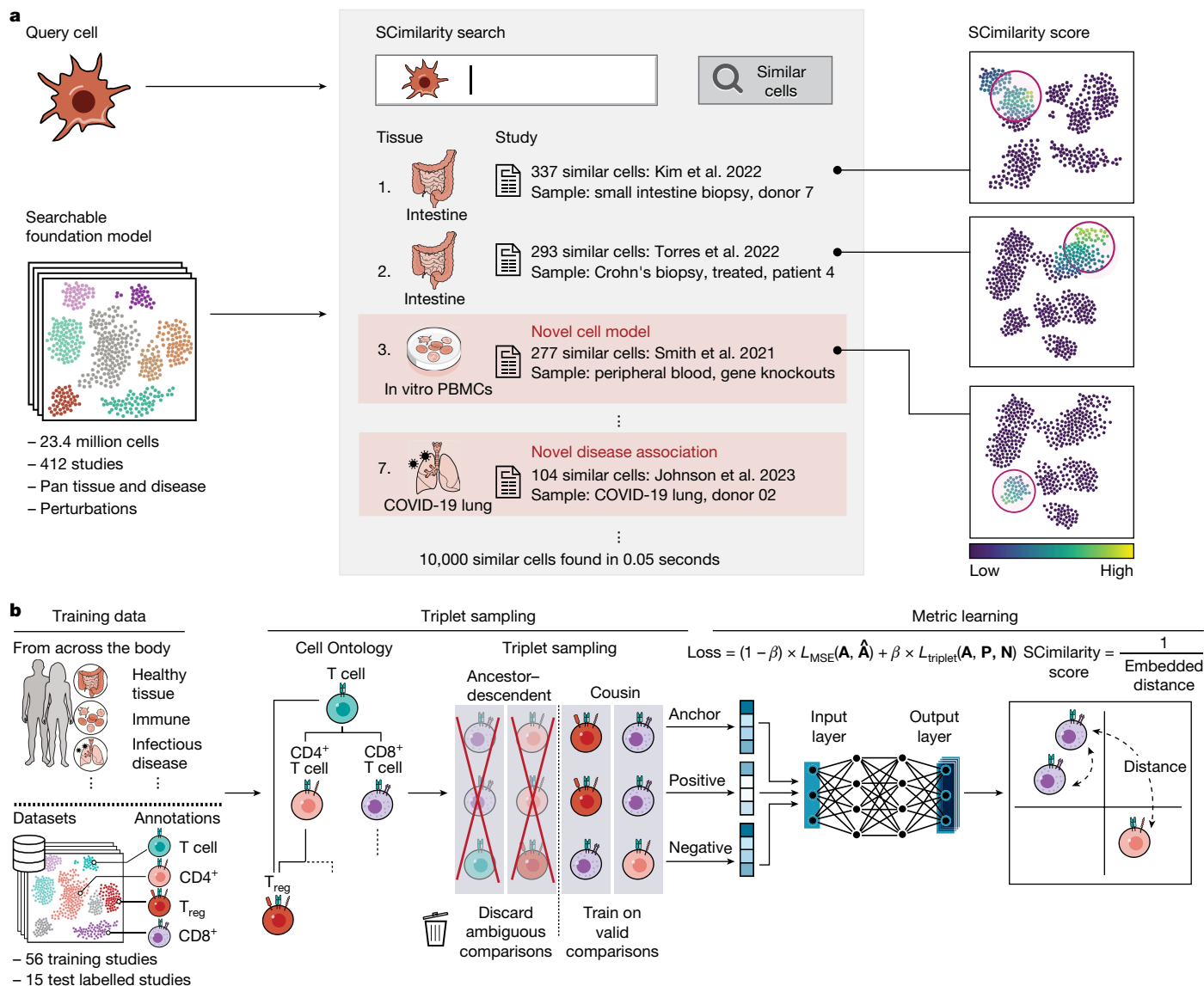
To leverage and query the massive scale and richness of single-cell atlases, we need (1) a foundation model of cell states with an effective representation for single-cell profiles usable across applications without retraining; and (2) a measure of cell similarity that is robust to technical noise, scales to hundreds of millions of cells, and generalizes to datasets and cell states not observed during training. Unsupervised methods, such as principal component analysis or autoencoders<sup>8–11</sup>, faithfully preserve information from the input<sup>8–11</sup>, but do not learn universal features that encode cells and the similarity between cells

needed to query new datasets. Conversely, other machine-learning methods, especially in image processing, have successfully learned representations of diverse entities and their similarity. In particular, metric-learning models for facial recognition are trained to embed images into a low-dimensional space where images of the same person are closer than images of different people<sup>12</sup>. Users query-trained models with an image not in the training set to find additional images that are nearby in the embedding and depict the same person. Analogously, metric learning could provide a meaningful metric for the similarity between cells, by training a model using annotated sc/snRNA-seq data to learn a low-dimensional representation that places similar cell profiles near each other and dissimilar ones farther apart. If learned from a sufficient diversity of cell profiles, such a representation should provide a foundation model of cells that allows efficient searches for cells with similar expression states (Fig. 1a).

Here we introduce SCimilarity—a deep-metric-learning foundation model that quantifies similarity between single-cell profiles and provides a single-cell reference to query for comparable cell states across tissues and diseases. We illustrate the power of SCimilarity by searching a learned reference of 23.4 million cells with query profiles of macrophage and fibroblast subsets from interstitial lung disease (ILD)<sup>1</sup>, showing how SCimilarity provides a powerful framework for scalable cell search across organs, systems and conditions to generate

<sup>1</sup>Biology Research, AI Development, gRED Computational Sciences, Genentech, San Francisco, CA, USA. <sup>2</sup>Department of Immunology Discovery, Genentech, San Francisco, CA, USA. <sup>3</sup>Roche Informatics, F. Hoffmann–La Roche, Mississauga, Ontario, Canada. <sup>4</sup>Department of Regenerative Medicine, Genentech, San Francisco, CA, USA. <sup>5</sup>OMNI Bioinformatics, gRED Computational Sciences, Genentech, San Francisco, CA, USA. <sup>6</sup>Research and Early Development, Genentech, San Francisco, CA, USA. <sup>7</sup>These authors contributed equally: Graham Heimberg, Tony Kuo.

<sup>8</sup>These authors jointly supervised this work: Josh Kaminker, Jason A. Vander Heiden, Aviv Regev. <sup>✉</sup>e-mail: heimberg@gene.com; kaminker@gene.com; vanderheiden.jason@gene.com; regev.aviv@gene.com



**Fig. 1 | SCSimilarity metric learning enables cell search in large human scale atlases. a**, Cell querying with SCSimilarity. Left, a query cell profile is compared to a searchable reference foundation model of 23.4 million profiles from 412 studies. Middle, samples with similar cells are identified and returned with information about the original sample conditions, including tissue, in vitro or diseases contexts. Right, a SCSimilarity score is computed between the query cell and each cell within a tissue sample. **b**, Triplet loss training. Left, 56 training and 15 test datasets with Cell Ontology annotations from across the body are

used as input. Middle, cell triplets are sampled, each consisting of an anchor cell (**A**), a positive cell (**P**, anchor-similar) and a negative cell (**N**, anchor-dissimilar), based on Cell Ontology annotations. Only non-ambiguous relationships are allowed. Right, triplets are used to train a neural network that embeds similar cells closer than dissimilar ones, forming a foundation model. T<sub>reg</sub>, regulatory T cells. The loss function is computed using a cell triplet, a reconstructed anchor cell profile ( $\hat{\mathbf{A}}$ ), and a weighting parameter ( $\beta$ ) to balance the triplet loss ( $L_{\text{triplet}}$ ) and the mean squared error loss ( $L_{\text{MSE}}$ ).

biological insights and experimentally testable hypotheses from the Human Cell Atlas.

### A similarity metric for scRNA-seq

SCSimilarity blends unsupervised representation learning and supervised metric learning through simultaneously optimizing two objectives: (1) a supervised triplet loss function, which is used to embed expression profiles from matching cell types close together, integrating cells of the same type across studies<sup>13–15</sup>; and (2) an unsupervised mean squared error (MSE) reconstruction loss function, which encourages the model to preserve variation from the input expression profiles, capturing subtler differences in expression patterns within cells of the same type (Fig. 1b and Methods). Increasing the relative weight of the reconstruction loss improves querying performance, while increasing

the relative weight of the triplet loss improves performance on dataset integration metrics<sup>16</sup>. We focused on a single ( $\beta = 0.001$ ) model that best combined query sensitivity and integration performance (below).

We trained SCSimilarity with tens of millions of cell triplets sampled from data with author-provided standardized cell type annotations from the Cell Ontology<sup>17</sup> (Fig. 1b and Methods). Each triplet consists of an anchor, a positive and a negative cell: the anchor and positive cells are similar cells (that is, the same cell type) from different studies, while the anchor and negative cells are dissimilar (that is, different cell types; from the same or a different study). Even with standardized Cell Ontology terms, some cell type comparisons are ambiguous due to differences in annotation granularity (for example, it is ambiguous whether cells annotated as ‘T cell’ in one study and ‘CD4<sup>+</sup> T cell’ in another are similar or dissimilar). SCSimilarity therefore excludes triplets with positive and negative labels that have a vertical,

ancestor-descendant relationship in the Cell Ontology, and learns only from cells that are either explicitly similar or unambiguously dissimilar (Fig. 1b and Methods). This eliminates the need to manually flatten or harmonize every cell type annotation and seamlessly scales the training set across studies.

### Training on a large, diverse atlas

To test the SCimilarity framework, we aggregated sc/snRNA-seq datasets across human biology. We focused on studies generated using one experimental platform (10x Genomics Chromium droplet-based sc/snRNA-seq), mostly sourced from the Gene Expression Omnibus (GEO)<sup>18</sup> or CELLxGENE<sup>19</sup>. These data were generated with similar library preparation protocols and computational pre-processing pipelines<sup>20</sup>. There were 753 datasets matching our criteria as of 23 March 2021. The number of samples and cells matching our criteria has at least doubled every 6 months between December 2018 and March 2021 (Extended Data Fig. 1a,b). We programmatically downloaded 13,401,599 cell profiles from 333 of the studies with their respective GEO metadata and unnormalized gene count matrices (Methods and Supplementary Table 1), and manually ingested another 66 studies from either CELLxGENE<sup>19</sup> or other large studies and consortia (Methods) to a corpus of 412 studies comprising 23,381,150 cells from 5,142 tissue samples with 184 unique Tissue Ontology terms<sup>21</sup> and 132 Disease Ontology terms<sup>22</sup> (Fig. 2, Extended Data Fig. 1c and Supplementary Table 1).

We trained SCimilarity models with a training set of 7,886,247 single-cell profiles from 56 studies (46 scRNA-seq and 10 snRNA-seq) with 203 Cell Ontology author-annotated terms<sup>17</sup> (each appearing in at least two datasets) (Extended Data Fig. 1d and Supplementary Table 1). We sampled 50,000,000 of the most informative cell triplets (Methods) weighted by study and cell type (to mitigate dataset size imbalances), requiring that the anchor and positive cells in each triplet are from two different studies, and using hard triplet mining<sup>12</sup>, so that only the most informative triplets are used when updating model gradients (Methods). Cell Ontology annotations are required only in training, but using a trained SCimilarity model on new datasets requires neither author labels nor fine-tuning. For evaluation, we withheld 15 validation studies (13 scRNA-seq and 2 snRNA-seq) from training, comprising 1,415,962 cells with Cell Ontology annotations (Fig. 2). We excluded samples profiling tumours, cell lines or induced pluripotent stem cell-derived cells from the training and test sets, because their cell identity may be ambiguous.

### Loss functions for sensitive cell search

Testing 18 different parameter combinations for SCimilarity's objective function, varying the margin ( $\alpha$ ) and relative weighting of the reconstruction and triplet loss functions ( $\beta$ ) revealed that the two loss function components gave rise to different model behaviours (Extended Data Fig. 2a–c). Using the 15 validation studies, we assessed the models ability to search for cells similar to an input profile (query) and to mix similar cells across studies in a low-dimensional space (integration) (Extended Data Fig. 2b,c). We reasoned that a good similarity metric should both allow searching for similar cells and group together similar cells from different studies.

To evaluate querying, we compared searches with SCimilarity to gene signature scoring (Methods), aiming for a higher correlation between these two quantities (however, cell querying does not depend on predefined signatures or annotations). To evaluate integration across datasets, without the need to harmonize cell type annotations, we applied several benchmarks: an ontology-aware variation of average silhouette width<sup>16</sup> (ASW) and the established normalized mutual information (NMI), adjusted Rand index (ARI) and graph connectivity benchmarks, which measure the extent of study mixing within each cluster (Methods).

Models with higher reconstruction loss weighting (lower  $\beta$ ) performed better on the query task, whereas those with higher triplet-loss weighting (higher  $\beta$ ) scored higher on integration benchmarks (Extended Data Fig. 2c). Pure triplet loss ( $\beta = 1.0$ ) does not reliably preserve subtle cell state differences but does cluster cells of the same type closely together. MSE loss complements this by preserving subtle gene expression patterns. We selected a SCimilarity model that optimized the combined query and integration task scores ( $\beta = 0.001$  and margin = 0.05; Methods and Extended Data Fig. 2b,c).

For querying, SCimilarity's metric learning architecture more faithfully encoded cell similarities in the latent space than existing foundation models. SCimilarity's prediction of similarity to the query cell state matched the retrieval gene signature scores much more highly (Spearman's  $\rho = 0.77$ ) than previous foundation models ( $\rho = 0.54$  for scFoundation and  $\rho = 0.59$  for scGPT; Extended Data Fig. 2d) with far fewer cells incorrectly scored highly (Extended Data Fig. 2e).

For integration, we compared SCimilarity's pretrained representation to Harmony<sup>23</sup>, scVI<sup>10</sup>, scanorama<sup>24</sup> and scArches<sup>11</sup> on two kidney datasets<sup>25,26</sup>, two peripheral blood mononuclear cell (PBMC) datasets<sup>27,28</sup>, two lung datasets<sup>1,29</sup> and all 15 held-out datasets. In all four cases, SCimilarity had more coherent cell type clusters as measured by higher cell type ASW, comparable graph connectivity, but less mixing between studies in low dimensions (higher NMI, ARI and batch ASW; all measures of batchiness) (Fig. 2b), albeit comparable to many of these dedicated integration methods (which, by definition, see the test data in their training). As a negative control, SCimilarity, along with Harmony and scArches, did not artificially mix distinct B cell and regulatory T cell populations filtered from two different datasets (Extended Data Fig. 3g). Scanorama and scVI experienced such cross-population mixing. Notably, SCimilarity's integrated simply by embedding the cells in the common space without learning the integration from the data or fine tuning.

Thus, SCimilarity's loss function decouples faithful cell representation (query) from sample mixing (integration) and learns features that capture meaningful biology, reduce technical noise and generalizes to data held out of the training set.

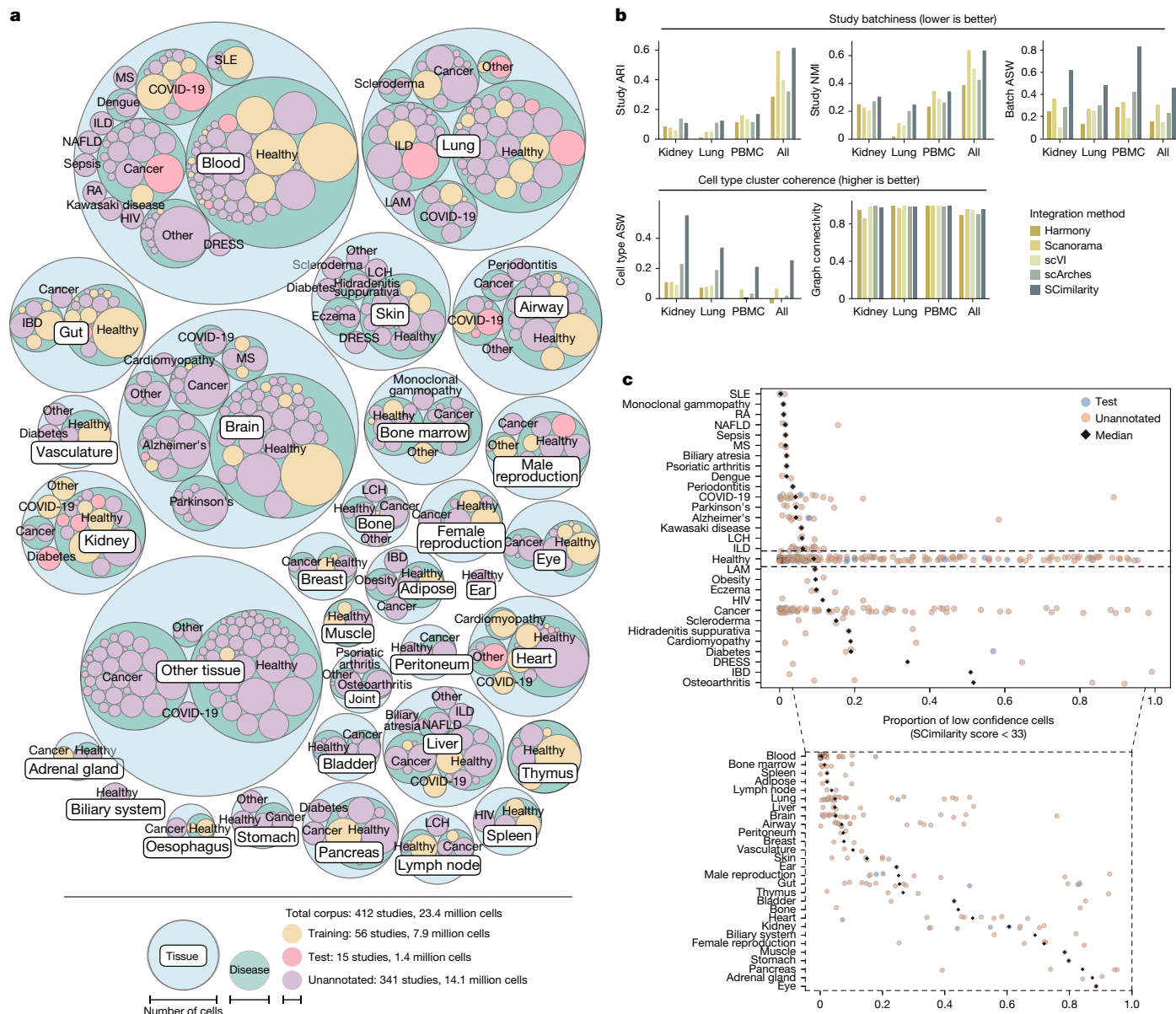
### Generalization across platforms

SCimilarity was trained on both scRNA-seq and scRNA-seq studies (Supplementary Table 1) and embeds both data types well, as demonstrated for profiles generated for the same human sample using multiple sc/snRNA-seq protocols<sup>30</sup>. Within SCimilarity-annotated cell types, the pairwise embedding distances were only slightly higher for nucleus-to-cell profile comparisons than for nucleus–nucleus or cell–cell distances (Extended Data Fig. 3a).

SCimilarity's learned representation also generalizes well to test data from multiple other profiling platforms, based on the embedding distances and annotation precision for a human PBMC sample that was profiled using seven platforms and chemistries<sup>31</sup> (10x Chromium v2, 10x Chromium v3, CEL-Seq2, Drop-Seq, Seq-well, SMART-Seq2 and InDrops) (Extended Data Fig. 3b–f). Data from all platforms were embedded effectively, although average within-platform nearest-neighbour embedding distances were slightly higher in non-10x platforms, with the highest distances for Seq-well and the non-UMI, full-length SMART-Seq2 data (Extended Data Fig. 3c,d). Cross-platform annotation precision was consistent for most cell types (except rare conventional (cDCs) and plasmacytoid dendritic cells) (Extended Data Fig. 3e,f). Thus, while SCimilarity was trained exclusively on 10x Genomics Chromium data, it effectively generalized to other single-cell profiling platforms.

### Integration without batch correction

SCimilarity quantifies a confidence level for each cell's representation, providing both outlier detection and an assessment of the



**Fig. 2 | SC similarity learns a universal representation that generalizes to new datasets.** **a**, A large-scale reference database of public gene expression datasets across tissues and diseases. The number of cells (circle size) across tissues (outermost light blue circles) and disease states (middle green circles) across individual studies (innermost circles) in the training (gold), test (pink) or unannotated (purple) datasets is shown. SLE, systemic lupus erythematosus; RA, rheumatoid arthritis; NAFLD, non-alcoholic fatty liver disease; MS, multiple sclerosis; LCH, Langerhans cell histiocytosis; LAM, lymphangioleiomyomatosis; IBD, inflammatory bowel disease. **b**, Benchmarking SC similarity against established data integration models. Ontology-aware ARI (study ARI, y axis,

representation's relevance to new data. Using SC similarity's score to quantify how distant a query cell is from the training data distribution provides a heuristic about the quality of the representation—a cell scoring highly similar to cells seen during training can be more confidently represented. Overall, 79.5% of in vivo holdout cells had high representation confidence. Tissue samples with low representation confidence, such as stomach ( $n = 0$  training studies), fetal gut ( $n = 1$ ) and bladder ( $n = 0$ ) were either absent or poorly represented in training (Fig. 2c and Methods). Similarly, 43.8% of in vitro cell profiles had low confidence due to poor matching to the training set (which excluded in vitro samples). Leveraging this ability, we assembled an atlas of

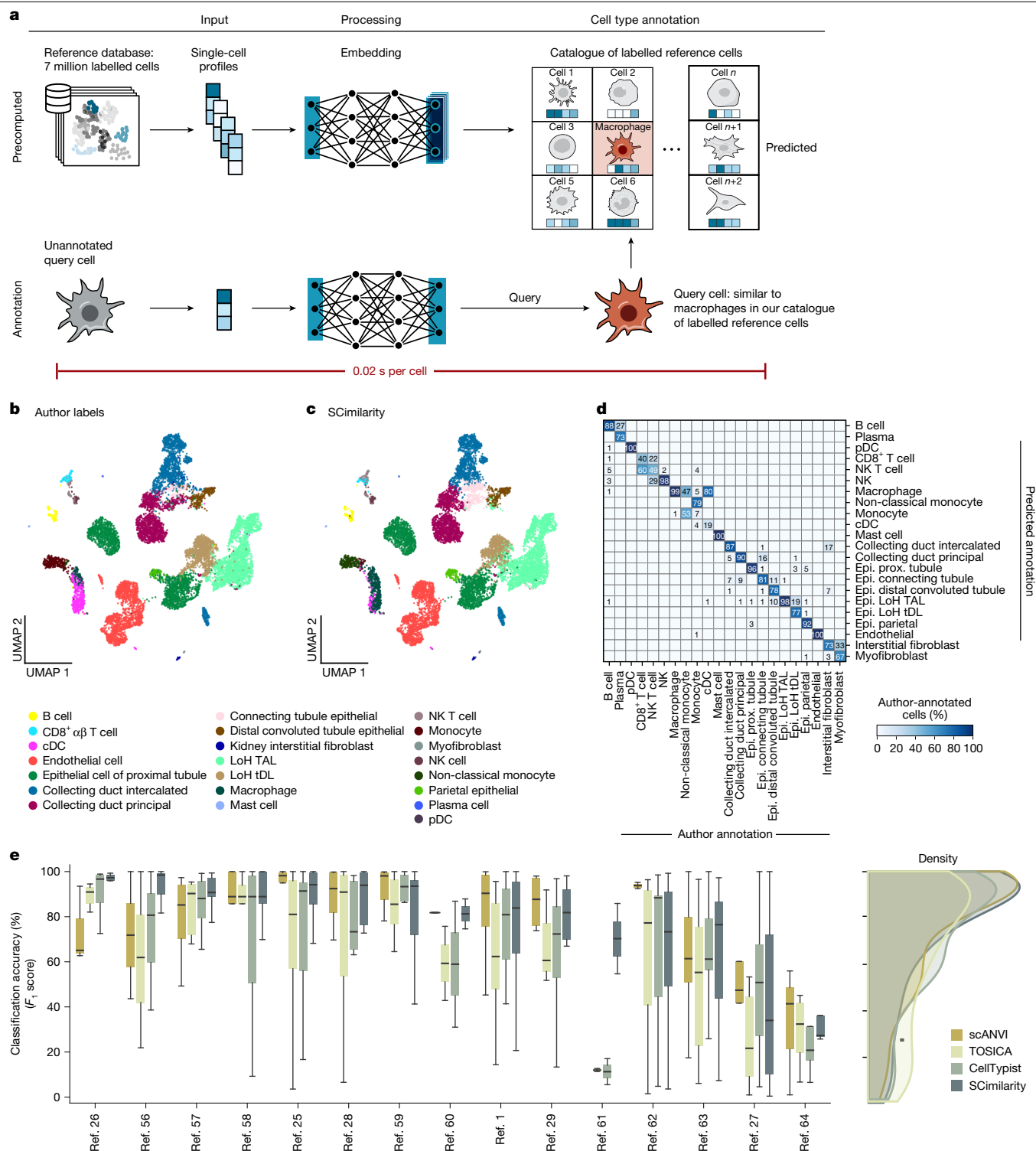
top left), NMI (study NMI, y axis, top middle), batch ASW (y axis, top right), cell type ASW (y axis, bottom left) and graph connectivity (y axis, bottom right) for different integration methods and SC similarity (coloured bars), each applied to integrate two kidney datasets, two lung datasets, two PBMC datasets and all 15 held out (test) datasets (x axis) are shown. **c**, SC similarity generalizes new datasets and flags outlier cells across different tissues and conditions. The fraction (x axis) of cells with low similarity to training data (SC similarity score of <50) in each study (points) from different diseases (y axis, top) or healthy tissues (y axis, bottom callout) is shown.

30 human tissues (Supplementary Table 2) and shared their embeddings as part of the SC similarity distribution.

### Cell type matching through similarity

SC similarity annotated query cell types by finding the cells in the annotated reference that are most similar to their profiles (Fig. 3a and Methods). This approach differs from established annotation methods because it (1) relies on a large, pan-body annotated cell repository; (2) uses a measure of expression similarity; and (3) annotates at the single-cell level rather than at the subset level. Thus, users can see which





**Fig. 3 | SCimilarity accurately annotates cell types across the human body.**

**a**, SCimilarity cell annotation. A new unannotated cell (grey, bottom left) is embedded in SCimilarity's common low-dimensional space and compared against the precomputed reference for cell type annotation (0.02 s per cell). **b–d**, SCimilarity annotation of a kidney scRNA-seq dataset. **b, c**, Uniform manifold approximation and projection (UMAP) embedding of cell profiles (dots) from SCimilarity's latent representation of a held-out kidney dataset<sup>25</sup> coloured by author-provided (**b**) or SCimilarity-predicted (**c**) cell type annotations. LoH TAL, loop of Henle thick ascending limb; LoH tDL, loop of Henle thin descending limb. **d**, The percentage (colour bar and number) of author-annotated cells

(columns) with each SCimilarity annotation (rows). **e**, Cell type annotation performance. Left, the accuracy (percentile  $F_1$  scores, higher is better; y axis) of SCimilarity and each of three annotation methods (colour bars) in matching author annotations in each of 15 test datasets (x axis) withheld from SCimilarity training. Right, the distribution of percentile  $F_1$  scores for each method (colour) across all 15 datasets. The box plots show the upper/lower quartiles (box limits), minimum/maximum values (whiskers) and median (centre line).  $F_1$  scores are calculated using a random sample of  $n = 10,000$  cells per study. Data from refs. 1,25–29,56–64. Epi., epithelial; prox., proximal; pDC, plasmacytoid DC.

individual cells, studies and tissues are driving the annotation. As each cell is annotated independently, no clustering is required. A user can annotate a cell's profile by comparing it to a desired subset (for example, for a tissue-specific query) or to the entire annotated cell reference. Finding the most similar cells is the same as retrieving the query cell's nearest neighbours. This is extremely efficient with hnswnlib<sup>32</sup>, where searching a precomputed approximate nearest-neighbour index of SCimilarity's full annotated reference takes just 20 ms (Methods).

A single, pretrained SCimilarity model annotated cell types competitively with tissue-specific models from established methods. For example, when limiting potential cell types to author-selected labels, 86.5% of SCimilarity's predicted labels from healthy kidney samples<sup>19</sup> matched the author-provided ones (Fig. 3b–d and Methods), comparable to the accuracy of scANVI (85.2%), CellTypist (90.4%) and TOSICA (87.2%) models directly trained on this dataset (Extended Data Fig. 4c–h). In closely related cells (monocytes versus macrophages versus DCs, fibroblasts versus myofibroblasts, natural killer (NK) cells versus NK T cells versus CD8<sup>+</sup> T cells), all methods showed considerable discordance with author-provided labels, suggesting that those annotations may be imprecise. Indeed, the author-annotated cDCs expressed a mix of macrophage (*CD68*, *CD163*, *CIQA*, *MS4A7*) and DC (*CD1C*, *CLEC9A*, *CLEC10A*, *FCERIA*) markers, and each method resolved this ambiguity differently (Extended Data Fig. 4i,j). SCimilarity also competitively recovered fine-grained author annotations supported by surface protein markers, performing on par or better than other methods across 22 immune cell subsets from a CITE-seq dataset that was held out of training<sup>33</sup> (Methods) with an annotation accuracy (75.3%) outperforming scANVI (52.2%), CellTypist (59.1%) and TOSICA (44.4%) (Extended Data Fig. 5a–i). Some closely related states (memory versus naive T cells; CD56<sup>bright</sup> versus CD56<sup>dim</sup> NK cells) were less precisely predicted by all methods, and may not be fully resolved by surface markers (Extended Data Fig. 5j). Similarly, author-provided and SCimilarity annotations matched well across all 15 test datasets, spanning 73 Cell Ontology terms, on par or better than other annotation methods (Fig. 3e).

We used SCimilarity's cell type assignment to rapidly annotate all 23.4 million cell profiles using one model, labelling 14,078,941 unannotated profiles and reannotating 9,302,209 author-annotated profiles (Methods) to a common set spanning 74 cell type labels across 21 coarse-grain lineages from 30 simplified tissue categories (Extended Data Fig. 6a).

## Interpretable features drive SCimilarity

To probe SCimilarity's model and annotations, we quantified the importance of each gene for each cell type using Integrated Gradients<sup>34</sup>—an explainability method that identifies the impact on model predictions from small disturbances to the input expression profiles (Methods). For example, the top gene attributions that distinguish lung alveolar type 2 (AT2) cells are surfactant genes *SFTPA2*, *SFTPA1*, *SFTPB* and *SFTPC*, consistent with known AT2 cell function<sup>35</sup>. SCimilarity learned these without previous knowledge of cell-type-specific genes, signatures or highly variable genes. Overall, SCimilarity's top importance genes agreed well with differentially expressed marker genes for 17 different matched types<sup>5</sup> with the exception of rare neuroendocrine cells ( $n = 90$  cells in training) (average area under the curve (AUC) = 0.84; Extended Data Fig. 6b and Supplementary Table 3).

## Cell search across tissues and diseases

We used SCimilarity's embedding to query for cells across the 23.4-million-cell reference (Fig. 4a), leveraging the fact that, with metric learning, the most similar cells are the nearest-neighbours of a query cell. As a query, the user can select either an individual cell profile or a centroid of multiple cell profiles. The SCimilarity software provides

tools for calculating query profiles, performing searches, filtering results by metadata and absolute distance, and evaluating the query and the results, including metrics to assess whether the query population is homogenous enough to yield reliable results, and how novel their query profile is (Methods).

As case studies, we focused on macrophages and fibroblasts in ILD, given their roles in tissue repair, regeneration and fibrosis<sup>36,37</sup>. In particular, recent scRNA-seq studies in many fibrotic diseases, including lung fibrosis, cancer, obesity and COVID-19, have reported seemingly similar *SPPI*<sup>+</sup> fibrosis-associated macrophage (FM) populations<sup>1,38–45</sup>. However, because each study defined them with different nomenclatures and gene signatures, it is unclear how similar they are and whether the same cells are broadly present across tissues, especially fibrotic conditions.

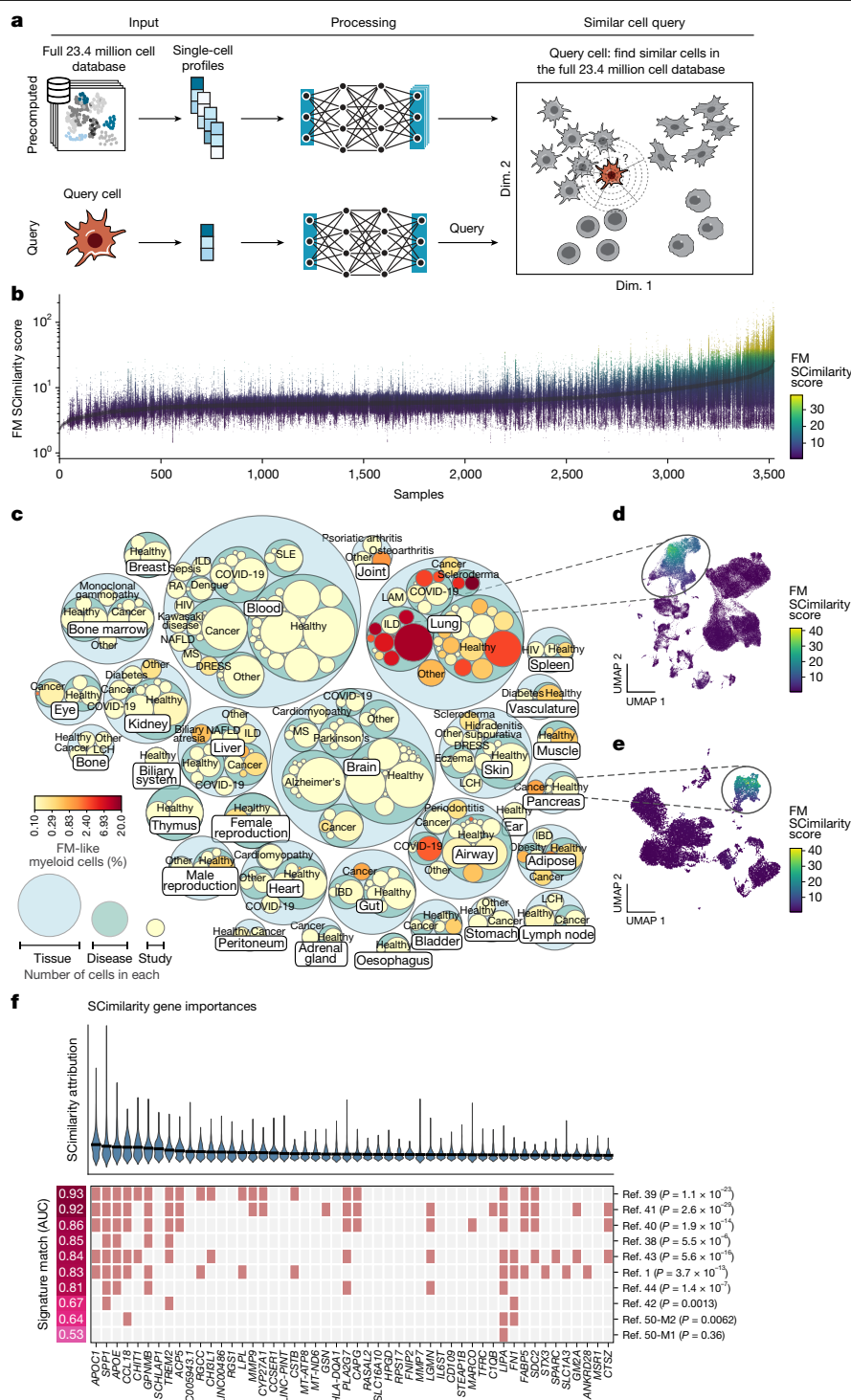
To study this, we searched our model with an FM cell profile across 2,507,171 in vivo cell profiles annotated by SCimilarity as monocytes or macrophages (Fig. 4a,b). As a query, we input the centroid of a macrophage cell subset<sup>1</sup> (query coherence: 94.7%), chosen using a gene signature of extracellular matrix remodelling and fibrosis-associated genes (*SPPI*, *TREM2*, *GPNMB*, *MMP9*, *CHIT1* and *CHI3L1*; Methods). In 2 s, SCimilarity computed the pairwise similarity of our query profile to each of the 2.5 million profiles (Fig. 4b). Alternatively, identifying the 10,000 cells with the highest SCimilarity score out of the 23.4-million-cell reference takes 0.05 s (Methods). By comparison, scoring each cell in the corpus with a literature-defined FM gene signature took 2 h and 46 min (not shown). The gene signature and SCimilarity scores are broadly correlated ( $r = 0.50$ ,  $P < 10^{-300}$ ; Extended Data Fig. 8a–c), showing that this granular cell state, not just the cell type, is well represented in SCimilarity's query score and embedding.

The SCimilarity search showed that FMs are common in ILD lung samples, as well as present in some cancers, including uveal melanoma, pancreatic ductal adenocarcinoma (PDAC) and colon cancer (Fig. 4c–e and Supplementary Table 4). Of the top 1% of monocytes and macrophages most similar to our query, 93.7% were from lung tissue and 81.2% from ILD and COVID-19 lung samples. The prevalence of FM-like cells in the lung varied by disease: FM-like cells were 20% and 4% of monocytes and macrophages in two systemic sclerosis (SSc) studies, 6.6% on average (s.d. = 4.8%) across 12 ILD studies (excluding SSc), 0.97% on average across six COVID-19 lung studies (s.d. = 0.25%, 0% in non-lung COVID-19 data) and 0.40% in 22 lung studies annotated as healthy, normal or with no disease annotation (s.d. = 0.15%). While abundant in SSc lungs, FM-like cells were much rarer (0.14% of myeloid cells) in SSc skin<sup>46</sup>. Notably, there were some FM-like cells in other fibrotic diseases and tissues, such as one primary PDAC tumour<sup>47</sup> (0.85% of 1,171 myeloid cells) and one liver metastasis<sup>48</sup> of PDAC (0.5% of 1,199 cells). Thus, while our query was derived from IPF samples, it identified FM-like cells in many contexts, confirming previous observations of FMs in lung injury<sup>45,49</sup> and suggesting a role for FM-like cells across other organs and diseases.

Searching for multiple cell states helps relate them across tissues, as we found by querying a fibrosis-associated myofibroblast query profile, defined as the centroid of cells<sup>1</sup> expressing a corresponding gene signature (*ACTA2*, *CDH11*, *ELN*, *LOXL1*, *TNC*, *ASPN*, *COMP*, *CTHRC1*, *POSTN*, *COL1A1*, *COL3A1* and *COL8A1*; query coherence: 77.0%). SCimilarity distances were substantially more correlated with the myofibroblast gene signature scores ( $\rho = 0.36$ ) compared with those of scGPT ( $\rho = -0.19$ ) and scFoundation ( $\rho = -0.17$ ) (Extended Data Fig. 7c) and captured relevant cell types more specifically (Extended Data Fig. 7d). The presence of myofibroblasts correlated with the presence of FMs in other ILD datasets, COVID-19 and PDAC ( $r^2 = 0.48$ ; Extended Data Fig. 7a,b).

## Important FM features match known signatures

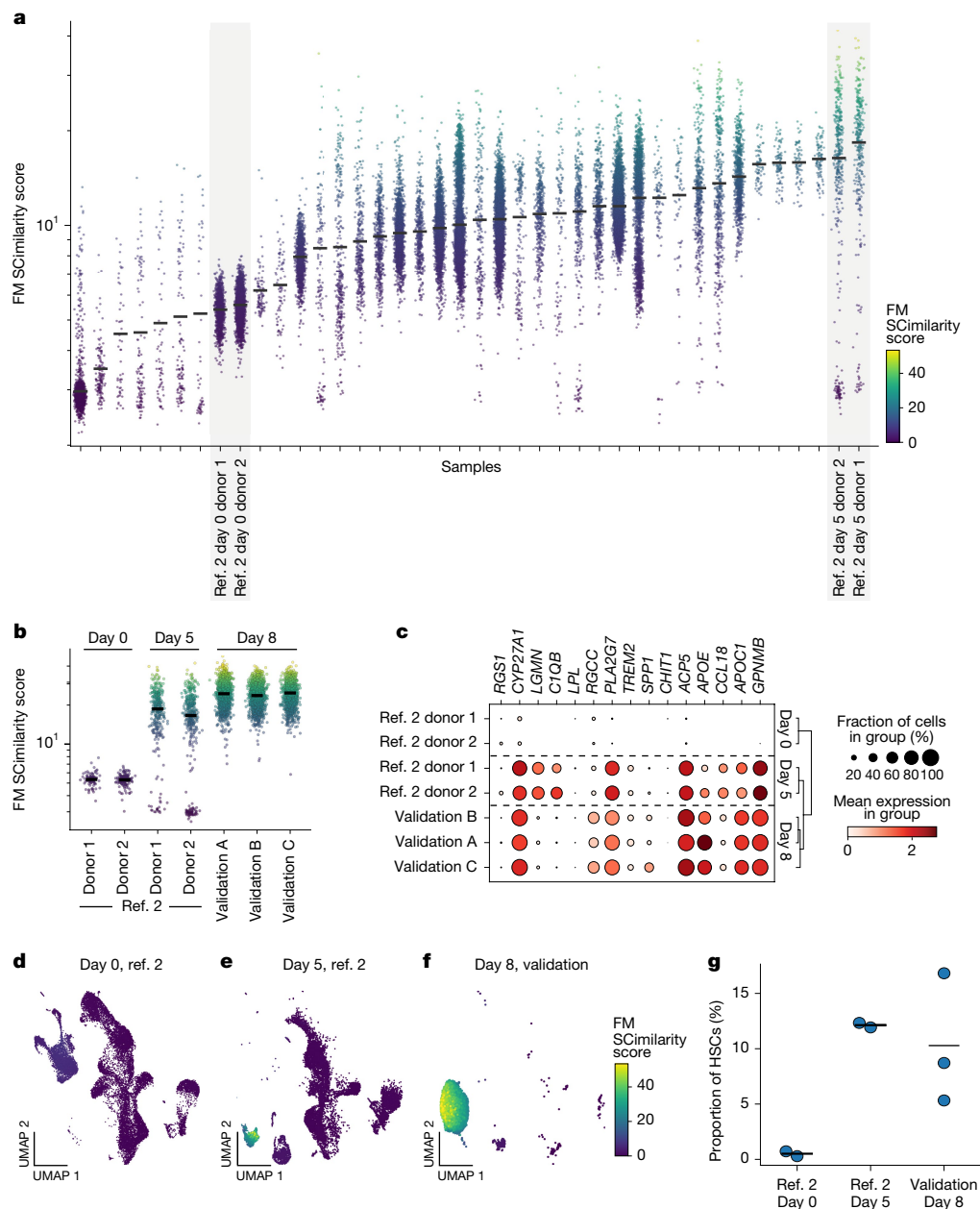
We hypothesized that SCimilarity's detection of FM-like cells across ILD studies reflects a shared biological state, despite varying markers



**Fig. 4 | SCSimilarity cell search reveals FMs across ILD and other diseases.**

**a**, SCSimilarity cell search. A query cell profile (bottom left) is embedded into the SCSimilarity representation with 23.4 million reference cells. Its nearest neighbours by distance are tabulated by study, tissue and disease. **b–e**, Identification of FMs across tissues. **b**, SCSimilarity scores (y axis, log<sub>10</sub> scale and colour bar) against an FM query profile for all monocytes and macrophages (dots) from 1,041 in vivo tissue samples from 143 studies (x axis), ordered by the mean SCSimilarity score. **c**, The number of cells (circle size) across tissues (outermost light blue circles), disease states (middle green circles) and individual studies (innermost circles, coloured by the fraction of monocytes and macrophages with SCSimilarity scores >99th percentile of all FM SCSimilarity scores (log-scaled colour bar)). Circle sizes for disease and individual study are scaled relative to other diseases in the same tissue or studies in the same disease. **d,e**, UMAP of all single-cell profiles (macrophages and otherwise, dots) from the

SCSimilarity representation for ILD<sup>40</sup> (**d**) and PDAC<sup>47</sup> (**e**) studies, coloured by FM query SCSimilarity scores (colour bar). **f**, SCSimilarity's explainability framework scores FM-associated genes by importance. The distribution of Integrated Gradients attribution scores (y axis, top; horizontal bars show the mean) for genes (x axis, top; columns, bottom) with the top 50 scores for FMs versus lung macrophages and their membership (red, presence; grey, absence) in published macrophage signatures (bottom, rows). The left colour bar represents the AUC for the attribute score match to published signatures. The signature publication source and *P* value (two-sided Mann–Whitney *U*-tests; in signature > not in signature) across the top 3,000 genes by mean attribution score are shown on the right. Attribution scores, AUC values and *P* values were calculated using the *n* = 500 cells most similar to FMs against *n* = 500 randomly sampled cells from the full *n* = 2,578,221 cell monocyte and macrophage query set.



**Fig. 5 | SCSimilarity cell search identifies in vitro cells matching an in vivo FM state and a novel in vitro disease model.** **a**, Identification of FM-like cells across in vitro samples with a SCSimilarity cell search. SCSimilarity scores (y axis, log<sub>10</sub> scale, colour bar) against a FM query profile for each annotated myeloid cell (dot) from  $n = 40$  in vitro samples (x axis) from  $n = 17$  studies, ordered by the mean SCSimilarity score. The grey boxes show day 0 and day 5 samples from a 3D-hydrogel culture system<sup>2</sup>. **b–f**, 3D conditions yield FM-like cells in vitro in validation experiments. **b**, SCSimilarity scores (y axis, log<sub>10</sub> scale, colour bar) against a FM query profile for each annotated myeloid cell (dot) in the original 3D-hydrogel culture system dataset<sup>2</sup> from  $n = 2$  independent donors at day 0 and day 5 and from  $n = 3$  independent donors in the day 8 validation experiment

and nomenclature. To explore this, we used Integrated Gradients to quantify gene importance in distinguishing FMs (Methods), yielding genes enriched in fibrotic processes (for example, *MMP7*, *FNI*), lipid metabolism (such as *APOE*, *LPL*) and damage recognition (for example, *MARCO*, *MSR1*) (Fig. 4f, Extended Data Fig. 8d and Supplementary Table 5). These include known markers (*TREM2*) and novel genes (*HLA-DQA1* and *RGS1*) with higher detection rates in FM-like cells (Extended Data Fig. 8e–g).

(x axis). **c**, The mean expression (dot colour) and percentage of cells (dot size) expressing genes (rows) with a high SCSimilarity attribution score for distinguishing FMs in vivo (as in **f**) in myeloid cells in the original 3D-hydrogel culture system<sup>2</sup> and in the validation experiment (columns). **d–f**, UMAP embedding from SCSimilarity's query model latent space of cell profiles (dots) from day 0 (**d**) or day 5 (**e**) of the original 3D-hydrogel culture system<sup>2</sup>, or from day 8 (**f**) of the replication experiment, coloured by FM SCSimilarity score (colour bar). **g**, Replication of original finding of HSC expansion. The proportion of HSCs in  $n = 2$  donors from ref. 2 at day 0 and day 5 and  $n = 3$  donors from the day 8 validation experiment.

The most important genes significantly overlapped with published gene signatures describing similar macrophage populations or with genes of which the differential expression defined each study's macrophage population of interest (Supplementary Table 6). Published signatures derived from seven studies had a high signature match (AUC > 0.8), while negative control signatures of M2 and M1 macrophages<sup>50</sup> ranked in the bottom three (AUC = 0.64 ( $P = 0.0062$ ) and 0.53 ( $P = 0.36$ ), respectively; Fig. 4f).



## Search for ex vivo human cell model

Researching the role of novel cell states like FMs in disease requires modelling, perturbing and studying them in vitro, but identifying culture conditions remains challenging. To address this, we used SCimilarity to find FM-like cells within in vitro samples. After relaxing the SCimilarity score threshold to account for differences between in vitro and in vivo cells, we identified 41,926 monocytes and macrophages from 40 samples across 17 studies, from lung organoids to ex vivo treated leukaemia cells<sup>51</sup>, to stimulated PBMCs<sup>52</sup>.

The cells most similar to our query were from PBMCs cultured for 5 days in a 3D hydrogel system designed to expand haematopoietic stem cells (HSCs)<sup>2</sup> (Fig. 5a and Supplementary Table 7). This was a surprising result, as this study was unrelated to lung biology, the cells are rare in peripheral blood and there were no findings reported about myeloid cells. While no FM-like cells were present among myeloid cells on day 0, 15% of cells grown for five or more days were similar to FMs (SCimilarity score of >25) and expressed *TREM2*, *GPNMB*, *CCL18* and *MMP9* (Fig. 5b–e).

We validated SCimilarity's prediction by experimentally replicating the 3D hydrogel system<sup>2</sup> and profiling cultured PBMCs by scRNA-seq (Fig. 5b,c,f). While the relative cellular abundances differed between the original day 5 data<sup>2</sup> and our day 8 replication (Methods), 10.1% of all cells in the day 8 experiment were predicted as HSCs by SCimilarity (Fig. 5g), and 41.5% of the myeloid cells were predicted as FM-like macrophages (Fig. 5b,f;  $n = 3$  donors; 37.1%, 42.5% and 44.9%; SCimilarity score > 25) and enrichment for FM hallmark genes, such as *CCL18*, *GPNMB*, *SPPI* and *TREM2* (Fig. 5c). This demonstrates SCimilarity's ability to interrogate publicly available data at scale, query a reference of in vivo and in vitro data for biologically similar conditions, and help to identify experimental conditions to reproduce those results in the laboratory.

## Discussion

SCimilarity offers a unique approach based on metric learning for cell searches across hundreds of studies, thousands of samples and tens of millions (and more) of cells. Query cell states can be defined based on an individual cell profile (although these may lack robustness), metacells<sup>53</sup>, clusters or a group of highly similar cells defined by a gene signature. To ensure reliable results, SCimilarity assesses a query's coherence and the model's confidence in the cell's representation. Using a cell's full expression profile captures its full complexity, bypassing the need for curated (and biased) gene signatures. SCimilarity can generate a robust signature for a cell state using an explainability technique. As public data are diverse and different biological questions may have different assumptions, SCimilarity enables users to make case-by-case decisions on proper study, sample, or cell filtering and SCimilarity score cut-offs appropriate to their investigation. To ensure high quality, we have removed any sample duplication across training and test sets; however, there are duplicated samples within our full reference dataset as a consequence of including published datasets in toto. We made SCimilarity available as an open-sourced Python API with tutorials for querying, embedding, annotating and ranking cell profiles. The API facilitates tailored queries by  $k$ -nearest neighbours ( $k$ -NN), exhaustive searches, metadata filtering, score-based filtering and visualization tools, and each query result is traceable to the original dataset for further analysis.

SCimilarity's cell queries open the way for systematic exploration of transcriptionally similar populations across the vast Human Cell Atlas by showing that an identified population is reproducibly present in other studies<sup>54</sup>; connecting results from independent studies, such as observational and functional ones; and identifying contexts in which the same population may be active. We illustrated this with our search for FM-like cells across the atlas, leading to explanatory marker genes, a cell culture system that elicits a similar state in vitro, and identification of similar cells in other fibrotic lung diseases, COVID-19 and multiple

tumour types (especially PDAC<sup>55</sup>), suggesting a broader role for these cells in the damage response and tissue remodelling processes. Notably, previous foundation models did not perform well on identifying cells similar to FM or myofibroblasts, both expanding to less similar cells, and missing more similar ones.

As SCimilarity can generalize to cells and datasets not seen in the training, cell profiles can be filtered or added without recomputing the existing embeddings. Downstream tasks, such as cell type annotation, cell queries and gene signature derivation all are simplified using SCimilarity's generalized representation and can be applied to cells not seen during training without informing the model about the importance or variability of specific genes. We trained SCimilarity on both scRNA-seq and snRNA-seq data collected by 10x Genomics Chromium data (of varying tissue coverage) and it was able to handle test data from profiles collected by other scRNA-seq platforms that were not included in training. Nevertheless, users should always interpret cross-technology integrations with care. The strong performance of SCimilarity's learned representation for both the integration and querying tasks may suggest that it can perform well for other tasks, but these need to be assessed in future studies.

By training on Cell Ontology annotations from many published studies, SCimilarity learns a consensus of how experts define a given cell type. For annotation tasks, the set of labels that SCimilarity can predict is by necessarily limited by available Cell Ontology terms and experimental observation of cell states across studies. Conversely, cell querying is annotation independent, and can use any profile, irrespective of whether the cell state is in the Cell Ontology or observed in training. Note that we deliberately withheld cancer cells and cell lines from training due to lack of clear cell type identity and these may not be well represented in the current model. In our experience, we see poor performance on fetal samples, granulocytes, haematopoietic stem and progenitor cells, and intermediate precursor cell states, probably because most training data were sourced from adult tissues and due to ambiguity in lineage commitment of non-differentiated populations, respectively. While SCimilarity's API provides guidance to assess the coherence of a query cell profile, the quality of query results ultimately depends on the assumptions and quality of the input profile. An input cell profile can be derived from a single cell, the centroid of a cluster, or an aggregation of cells scored and filtered by a user-defined gene signature—all of which require some subjective selections that can influence downstream analyses. As larger SCimilarity representations are trained on the growing Human Cell Atlas, the model will allow querying and searches on expanded swaths of human biology.

## Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41586-024-08411-y>.

1. Adams, T. S. et al. Single-cell RNA-seq reveals ectopic and aberrant lung-resident cell populations in idiopathic pulmonary fibrosis. *Sci. Adv.* **6**, eaba1983 (2020).
2. Xu, Y. et al. Efficient expansion of rare human circulating hematopoietic stem/progenitor cells in steady-state blood using a polypeptide-forming 3D culture. *Protein Cell* <https://doi.org/10.1007/s13238-021-00900-4> (2022).
3. Rood, J. E., Maartens, A., Hupalowska, A., Teichmann, S. A. & Regev, A. Impact of the Human Cell Atlas on medicine. *Nat. Med.* **28**, 2486–2496 (2022).
4. Rosen, Y. et al. Toward universal cell embeddings: integrating single-cell RNA-seq datasets across species with SATURN. *Nat. Methods* **21**, 1492–1500 (2024).
5. Cui, H. et al. scGPT: toward building a foundation model for single-cell multi-omics using generative AI. *Nat. Methods* **21**, 1470–1480 (2024).
6. Theodoris, C. V. et al. Transfer learning enables predictions in network biology. *Nature* <https://doi.org/10.1038/s41586-023-06139-9> (2023).
7. Shen, H. et al. Generative pretraining from large-scale transcriptomes for single-cell deciphering. *iScience* **26**, 106536 (2023).

8. Eraslan, G. et al. Single-nucleus cross-tissue molecular reference maps toward understanding disease gene function. *Science* **376**, eabl4290 (2022).
9. Heimberg, G., Bhatnagar, R., El-Samad, H. & Thomson, M. Low dimensionality in gene expression data enables the accurate extraction of transcriptional programs from shallow sequencing. *Cell Syst.* **2**, 239–250 (2016).
10. Lopez, R., Regier, J., Cole, M. B., Jordan, M. I. & Yosef, N. Deep generative modeling for single-cell transcriptomics. *Nat. Methods* **15**, 1053–1058 (2018).
11. Lotfollahi, M. et al. Mapping single-cell data to reference atlases by transfer learning. *Nat. Biotechnol.* **40**, 121–130 (2022).
12. Schroff, F., Kalenichenko, D. & Philbin, J. FaceNet: a unified embedding for face recognition and clustering. In *Proc. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 815–823 (IEEE, 2015).
13. Simon, L. M., Wang, Y.-Y. & Zhao, Z. Integration of millions of transcriptomes using batch-aware triplet neural networks. *Nat. Mach. Intell.* **3**, 705–715 (2021).
14. Yang, M. et al. Contrastive learning enables rapid mapping to multimodal single-cell atlas of multimillion scale. *Nat. Mach. Intell.* **4**, 696–709 (2022).
15. Yu, X., Xu, X., Zhang, J. & Li, X. Batch alignment of single-cell transcriptomics data using deep metric learning. *Nat. Commun.* **14**, 960 (2023).
16. Luecken, M. D. et al. Benchmarking atlas-level data integration in single-cell genomics. *Nat. Methods* **19**, 41–50 (2022).
17. Diehl, A. D. et al. The Cell Ontology 2016: enhanced content, modularization, and ontology interoperability. *J. Biomed. Semantics* **7**, 44 (2016).
18. Edgar, R., Domrachev, M. & Lash, A. E. Gene Expression Omnibus: NCBJ gene expression and hybridization array data repository. *Nucleic Acids Res.* **30**, 207–210 (2002).
19. CZ CELLxGENE Discover. CELLxGENE data portal (Chan Zuckerberg Initiative, 2022); <https://cellxgene.cziscience.com/collections/a98b828a-622a-483a-80e0-15703678befd>.
20. Zheng, G. X. Y. et al. Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* **8**, 14049 (2017).
21. Gremse, M. et al. The BRENDA Tissue Ontology (BTO): the first all-integrating ontology of all organisms for enzyme sources. *Nucleic Acids Res.* **39**, D507–D513 (2011).
22. Schriml, L. M. et al. Human Disease Ontology 2018 update: classification, content and workflow expansion. *Nucleic Acids Res.* **47**, D955–D962 (2019).
23. Korsunsky, I. et al. Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat. Methods* **16**, 1289–1296 (2019).
24. Hie, B., Bryson, B. & Berger, B. Efficient integration of heterogeneous single-cell transcriptomes using Scanorama. *Nat. Biotechnol.* **37**, 685–691 (2019).
25. Kretzler, M., Otto, E., O'Connor, C., Bitzer, M. & Menon, R. HCA seed network precise tumor-nephrectomy samples. *Human Cell Atlas Data Portal* <https://explore.data.humancellatlas.org/projects/29ed827b-c539-4f4c-bb6b-ce8f9173dfb7> (2022).
26. Muto, Y. et al. Single cell transcriptional and chromatin accessibility profiling redefine cellular heterogeneity in the adult human kidney. *Nat. Commun.* **12**, 2190 (2021).
27. Deng, Q. et al. Characteristics of anti-CD19 CAR T cell infusion products associated with efficacy and toxicity in patients with large B cell lymphomas. *Nat. Med.* **26**, 1878–1887 (2020).
28. Szabo, P. A. et al. Longitudinal profiling of respiratory and systemic immune responses reveals myeloid cell-driven lung inflammation in severe COVID-19. *Immunity* **54**, 797–814 (2021).
29. Vieira Braga, F. A. et al. A cellular census of human lungs identifies novel cell states in health and in asthma. *Nat. Med.* **25**, 1153–1163 (2019).
30. Slyper, M. et al. A single-cell and single-nucleus RNA-Seq toolbox for fresh and frozen human tumors. *Nat. Med.* **26**, 792–802 (2020).
31. Ding, J. et al. Systematic comparison of single-cell and single-nucleus RNA-sequencing methods. *Nat. Biotechnol.* **38**, 737–746 (2020).
32. Malkov, Y. A. & Yashunin, D. A. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. *IEEE Trans. Pattern Anal. Mach. Intell.* **42**, 824–836 (2020).
33. Chan Zuckerberg Initiative Single-Cell COVID-19 Consortia et al. Single cell profiling of COVID-19 patients: an international data resource from multiple tissues. Preprint at *bioRxiv* <https://doi.org/10.1101/2020.11.20.20227355> (2020).
34. Sundararajan, M., Taly, A. & Yan, Q. Axiomatic attribution for deep networks. In *Proc. 34th International Conference on Machine Learning* Vol. 70 (eds Precup, D. & Teh, Y. W.) 3319–3328 (PMLR, 2017).
35. Beers, M. F. & Moodley, Y. When is an alveolar type 2 cell an alveolar type 2 cell? A conundrum for lung stem cell biology and regenerative medicine. *Am. J. Respir. Cell Mol. Biol.* **57**, 18–27 (2017).
36. Wynn, T. A. & Vannella, K. M. Macrophages in tissue repair, regeneration, and fibrosis. *Immunity* **44**, 450–462 (2016).
37. Lis-López, L., Bauset, C., Seco-Cervera, M. & Cosin-Roger, J. Is the macrophage phenotype determinant for fibrosis development? *Biomedicines* **9**, 1747 (2021).
38. Ayaub, E. A. et al. Single cell RNA-seq and mass cytometry reveals a novel and a targetable population of macrophages in idiopathic pulmonary fibrosis. Preprint at *bioRxiv* <https://doi.org/10.1101/2021.01.04.425268> (2021).
39. Jaitin, D. A. et al. Lipid-associated macrophages control metabolic homeostasis in a Trem2-dependent manner. *Cell* **178**, 686–698 (2019).
40. Morse, C. et al. Proliferating SPPI/MERTK-expressing macrophages in idiopathic pulmonary fibrosis. *Eur. Respir. J.* **54**, 1802441 (2019).
41. Mulder, K. et al. Cross-tissue single-cell landscape of human monocytes and macrophages in health and disease. *Immunity* **54**, 1883–1900 (2021).
42. Ramachandran, P. et al. Resolving the fibrotic niche of human liver cirrhosis at single-cell level. *Nature* **575**, 512–518 (2019).
43. Reyfman, P. A. et al. Single-cell transcriptomic analysis of human lung provides insights into the pathobiology of pulmonary fibrosis. *Am. J. Respir. Crit. Care Med.* **199**, 1517–1536 (2019).
44. Wendisch, D. et al. SARS-CoV-2 infection triggers profibrotic macrophage responses and lung fibrosis. *Cell* **184**, 6243–6261 (2021).
45. Gao, X. et al. Osteopontin links myeloid activation and disease progression in systemic sclerosis. *Cell Rep. Med.* **1**, 100140 (2020).
46. Mirizio, E. et al. Single-cell transcriptome conservation in a comparative analysis of fresh and cryopreserved human skin tissue: pilot in localized scleroderma. *Arthritis Res. Ther.* **22**, 263 (2020).
47. Lin, W. et al. Single-cell transcriptome analysis of tumor and stromal compartments of pancreatic ductal adenocarcinoma primary tumors and metastatic lesions. *Genome Med.* **12**, 80 (2020).
48. Kemp, S. B. et al. Pancreatic cancer is marked by complement-high blood monocytes and tumor-associated macrophages. *Life Sci. Alliance* **4**, e202000935 (2021).
49. Bhattacharya, M. Insights from transcriptomics: CD163<sup>+</sup> profibrotic lung macrophages in COVID-19. *Am. J. Respir. Cell Mol. Biol.* **67**, 520–527 (2022).
50. Martinez, F. O., Gordon, S., Locati, M. & Mantovani, A. Transcriptional profiling of the human monocyte-to-macrophage differentiation and polarization: new molecules and patterns of gene expression. *J. Immunol.* **177**, 7303–7311 (2006).
51. Duy, C. et al. Chemotherapy induces senescence-like resilient cells capable of initiating AML recurrence. *Cancer Discov.* **11**, 1542–1561 (2021).
52. Karagiannis, T. T. et al. Single cell transcriptomics reveals opioid usage evokes widespread suppression of antiviral gene program. *Nat. Commun.* **11**, 2611 (2020).
53. Baran, Y. et al. MetaCell: analysis of single-cell RNA-seq data using K-nn graph partitions. *Genome Biol.* **20**, 206 (2019).
54. Regev, A. et al. The Human Cell Atlas. *eLife* **6**, e27041 (2017).
55. Liou, G.-Y. Inflammatory cytokine signaling during development of pancreatic and prostate cancers. *J. Immunol. Res.* **2017**, 7979637 (2017).
56. Kuppe, C. et al. Spatial multi-omic map of human myocardial infarction. *Nature* **608**, 766–777 (2022).
57. Wilson, P. C. et al. Multimodal single cell sequencing implicates chromatin accessibility and genetic background in diabetic kidney disease progression. *Nat. Commun.* **13**, 5253 (2022).
58. Olah, M. et al. Single cell RNA sequencing of human microglia uncovers a subset associated with Alzheimer's disease. *Nat. Commun.* **11**, 6129 (2020).
59. Van Der Wijst, M. G. P. et al. Type I interferon autoantibodies are associated with systemic immune alterations in patients with COVID-19. *Sci. Transl. Med.* **13**, eab2624 (2021).
60. Szabo, P. A. et al. Single-cell transcriptomics of human T cells reveals tissue and activation signatures in health and disease. *Nat. Commun.* **10**, 4706 (2019).
61. Ravindra, N. G. et al. Single-cell longitudinal analysis of SARS-CoV-2 infection in human airway epithelium identifies target cells, alterations in gene expression, and cell state changes. *PLoS Biol.* **19**, e3001143 (2021).
62. Henry, G. H. et al. A cellular anatomy of the normal adult human prostate and prostatic urethra. *Cell Rep.* **25**, 3530–3542.e5 (2018).
63. Fawcner-Corbett, D. et al. Spatiotemporal analysis of human intestinal development at single-cell resolution. *Cell* **184**, 810–826.e23 (2021).
64. Cano-Gamez, E. et al. Single-cell transcriptomics identifies an effectorness gradient shaping the response of CD4<sup>+</sup> T cells to cytokines. *Nat. Commun.* **11**, 1801 (2020).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2024

## Methods

### SCimilarity model design

**Model architecture.** The SCimilarity model consists of one fully connected encoder and one decoder stage and reuses the same encoding network three times per training triplet, such that updates to the model after each batch are shared equally for each subsequent batch of training triplets. The decoder stage is not part of the conventional triplet loss architecture, but is included to compute a MSE reconstruction loss.

Expression profiles are reduced through an encoder network, starting from 28,231 genes through four hidden layers with dimensions 1,024, 1,024, 1,024 and 128. The 128-dimensional outputs are unit length normalized, forcing all low-dimensional cell representations to lie on the surface of a hypersphere. During training, the input layer is subjected to 40% dropout, zeroing out many gene expression values at random and each hidden layer is subjected to 50% dropout rates for maximum regularization<sup>65</sup>.

While hyperspheric spaces have been infrequently used for representation of single-cell profiles<sup>66</sup>, the triplet-loss model often uses hypersphere embeddings to ensure consistency between the model hyperparameters<sup>12</sup>. During triplet-loss training, the objective is to place cells of different types sufficiently far apart. The minimum desired distance between cells of different types is called the margin. By fixing the volume of the embedding space to the surface of a unit length 128-dimensional hypersphere, the margin is interpreted consistently between model runs. Without normalization, cells can be placed up to an infinite distance apart, rendering the margin meaningless.

**Triplet-loss training.** To learn features that place datapoints considered similar near each other, the loss function depends on distances between data points embedded in a learned low dimensional latent space, described with equation (1):

$$d(\mathbf{x}, \mathbf{y}) = \|f(\mathbf{x}) - f(\mathbf{y})\|_2^2 \quad (1)$$

where  $\mathbf{x}$  and  $\mathbf{y}$  are two high-dimensional vectors (here, cell profiles), passed through a neural network encoder  $f()$ .

The triplet-loss model learns from three vectors at a time: the anchor ( $\mathbf{x}_i^a$ ), positive ( $\mathbf{x}_i^p$ ) and negative ( $\mathbf{x}_i^n$ ). The anchor and positive vectors are considered to be similar, whereas the anchor and negative are dissimilar.

The model parameters are iteratively updated to decrease the number of triplets where the distance between the anchor and negative data vectors is insufficiently large relative to the distance between the anchor and the positive points, therefore minimizing the triplet-loss function defined in equation (2):

$$L_{\text{triplet}} = \frac{\sum_i^N \max(d(\mathbf{x}_i^a, \mathbf{x}_i^p) - d(\mathbf{x}_i^a, \mathbf{x}_i^n) + \alpha, 0)}{N} \quad (2)$$

where  $\alpha$  is the margin, which denotes how much further the negatives should be from the anchor than the positives, and  $i$  is the index of the triplet.

**Reconstruction loss training.** The reconstruction loss is computed on the anchor cell only, because each anchor cell is used only once as an anchor within a batch. The reconstruction loss is defined in equation (3):

$$L_{\text{MSE}} = \frac{\sum_i^N \|\mathbf{x}_i^a - g(f(\mathbf{x}_i^a))\|_2^2}{N} \quad (3)$$

where  $N$  is the number of anchor cells in a batch, set to  $N = 1,000$  in SCimilarity, and  $g()$  is the function learned by the neural network decoder stage.

**Combined loss function.** Adding a reconstruction loss to classification models has been shown to improve generalization<sup>67</sup> through a regularization effect. The SCimilarity loss function combines the triplet loss and reconstruction loss functions as defined in equation (4):

$$L = (1 - \beta) \times L_{\text{MSE}} + \beta \times L_{\text{triplet}} \quad (4)$$

where  $\beta$  is a weighting term in  $[0, 1]$ . Training and validation curves for triplet loss, reconstruction loss and the percentage of hard triplets were constructed for varying values of  $\beta$  in  $[0, 1]$  (Extended Data Fig. 2a), where  $\beta = 0$  corresponds to a conventional autoencoder and  $\beta = 1$  corresponds to a pure triplet-loss model. Empirically,  $\beta = 0.001$  performed best on the cell search task (query model) and  $\beta = 1$  performed best on batch integration (Extended Data Fig. 2c).

### Cell Ontology terms and relationships

Authors may annotate cell types at different granularities, which confounds triplet sampling by introducing cell type annotations with hierarchical relationships that cannot be unambiguously defined as either similar or dissimilar. As such, cell type annotations used for training are defined using standardized Cell Ontology terms and valid triplets are restricted to cells without vertical Cell Ontology relationships between the members of the triplet. A vertical relationship is defined as any directed path of one or more ancestor–descendant relationships in the Cell Ontology network. Thus, there are three binary relationships defined for annotation: (1) similar pairs with identical annotations (for example, T cell and T cell); (2) dissimilar pairs with non-vertical ontology relationships (for example, 'CD4-positive,  $\alpha\beta$  T cell' and 'CD8-positive,  $\alpha\beta$  T cell'); and (3) ambiguous pairs with vertical relationships (for example, 'T cell' and 'CD4-positive,  $\alpha\beta$  T cell'). Positives are drawn from cells similar to the anchor, negatives are drawn from cells dissimilar to the anchor and cells that are ambiguous to the anchor are excluded from sampling.

### GEO data aggregation

In total, 334 human sc/snRNA-seq datasets were obtained from the GEO<sup>18</sup>. Multiple filtering steps were used to restrict the datasets analysed to samples from human tissue that were generated using the 10x Chromium platform and that reported unnormalized gene count data that could be automatically processed. To select appropriate datasets, search criteria were designed for the Biopython Entrez search tool<sup>68</sup> to find GEO studies that had specific properties, such as metadata keywords, file formats and species. Then, using GEOparse<sup>69</sup>, the GEO text metadata were downloaded for each sample and searched for blacklisted words in the metadata or download URLs (for example, smartseq, trizol and fasta) to further filter out samples that were not generated using 10x Chromium. Data for samples and corresponding download links that passed the metadata filter stage were automatically downloaded. No datasets were realigned. In total, 753 studies were identified for download. A set of import functions was designed for the most common file type formats (.mtx, .h5ad and gene expression matrices in .tsv or .csv). Any dataset that could not be successfully downloaded or read in was discarded. Once read in, each sample was automatically tested for count data and gene names that match a reference gene list or gene name mapper before saving each file in a uniform .h5ad format for later processing. This resulted in a total of 334 published studies that were not duplicates of studies found in CELLxGENE<sup>19</sup> for use in our analysis. In the process of curating our reference data, we found that individual samples have been reposted across studies without references or a record of data provenance. We therefore advocate for improved data management practices in the field.

### Data preprocessing

All UMI count data were natural-log normalized per cell with a scaling factor of 10,000 using the scanpy.pp.normalize\_to\_target(adata, 10000) and scanpy.pp.log1p(adata) functions from scanpy<sup>70</sup>.

## Data aggregation and filtering

Datasets with author-provided cell type annotations used for training were obtained from Tabula Sapiens<sup>71</sup>, 10x Genomics<sup>20</sup>, the Single-Nucleus Cross-Tissue Atlas<sup>8</sup> and the Human Lung Cell Atlas<sup>72</sup> and subjected to the same preprocessing procedures as programmatically downloaded datasets. Cell type annotations were manually converted into terms contained within the Cell Ontology. Cells with annotations that did not clearly map to the Cell Ontology were not included in training.

Cell profiles previously annotated as doublets, scored as doublets by `infer_doublets` from Pegasus<sup>73</sup>, had >20% total UMI counts aligned to mitochondrial genes or had <500 total genes detected were removed.

## Preparation of training and test data

Training and test sets were chosen such that entire studies were held out of training (rather than holding out a subset of cells from each dataset) (Supplementary Table 1); there were 56 and 15 datasets in the training and test sets, respectively. This presents a harder generalization challenge and reflects how users are likely to use SCimilarity. Test datasets were selected to reflect the tissue diversity within the training sets.

## Cell Ontology term selection

Cell Ontology terms were selected for training if they were observed in at least two separate studies in the training set. Terms that appeared in only one study were not used because SCimilarity is trained by comparing cells across studies. To rescue single-study terms, the immediate parent terms were inspected across studies. If a single-study term's parent was observed in at least two other datasets, then the original cell type annotation was replaced with the coarser parent term (Supplementary Table 1) and used for ontology-aware triplet sampling. Otherwise, all cells with this annotation were removed from training. This process resulted in 203 Cell Ontology terms used in training (Supplementary Table 1). All 203 terms are available to the user for cell type prediction on new datasets using the SCimilarity software, with the default terms used for cell type prediction set to 81 manually curated terms with similar granularity, for convenience (Supplementary Table 1). As the size or annotation quality of training data grows, the number of Cell Ontology terms meeting the inclusion criteria are expected to increase.

## Semi-hard triplet mining

During training, batches of 1,000 cells are sampled from the training datasets. This sampling is weighted by study and cell type to have a similar number of observations per cell type from each study per batch.

Owing to the maximum operation within the loss function, not all viable triplets contribute to the gradient, and are categorized as easy, semi-hard or hard, based on their contribution to the gradient.

Easy negatives are defined by equation (5):

$$\|f(\mathbf{x}_i^a) - f(\mathbf{x}_i^p)\|_2^2 < \|f(\mathbf{x}_i^a) - f(\mathbf{x}_i^n)\|_2^2 + \alpha \quad (5)$$

Easy negatives provide no information to the gradient because the distances between the cells in the low-dimensional embedding already satisfy the objective, such that the maximum operation returns 0 to the triplet loss sum. As there are many easy triplets after training a small number of batches, randomly sampling triplets does not train models effectively. To accelerate training, triplets are mined to search for training triplets that are especially informative for model training<sup>12</sup>.

Hard negatives are defined by equation (6):

$$\|f(\mathbf{x}_i^a) - f(\mathbf{x}_i^p)\|_2^2 > \|f(\mathbf{x}_i^a) - f(\mathbf{x}_i^n)\|_2^2 + \alpha \quad (6)$$

Hard negatives contribute the largest quantity to the loss function, because they do not fit and are far from fitting the desired latent relationships. In practice, hard triplets are rarely useful for training, because

they contribute to model collapse during training<sup>12,74</sup>. Hard negatives may be enriched for incorrectly annotated cells.

Semi-hard negatives are defined by equation (7):

$$\|f(\mathbf{x}_i^a) - f(\mathbf{x}_i^n)\|_2^2 - \|f(\mathbf{x}_i^a) - f(\mathbf{x}_i^p)\|_2^2 < \alpha \quad (7)$$

Semi-hard negatives contribute small amounts to the loss function because they nearly satisfy the desired distances between cells in low-dimensional space. Meaning, the negative cell profile is further from the anchor cell than the positive cell, but by a less than the margin,  $\alpha$ . Semihard negatives are often used in triplet-loss models<sup>12</sup>.

Overall, we chose to train SCimilarity using only semi-hard negative triplets.

## Explainability framework

An explainability framework was used to identify genes of which the variation leads to the most significant variations of the learned features and, in turn, affects the relative distance between different cells.

An explanation for a pair of cells is defined as those genes that have the greatest impact on the relative distance between those cells in latent space. Given  $d(\mathbf{x}, \mathbf{y}) = \|f(\mathbf{x}) - f(\mathbf{y})\|_2^2$ , the distance between two cell profiles  $\mathbf{x}$  and  $\mathbf{y}$  in latent space  $f$ , the integrated gradient approach<sup>34</sup> was extended to compute the importance of each gene  $i$  in the comparison between cell profiles  $\mathbf{x}$  and  $\mathbf{y}$  as defined in equation (8):

$$\text{Importance}_i(\mathbf{x}, \mathbf{y}) = \left| \max((\mathbf{x}_i - \mathbf{y}_i), 0) \times \int_{a=0}^1 \frac{\partial d(\mathbf{y} + a \times (\mathbf{x} - \mathbf{y}), \mathbf{y})}{\partial x_i} \right| \quad (8)$$

Here  $a$  controls an interpolation process used to average gradients along a path. High values of  $\text{Importance}_i(\mathbf{x}, \mathbf{y})$  correspond to genes that are highly expressed in  $\mathbf{x}$ , and their modification (that is, gradient) affects  $d(\mathbf{x}, \mathbf{y})$  more. Intuitively, the expression of each gene in  $\mathbf{y}$  is gradually increased to match  $\mathbf{x}$  along the trajectory from  $\mathbf{x}$  to  $\mathbf{y}$ . Through this trajectory, the rate of change of  $d(\mathbf{x}, \mathbf{y})$  is computed for each gene, aggregating the results. To compute features relevant across broader contexts, the score is scaled by  $(\mathbf{x}_i - \mathbf{y}_i)$ , to achieve global explainability<sup>75</sup>. To identify genes that are upregulated in a subset of interest, genes  $i$  with expression  $\mathbf{x}_i < \mathbf{y}_i$  are ignored.

This approach differs in several key ways from the standard integrated gradient approach, because: (1) gradients are computed with respect to a learned distance instead of output features; (2) attributions where  $\mathbf{x}_i < \mathbf{y}_i$  are ignored; and (3) the sign of the integral is ignored due to the complex interactions between features.

To identify important genes for a cell type  $t$ , a set of cells  $T \in \{t_1, \dots, t_N\}$  with cell type  $t$  and a set of cells  $B \in \{b_1, \dots, b_N\}$  with cell types different from  $t$  are randomly sampled. Pairwise importances are computed for each pair of cells  $t_i$  in  $T$  and  $b_j$  in  $B$  and aggregated to obtain a signature that characterizes cell type  $t$  as defined in equation (9):

$$\text{Signature}_i(t) = \frac{1}{N} \sum_{c=1}^N \text{Importance}_i(t_c, b_c) \quad (9)$$

As the pairwise comparisons are averaging relative comparisons, the sampling of  $\{b_1, \dots, b_N\}$  impacts the signature scoring. To obtain general cell type markers, a background of all cell types is sampled. To obtain a cell-state-specific signature, a background of cells in other states of the same type are sampled. Confidence intervals for each gene  $i$  are computed as the standard error of the mean. This results in an attribution score for each gene.

## Attribution enrichment testing

Gene attributions were calculated using a set of foreground cells and a set of background cells. Foreground cells were the 500 cells most



similar to the query (FM) among the searched cells (for example, high confidence in vivo monocytes and macrophages). Background cells were selected by randomly sampling 500 cells outside of the top 10,000 cells (within in vivo monocytes and macrophages) by SCimilarity score to the query cell (FM). The AUC enrichment statistic was calculated based on the 3,000 genes with the highest attributions.

For each published signature, the AUC and one-sided  $P$  value were calculated using Mann–Whitney  $U$ -tests according to equation (10):

$$AUC = \frac{U}{n_1 n_2} \quad (10)$$

where  $U$  is the Mann–Whitney  $U$  statistic,  $n_1$  is the number of genes in the published signature among the 3,000 genes and  $n_2$  is the number of genes not in the published signature among the 3,000 genes.

### Training and evaluation metrics

**SCimilarity score.** The SCimilarity score is defined as the inverse of the cosine distance of two embedded cell profiles as in equation (11):

$$SCimilarity\ score = \frac{1}{1 - c_i \times c_j} \quad (11)$$

where  $c_i$  and  $c_j$  are the embeddings of the  $i$ th and  $j$ th cell profiles with unit length, respectively and  $i \neq j$ . The threshold for similarity varies in practice by question and cell types.

**Ontology-aware ASW.** ASW has been used to assess the performance of data integration tasks on multiple scRNA-seq studies<sup>16</sup> by quantifying how coherently each set of cells is grouped across studies after integration. For the batch ASW metric, the sets of cells are grouped by within study batches, so it is quantifying how coherently each batch is clustering (a lower score here is desired as it means greater mixing). For cell type ASW, where sets are defined by cell type, we introduce an ontology-aware modification. Here a higher score is desirable as it means that cells of the same type are more coherently clustered. The silhouette width of cell profile  $i$  of cell type  $t$  typically compares the average intracell type distances  $a(i)$  and the average inter-cell type distances  $b(i)$  between cells of type  $t$  and cells of the nearest cell type, defined by equations (12) and (13), respectively:

$$a(i) = \frac{1}{|C_t| - 1} \sum_{j \in C_t, i \neq j} d(i, j) \quad (12)$$

$$b(i) = \min_{j \neq t} \frac{1}{|C_j| - 1} \sum_{j \in C_j} d(i, j) \quad (13)$$

where, typically,  $C_t$  is the set of cells of author-annotated type  $t$  and  $C_j$  are the cells of all other cell types.

However, the ASW as typically formulated does not account for differences in granularity of cell type annotations across studies. To address those, a modified formulation is used where  $C_t$  contains cell type label  $t$  and all of its ontological descendants and  $C_j$  is the set of all other cell types, except cells of type  $t$  and any of its ontological descendants or ancestors. For example, if computing  $a(i)$  for a T cell, the distances between all types of T cell terms (CD4-positive,  $\alpha\beta$  T cell, CD8-positive,  $\alpha\beta$  T cell and CD4-positive, CD25-positive,  $\alpha\beta$  regulatory T cell and so on) are members of the T cell term. Ancestor terms of T cells, such as the term ‘lymphocytes’, are not members of the T cell class (nor a T cell subset) but are excluded from the summation indices in the calculations of  $a(i)$  and  $b(i)$ .

**Correlation with gene signatures.** To test how the SCimilarity distance represents distance between predefined cell states, a signature-based definition of cell state was correlated with the SCimilarity score (above).

For each cell in the test set, both the signature score<sup>76</sup> and a SCimilarity score versus the cell query are calculated, yielding two vectors, and the Pearson’s correlation coefficient is calculated between the vectors.

**Model selection.** Models were run in triplicate with 18 combinations of 3 different margins ( $\alpha \in \{0, 0.01, 0.05, 0.1\}$ ) and 6 different  $\beta$  parameters ( $\beta \in \{0, 0.0001, 0.001, 0.01, 0.1, 1.0\}$ ) and one query model and one integration model were selected based on two criteria. First, query performance was tested by how well cell similarities to a query FM profile correlated with a signature defining that same state (*TREM2*, *GPNNB*, *SPP1*, *CCL18*, *MMP9*, *CTSK*, *APOE*, *CHIT1*, *LIPA*, *CHI3L1*, *CD14*, *APOC1*). Second, ontology-aware ASW was used to quantify how well the cells of the same type from different studies intermixed in SCimilarity’s representation. The model with the highest summed query and integration score was selected as it performed much better on the query task than the other high integration models (Extended Data Fig. 2b,c). This selected integration model had more study mixing than the query model according to the study (NMI) and study ARI<sup>16</sup>.

**Data integration benchmarking.** SCimilarity was compared to four batch integration methods: Harmony<sup>23</sup> (harmonypy v.0.0.9), Scanorama<sup>24</sup> (v.1.7.4), scVI<sup>10</sup> (v.1.1.0rc2) and scArches<sup>11</sup> (scVI v.1.1.0rc2). The modified ASW (above), ARI and NMI were calculated as integration benchmark metrics. As ‘ground truth’ cell type annotations are required to assess preservation of biological signal, methods were benchmarked on the 15 test studies with author-provided cell type annotations held out during SCimilarity training.

Harmony and Scanorama were run using the wrapper functions in scanpy<sup>70</sup>. scVI and scArches were run using the scvi-tools workflow described in their online tutorials (<https://docs.scvi-tools.org>). As the scArches workflow requires a reference dataset, 101,133 cell profiles were sampled across all training datasets with uniform probability across studies for use as the reference.

Study ARI, study NMI and cell type ASW were calculated on four distinct integration tasks based on five different combination of validation datasets, four positive control tasks: (1) 143,638 cell profiles sampled from all 15 test datasets with uniform probability across studies; (2) two lung datasets<sup>1,29</sup>; (3) two kidney datasets<sup>25,26</sup>; and (4) two PBMC datasets<sup>27,28</sup>, all selected from the test studies, and one negative control task of integrating B cells from one PBMC dataset<sup>28</sup> with regulatory T cells from a different PBMC dataset<sup>27</sup>.

### Cell type annotation

Cell type assignments were performed by  $k$ -NN classification combined with an annotated reference set. SCimilarity’s reduced dimensionality latent space was used to determine  $k = 50$  nearest neighbours in the reference dataset to a query cell  $t$ , and the query cell was annotated either by tallying votes based each cell’s annotation with equal weights according to equation (14),

$$\text{Cell type}(t) = \arg\max_t \left( \sum_{i \in t} \frac{1}{n} \right) \quad (14)$$

or with weights by distance in SCimilarity’s reduced dimensionality latent space according to equation (15):

$$\text{Cell type}(t) = \arg\max_t \left( \sum_{y \in t} \frac{1}{d(\mathbf{x}, \mathbf{y})} \right) \quad (15)$$

To allow users to annotate new datasets from a restricted list of cell types of interest, specific cell types can be excluded (blocklisting) or annotations may be limited to specific cell types (safelisting). When feasible, blocklisting or safelisting is recommended to improve interpretability and reduce spurious annotations. However, extensive blocklisting or safelisting can slow the annotation process substantially,

# Article

because the pre-built  $k$ -NN indices are not optimized for a modified target cell type list.

## **$k$ -NN parameters for annotation and query**

Two separate  $k$ -NN indices were used for efficient and accurate queries. For cell type annotation, a 7.9-million-cell  $k$ -NN index was built using `hnsplib`<sup>32</sup> with `ef_construction = 1,000` and  $M = 80$ . Searching this  $k$ -NN found the 50 nearest neighbours (default behaviour) for cell type annotation ( $k = 50$ ) and `ef = 100`.

Cell query relied on a separate 23.4-million-cell  $k$ -NN index also built using `hnsplib`. This index was constructed with the following parameters: `ef_construction = 400` and  $M = 50$ . The search parameters are set by the user's request for how many similar cells to return. The default behaviour is set to  $k = 1,000$  and `ef = k` but, in practice,  $k$  can vary widely depending on the use case.

## **Cross-technology benchmarking**

Comparison of scRNA-seq and snRNA-seq SCimilarity embeddings was performed using the paired data for sample CLL1 from GEO GSE140819 (ref. 30). SCimilarity cell type annotation was constrained to 7 Cell Ontology terms that were most similar, but more granular than, the three author-provided annotations (B cell, T cell and macrophage). Pairwise distance distributions were calculated for up to 1,000 randomly sampled cell pairs (limited by cell numbers), without replacement, for the most abundant SCimilarity annotated cell types. Distributions were generated for pairs of selected populations within annotation and protocol (cell to cell or nucleus to nucleus), within annotation and across protocols (cell to nucleus) and across annotation (one cell type to another cell type) and within protocol.

Profiling platforms were compared using the data for the human PBMC sample from SCP424<sup>31</sup>. The distribution of nearest-neighbour SCimilarity scores was retrieved from the  $k$ -NN graph both irrespective of platform and constrained to within-platform and within-replicate neighbours. Cell type annotations were constrained to nine Cell Ontology terms most similar to the author-provided annotations. Annotation precision was calculated as the percent of cells with SCimilarity-predicted annotations identical to the Cell Ontology mapped author-provided annotations within each platform and replicate separately.

## **Cell type annotation benchmarking**

SCimilarity's cell type annotation was compared to three cell type prediction methods (CellTypist<sup>77</sup> v.1.6.2, TOSICA<sup>78</sup> v.1.0.0, and scANVI<sup>79</sup> from scVI v.1.1.0rc2) with three separate classification tasks: (1) annotating cells in a human kidney dataset<sup>25</sup>; (2) annotating cells in a human PBMC CITE-seq dataset<sup>33</sup>; and (3) annotating cell types across all 15 holdout datasets that had author-provided annotations. The same SCimilarity model was used for both evaluations. A separate model was trained for each task by each of the other three methods.  $F_1$  scores were calculated for each cell type in each test study.

For the ref. 25 kidney test dataset (12,190 cell profiles), cell type annotations were flattened to 22 Cell Ontology terms manually. CellTypist, TOSICA and scANVI models were trained using 89,520 cells obtained from four kidney SCimilarity training datasets that were annotated with cell type terms in the ref. 19 test dataset. For the CZI PBMC CITE-seq dataset of ref. 33 (94,811 cell profiles), four ambiguously defined cell populations were removed from the analysis (for example, exhausted B cells, immature B cells, proliferating T cells and proliferating NK cells) and cell type annotations were constrained to 22 Cell Ontology terms identical to the author provided annotations. scANVI was trained using the scvi-tools workflow (<https://docs.scvi-tools.org>). CellTypist was trained using the workflow for custom models (<https://colab.research.google.com/github/Teichlab/celltypist>). TOSICA was trained using the demo tutorial (<https://github.com/JackieHanLab/TOSICA>). Performance was assessed by  $F_1$  score for all cell type prediction methods.

For benchmarking across all 15 test datasets (Fig. 3e), 143,638 cell profiles were sampled with uniform probability across the 15 studies. These were then filtered to cell types found within the test set annotations. New CellTypist, TOSICA and scANVI models were learned with the remaining 103,116 training cell profiles sampled across all training datasets, weighted so that each study was equally represented in the complete training set.

## **Outlier filtering**

To filter outlier cells before visualization and downstream analysis, SCimilarity's score is used to flag cells that are out of distribution. Cells with a SCimilarity score  $< 33$  from the nearest cell in the training set were removed before further analysis. Many of these cells were from immortalized cell lines, and reflect their difference from primary cells (and absence in the training). Note that if out-of-distribution cells are not removed, these cells will not be accurately annotated and can confound visualization.

## **Macrophage query preprocessing**

To prepare a cell query for FM cells, a public dataset<sup>1</sup> (GSE136831 and <https://www.ipfcellatlas.com>) was preprocessed with the same steps for all ingested data and scored using Scanpy's `scanpy.tl.score_genes` function with a gene signature of *SPPI*, *TREM2*, *GPNMB*, *MMP9*, *CHIT1* and *CHI3L1* in Scanpy<sup>70</sup>. The average profile of the top 50 scoring cell was embedded using SCimilarity and used as the input query to SCimilarity's cell search model and used throughout analyses in Figs. 4 and 5.

## **Foundation model benchmarking**

SCimilarity, scGPT<sup>5</sup> (v.0.2.1, 23 June 2023 model) and scFoundation<sup>80</sup> (9 December 2023 model) were compared on dataset GSE128033 using the FM and myofibroblast gene signatures and a cell query profile derived from GSE136831. The query cell profile was defined as the centroid of the top 100 scoring cells using scanpy gene signature in GSE136831. The query profile and all cells in GSE128033 were embedded according to the scGPT reference mapping tutorial (<https://github.com/bowang-lab/scGPT>) and the scFoundation `get_embedding.py` script (<https://github.com/biomap-research/scFoundation>) documentation. Embedding distances were calculated using the Euclidean distance between the embedded query profile and all cells in GSE128033. Spearman rank correlation coefficient values ( $\rho$ ) were calculated between the gene signature score and distances to the query cell state across all cells in each model. Cell type annotations predicted by SCimilarity were constrained to 28 Cell Ontology terms present in lung tissue.

## **Quality control for query input**

The results of cell queries depend on the centroids used for the query. To help users generate effective cell state queries, a statistic is calculated from the query cells (that is, cells in a grouping and their centroid). For robust and meaningful query results, a cell state should be a centroid of a coherent, relatively homogeneous set of cells. To evaluate centroid's quality, its underlying cells are subclustered ( $k = 10$  clusters), 10 centroids are computed from the subclustering and a SCimilarity search is conducted for the most similar cells to each of the 10 centroids (default  $n = 100$  nearest neighbours). The mean overlap in cell query results between the parent centroid profile and each  $k$ -means subcluster centroid is reported as a measure of query stability.

## **Myofibroblast and FM co-occurrence**

Co-occurrence of two cell states was calculated using the results of two independent queries. The relative frequency of each query (for example, FMs and fibrosis-associated myofibroblasts) in each sample was quantified by counting the number of searched cells in that sample that were highly similar ( $\geq 95$ th percentile of SCimilarity scores) to each query profile, divided by the number of searched cells in the sample. 'Searched cells' for FMs were any subtype of monocyte or macrophage

(classical monocyte, intermediate monocyte, non-classical monocyte, macrophage or alveolar macrophage) (Fig. 4c). ‘Searched cells’ for fibrosis-associated myofibroblasts were all cells annotated as fibroblasts or myofibroblasts (Extended Data Fig. 5a). Only in vivo tissue samples with at least 50 macrophages and 50 fibroblasts were considered.

### Important genes and pathways in FMs

Important genes were identified using SCimilarity’s attribution score method. This method requires two cell groups to compare, identifying which genes differ between them. Here we used 500 cells that were considered to be similar to the average FM profile calculated from a previous study<sup>2</sup> as the FM-like group. To compare to the FM-like group, 500 cells were randomly sampled from the full 2.5-million-cell monocyte and macrophage query set.

Reactome pathways enriched for the 100 genes with the top importance scores for FMs were determined using the method provided in the ReactomePA<sup>81</sup> R package, with multiple-hypothesis correction using the Benjamini–Hochberg method and the background gene universe restricted to the approximately 28,000 genes included in SCimilarity. Pathways were considered to be significant if they met the criteria of adjusted *P* value (*Q*)  $\leq 0.05$  and gene count  $\geq 4$ .

### 3DCS culture of PBMCs

Peripheral blood samples from healthy volunteers were provided by the Samples for Science (S4S) donor program at Genentech; donors provided written informed consent and sample collection was approved by the Western-Copernicus Group Institutional Review Board. The samples were collected in heparin collection tubes and subsequently diluted 1:1 with a solution of PBS containing 2% FBS and 1 mM EDTA. Then, 30 ml of diluted blood was overlaid onto 15 ml of Lymphoprep (StemCell Technologies) in a 50 ml tube and centrifuged at 3,000 rpm for 20 min at 4 °C. PBMCs were isolated from the interphase after centrifugation and diluted with PBS containing 2% FBS and 1 mM EDTA and centrifuged at 300g for 10 min at 4 °C. The cell pellet was washed again with PBS containing 2% FBS and 1 mM EDTA. Red blood cell lysis was performed on the cell pellet by resuspending in RBC lysis buffer (Cell Signalling Technology) for 5 min at room temperature, followed by inactivation with addition of RPMI medium containing 10% FBS. Cells were pelleted by centrifugation at 300g for 10 min at 4 °C and subsequently washed with PBS containing 2% FBS and 1 mM EDTA. Cells were then resuspended in a 10% sucrose solution at a concentration of  $2 \times 10^6$  cells per ml right before plating into 3D hydrogel culture. Puramatrix hydrogel (Corning) was vortexed for 30 s and diluted 1:1 with a 20% sucrose solution. Then, 250  $\mu$ l of diluted Puramatrix hydrogel was mixed with 250  $\mu$ l of resuspended PBMCs and plated in a 24-well tissue culture plate. To induce gelation, RPMI medium was overlaid onto the hydrogel/PBMC mixture and incubated for 5 min in a 37 °C incubator with 5% CO<sub>2</sub>. Overlaid medium was aspirated off the 3D hydrogel and washed twice with RPMI medium, after which 600  $\mu$ l of 3DCS medium, formulated as previously described<sup>2</sup>, was overlaid onto the hydrogel. Cells were cultured in a 37 °C incubator with 5% CO<sub>2</sub> for 8 days, with medium exchanges every other day. On day 8, culture cells were recovered from the 3D hydrogel for scRNA-seq.

### scRNA-seq of the 3D culture system

Wells of the 3D hydrogel culture were washed with PBS, followed by recovery of the hydrogel and cells by gentle pipetting in PBS buffer. This solution was centrifuged for 5 min at 750g, and the hydrogel/PBMC pellet was resuspended in TrypLE solution (Thermo Fisher Scientific) and incubated at 37 °C for 10 min. RPMI medium with 10% FBS was added and the solution was centrifuged for 5 min at 750g. The resultant pellet was washed twice with PBS to remove hydrogel matrix debris. PBMCs were resuspended in PBS and passed through a 40  $\mu$ m filter, pelleted by centrifugation at 300g for 5 min and resuspended in RPMI medium with

10% FBS. The cell solution was subjected to FACS to isolate cells from any remaining hydrogel debris and recovered cells were concentrated to 1,000 cells per  $\mu$ l in RPMI medium with 10% FBS for downstream profiling by scRNA-seq.

scRNA-seq was performed using the Chromium Single Cell 3’ Library and Gel bead kit v3 (10x Genomics), according to the manufacturer’s user guide. In brief, the cell density and viability of the single-cell suspension were determined using the Vi-CELL XR cell counter (Beckman Coulter). The cell density was used to impute the volume of single-cell suspension needed in the reverse transcription master mix, aiming to achieve around 10,000 cells per sample. cDNAs and libraries were prepared according to the manufacturer’s user guide (10x Genomics). Libraries were profiled using the Bioanalyzer High Sensitivity DNA kit (Agilent Technologies) and quantified using the Kapa Library Quantification Kit (Kapa Biosystems). Libraries were sequenced on the NovaSeq 6000 (Illumina) system according to the manufacturer’s specifications with 28 + 90 bp paired-end reads at a depth of 101 million mate-pair reads. Sequencing reads were aligned to the GENCODE 27 Basic gene model on the human genome assembly GRCh38 using Cell Ranger v.6.0 (10x Genomics).

Individual samples were genetically demultiplexing using the singularity container provided with Souporecell (v.2.0)<sup>82</sup>. No genotype information was provided to the pipeline. As PBMCs were provided from three donors, a *k* value of 3 was used to cluster the samples into three genotypes. These samples were preprocessed consistently with the previously ingested samples and then embedded using SCimilarity to enable direct comparisons to ref. 2 as well as the rest of the public datasets.

SCimilarity cell type classification was applied to both public and validation cells using SCimilarity with the following safelist: B cell, CD4-positive,  $\alpha\beta$  T cell, CD8-positive,  $\alpha\beta$  T cell, conventional dendritic cell, haematopoietic stem cell, macrophage, monocyte, natural killer cell, plasma cell, plasmacytoid DC.

### Code performance benchmarking

Benchmarks were run on servers with 8 Intel Xeon E5-2650 v4 CPUs with 2.20 GHz cores and a total of 128 GB of RAM.

Query runtimes, using the prebuilt approximate *k*-NN index<sup>32</sup> to find the top *n* most similar cells, had an average runtime of 50 ms. Some API functions use the query and summarize the metadata within one function call. That function timing is dominated by summarizing metadata and computing statistics from the query results, which requires an additional 3.3 s. This performance differs from an exhaustive comparison (Fig. 5b), where the query was directly compared against 2.58 million monocytes and macrophages with a runtime of 2 s.

Cell signatures were calculated using scanpy.tl.score\_genes. The scanpy score\_genes function was applied to the already normalized data. This runtime totalled 2 h, 46 min and 20 s when it was applied across each .h5ad file (one file per tissue sample). Even though .h5ad files were not stored with any compression, file reading was a dominant factor in runtime.

### Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

### Data availability

In vitro data generated in this study have been deposited in the GEO under accession number GSE280632. Model weights, single-cell data embeddings, curated metadata and *k*-NN graphs have been deposited at Zenodo<sup>83</sup> (<https://doi.org/10.5281/zenodo.10685499>). Source repositories and accession numbers for the public sc/snRNA-seq studies used for model training, model testing or as part of the unlabelled referenced set are provided in Supplementary Table 1.

## Code availability

Code and tutorials are available at GitHub (<https://github.com/Genentech/scimilarity>). A snapshot of the code that accompanies this publication is available at Zenodo<sup>84</sup> (<https://doi.org/10.5281/zenodo.14087552>). Code license: Apache 2.0. Pretrained model weights,  $k$ -NN and pre-built indices license: CC-BY-SA 4.0.

65. Baldi, P. & Sadowski, P. Understanding dropout. In *Advances in Neural Information Processing Systems* Vol. 26 (eds Burges, C. J. C. et al.) 2814–2822 (Curran Associates, 2013).
66. Ding, J. & Regev, A. Deep generative model embedding of single-cell RNA-Seq profiles on hyperspheres and hyperbolic spaces. *Nat. Commun.* **12**, 2554 (2021).
67. Le, L., Patterson, A. & White, M. Supervised autoencoders: improving generalization performance with unsupervised regularizers. In *Advances in Neural Information Processing Systems* Vol. 31 (eds Bengio, S. et al.) 107–117 (Curran Associates, 2018).
68. Cock, P. J. A. et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**, 1422–1423 (2009).
69. Gumienny, R. GEOparse: Python library to access Gene Expression Omnibus Database (GEO). *GitHub* <https://github.com/guma44/GEOparse> (2021).
70. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.* **19**, 15 (2018).
71. Tabula Sapiens Consortium. The Tabula Sapiens: a multiple-organ, single-cell transcriptomic atlas of humans. *Science* **376**, eabl4896 (2022).
72. Travaglini, K. J. et al. A molecular cell atlas of the human lung from single-cell RNA sequencing. *Nature* **587**, 619–625 (2020).
73. Li, B. et al. Cumulus provides cloud-based data analysis for large-scale single-cell and single-nucleus RNA-seq. *Nat. Methods* **17**, 793–798 (2020).
74. Wu, C.-Y., Manmatha, R., Smola, A. J. & Krähenbühl, P. Sampling matters in deep embedding learning. In *2017 IEEE International Conference on Computer Vision (ICCV)* 2840–2848 (IEEE, 2017).
75. Ancona, M., Ceolini, E., Öztireli, C. & Gross, M. Towards better understanding of gradient-based attribution methods for deep neural networks. Preprint at <https://doi.org/10.48550/arXiv.1711.06104> (2018).
76. Satija, R., Farrell, J. A., Gennert, D., Schier, A. F. & Regev, A. Spatial reconstruction of single-cell gene expression data. *Nat. Biotechnol.* **33**, 495–502 (2015).
77. Domínguez Conde, C. et al. Cross-tissue immune cell analysis reveals tissue-specific features in humans. *Science* **376**, eabl5197 (2022).
78. Chen, J. et al. Transformer for one stop interpretable cell type annotation. *Nat. Commun.* **14**, 223 (2023).
79. Xu, C. et al. Probabilistic harmonization and annotation of single-cell transcriptomics data with deep generative models. *Mol. Syst. Biol.* **17**, e9620 (2021).
80. Hao, M. et al. Large-scale foundation model on single-cell transcriptomics. *Nat. Methods* **21**, 1481–1491 (2024).
81. Yu, G. & He, Q.-Y. ReactomePA: an R/Bioconductor package for reactome pathway analysis and visualization. *Mol. Biosyst.* **12**, 477–479 (2016).
82. Heaton, H. et al. Souporecell: robust clustering of single-cell RNA-seq data by genotype without reference genotypes. *Nat. Methods* **17**, 615–620 (2020).
83. Heimberg, G. et al. Data for ‘A cell atlas foundation model for scalable search of similar human cells’. *Zenodo* <https://doi.org/10.5281/zenodo.10685499> (2024).
84. Heimberg, G. et al. Code for ‘A cell atlas foundation model for scalable search of similar human cells’. *Zenodo* <https://doi.org/10.5281/zenodo.14087552> (2024).

**Acknowledgements** We thank A. Tripathi for coming up with the name ‘SCimilarity’ and J. Collier, G. Eraslan, J. Marioni and J. Freimer for their suggestions on the manuscript.

**Author contributions** G.H. conceived the method with input from A.R., J.A.V.H., H.C.B. and J.K.; G.H. and T.K. performed data ingest and model implementation with input from J.A.V.H., N.D., G.S., T.B., J.K. and A.R. Python API was developed by T.K. with help from J.A.V.H., O.S. and G.H. Interpretability was developed by N.D. and G.S. with input from H.C.B. and T.B.; J.A.V.H. conceived the biological application of the method with input from G.H., S.J.T., J.R.R. and D.J.D.; D.J.D. and T.H. performed experimental validation with guidance from J.R.R. and S.J.T.; G.H. wrote the manuscript with input from J.A.V.H., J.K., A.R. and H.C.B. All of the authors reviewed the manuscript.

**Competing interests** All of the authors are employees of Genentech or Roche. A.R. is a co-founder and equity holder of Celsius Therapeutics, an equity holder in Immunitas and, until 31 July 2020, was a scientific advisory board member of Thermo Fisher Scientific, Syros Pharmaceuticals, Neogene Therapeutics and Asimov. G.H., D.J.D., O.S., N.D., G.S., T.B., S.J.T., J.R.R., H.C.B., J.K., J.A.V.H. and A.R. have equity in Roche.

### Additional information

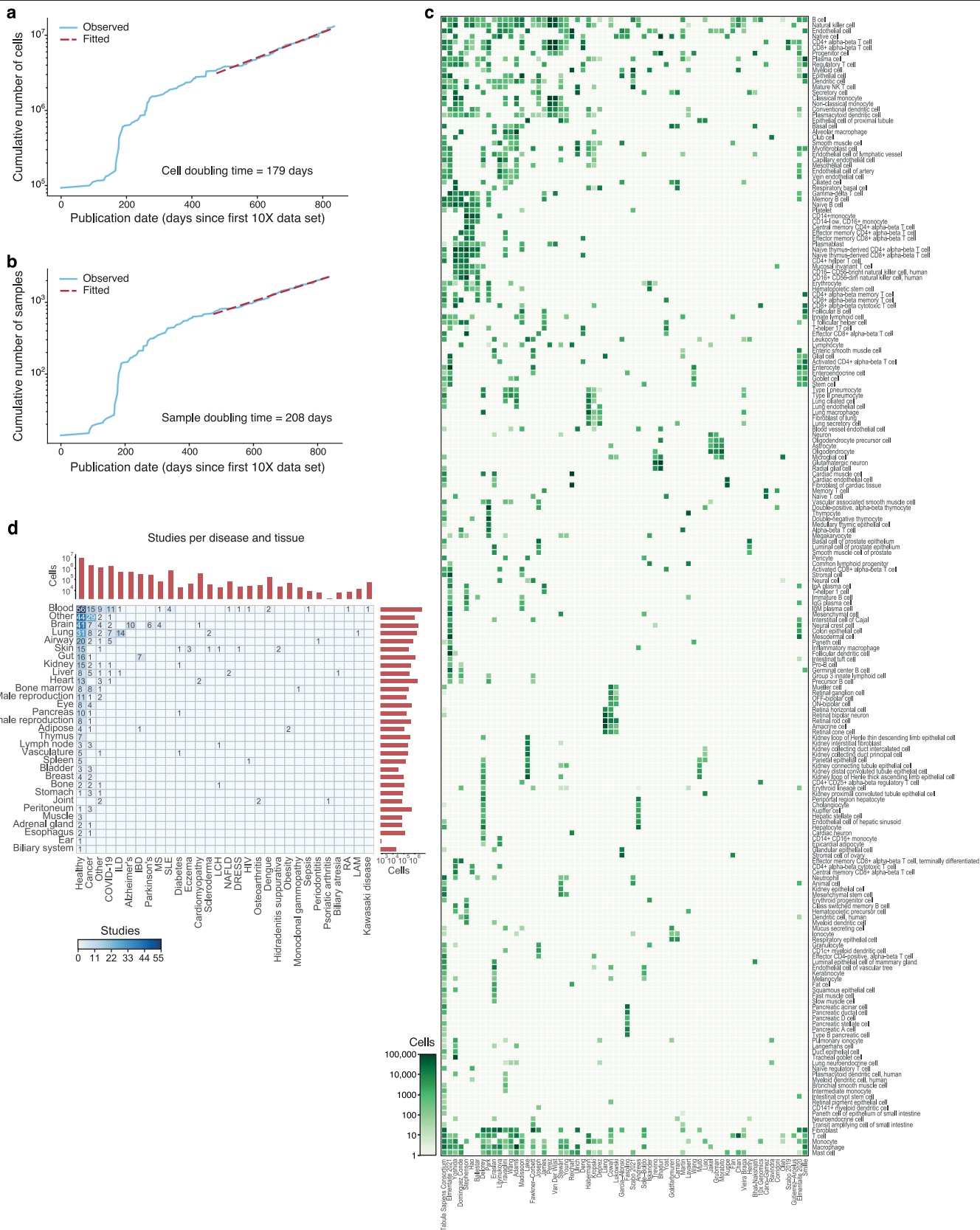
**Supplementary information** The online version contains supplementary material available at <https://doi.org/10.1038/s41586-024-08411-y>.

**Correspondence and requests for materials** should be addressed to Graham Heimberg, Josh Kaminker, Jason A. Vander Heiden or Aviv Regev.

**Peer review information** *Nature* thanks the anonymous reviewers for their contribution to the peer review of this work.

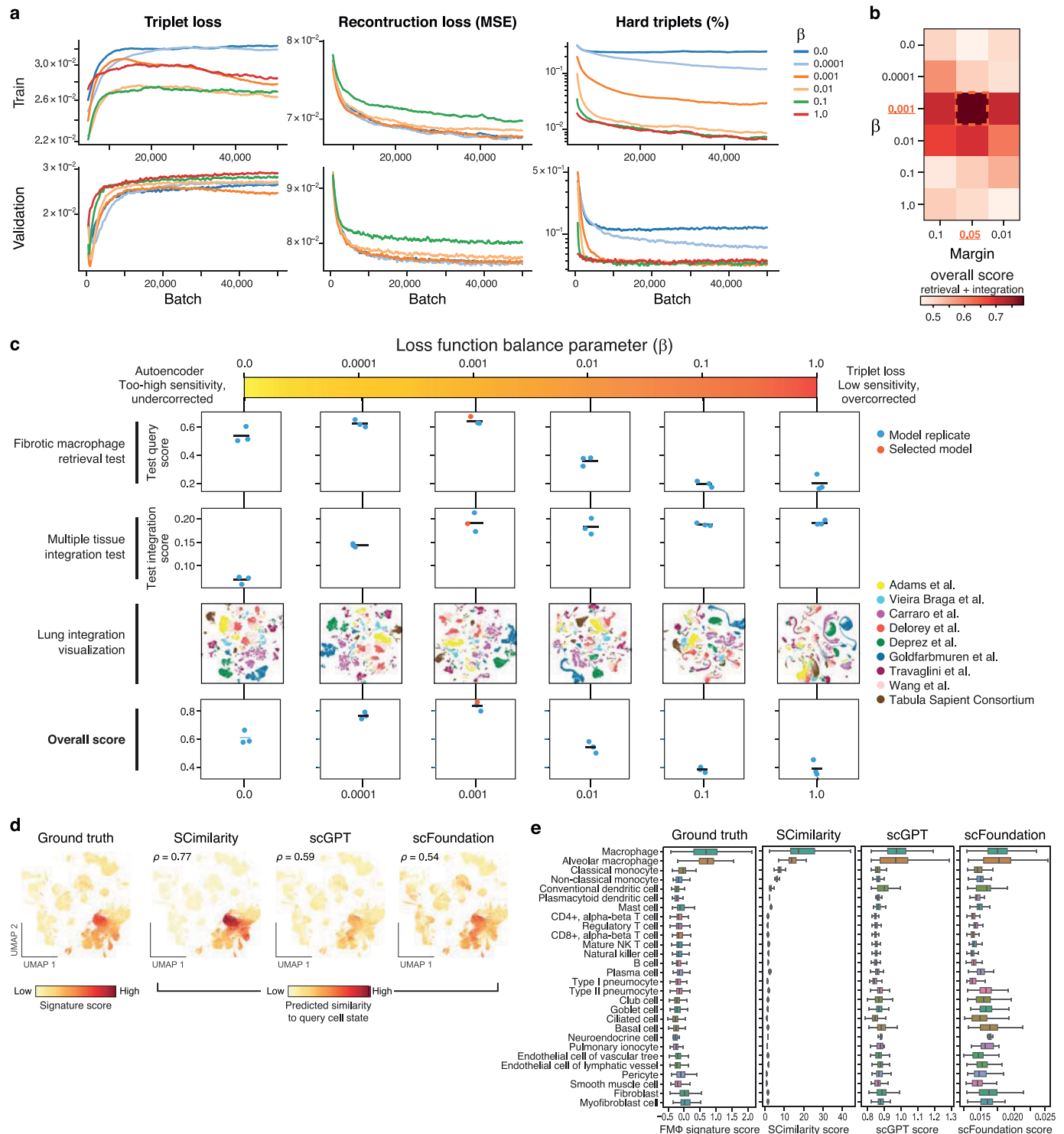
**Reprints and permissions information** is available at <http://www.nature.com/reprints>.





**Extended Data Fig. 1 | Data compendium to assemble a pan-human reference. a, b.** Cumulative number of (a) cells (y axis) and (b) samples (y axis) profiled by sc/snRNA-seq (and matching our filters; **Methods**) over time (x axis). Doubling time is calculated based on the publication date from the most recent 150 data points (dashed red line). **c.** Author-annotated cell types used in training.

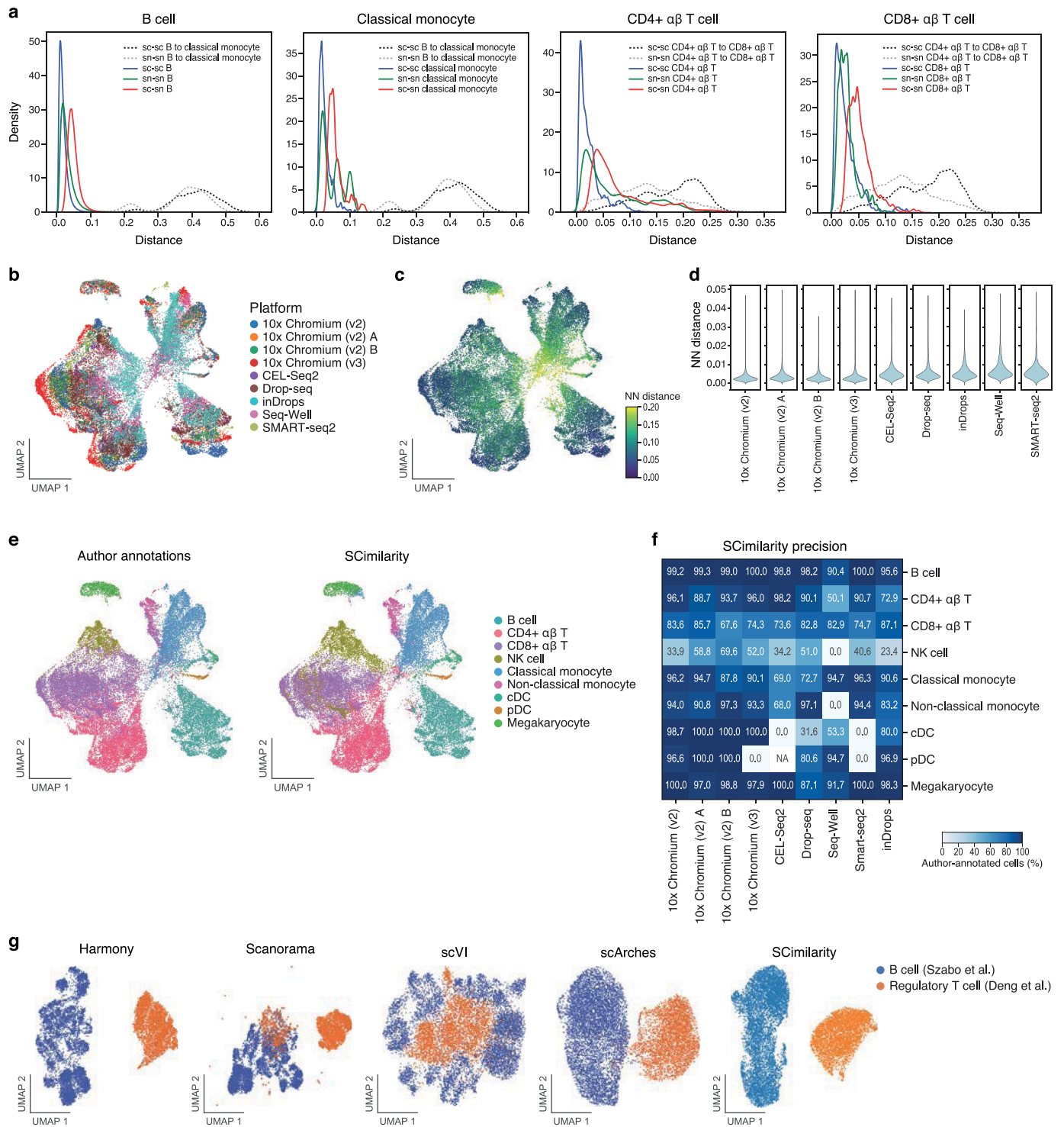
Number of author-annotated cells (colour bar) from each Cell Ontology type (rows) and study (columns) used for SCimilarity model training. **d.** Tissues and diseases used in training. Number of studies (heatmap tiles, text and colour bar) and cells (margins, y or x axis) used for model training from each tissue (rows, right y axis) and disease (columns, top x axis).



### Extended Data Fig. 2 | SCimilarity training and hyperparameters.

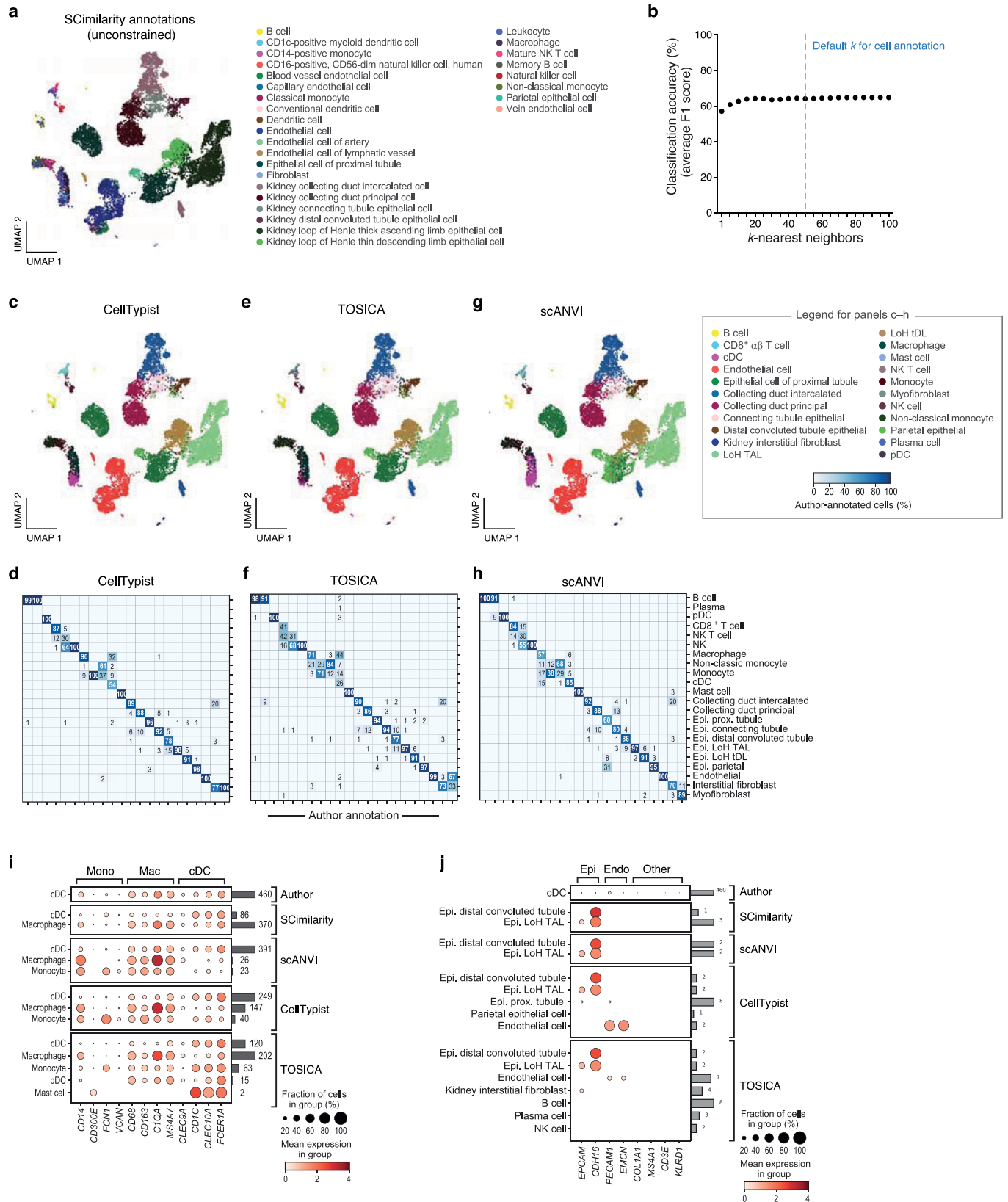
**a**, Training and validation curves. Triplet loss (y axis, left), reconstruction loss (mean squared error (MSE), y axis, middle), and percent of hard triplets (y axis, right) across training (top) or validation (bottom) batches (x axis), for SCimilarity models with margin=0.05 across six  $\beta$  values (colour). Reconstruction loss for pure triplet loss ( $\beta = 1$ ) not shown. **b**, Impact of hyperparameter selection on model performance. Overall model score (colour) across margins (columns) and loss weightings ( $\beta$ , rows), ( $\beta = 0$ : pure reconstruction loss;  $\beta = 1$ : pure triplet loss). Model score is the sum of query score for FMΦ retrieval (correlation between signature and SCimilarity score of retrieved FMΦs) and ontology-aware average silhouette width of integration (higher score reflects more coherent clusters by cell type). **c**, Test metrics for SCimilarity models across  $\beta$  values. FMΦ retrieval (first row), ontology-aware average silhouette width of integration (second

row), UMAP embedding of cells from nine lung datasets coloured by study (third row), and sum of retrieval and integration scores (y axis, fourth row) for models trained with increasing  $\beta$  (leftmost: traditional autoencoder; rightmost: triplet loss only) across  $n = 3$  model replicates for each  $\beta$ . **d**, **e**, SCimilarity better captures an FMΦ query. **d**, UMAP of cells from the ILD study GSE128033 with cells coloured by FMΦ signature score (ground truth) or similarity to the FMΦ query for SCimilarity (right, first), scGPT (right, second), or scFoundation (right, third). Top left: Spearman's  $\rho$  between signature score rankings and distances to the query cell. **e**, Distribution of FMΦ signature (first), SCimilarity (second), scGPT (third), and scFoundation (fourth) scores as in (d) for  $n = 28$  SCimilarity predicted cell types across  $n = 58,530$  cells (outliers removed). Boxplot: upper/lower quartiles (box), min/max values (whiskers), and median (center line).



**Extended Data Fig. 3 | SCimilarity integrates and annotates across profiling methods.** **a**, SCimilarity integrates snRNA-seq and scRNA-seq. Distribution of pairwise SCimilarity embedding distances for randomly sampled cell (sc-sc), nucleus (sn-sn) or cell-nucleus (sc-sn) profile pairs (max  $n = 1000$ , without replacement) within SCimilarity-annotated B cells (first), classical monocytes (second), CD4+ T cells (third), or CD8+ T cells (fourth) from patient tumour CLL1 in Slyper et al., 2020<sup>30</sup>; overlaid with similarly sampled cell or nucleus pairwise embedding distances between B cells and classical monocytes (first, second) or CD4+ T cells and CD8+ T cells (third, fourth). **b-f**, SCimilarity generalizes well to scRNA-seq test data collected by seven different methods. UMAP embedding

of PBMC profiles from one sample profiled by seven different scRNA-seq methods<sup>31</sup> coloured by platform and replicate (**b**) and nearest-neighbour distance in SCimilarity's latent space (**b**); **d**, Distribution of nearest-neighbour distances (y axis, range limited to  $\leq 0.05$ ) for each platform and replicate (x axis). **e**, UMAP embedding as in **b**, coloured by author (left) or SCimilarity (right) annotations. **f**, Percentage (colour bar) of author-annotated cells (rows) matching annotations predicted by SCimilarity for each platform and replicate (columns). **g**, Negative control benchmark of data integration. UMAP embedding of B cell profiles (from Szabo et al.<sup>28</sup>) and T<sub>reg</sub> profiles (from Deng et al.<sup>27</sup>), coloured by cell type after integration with each of five methods.

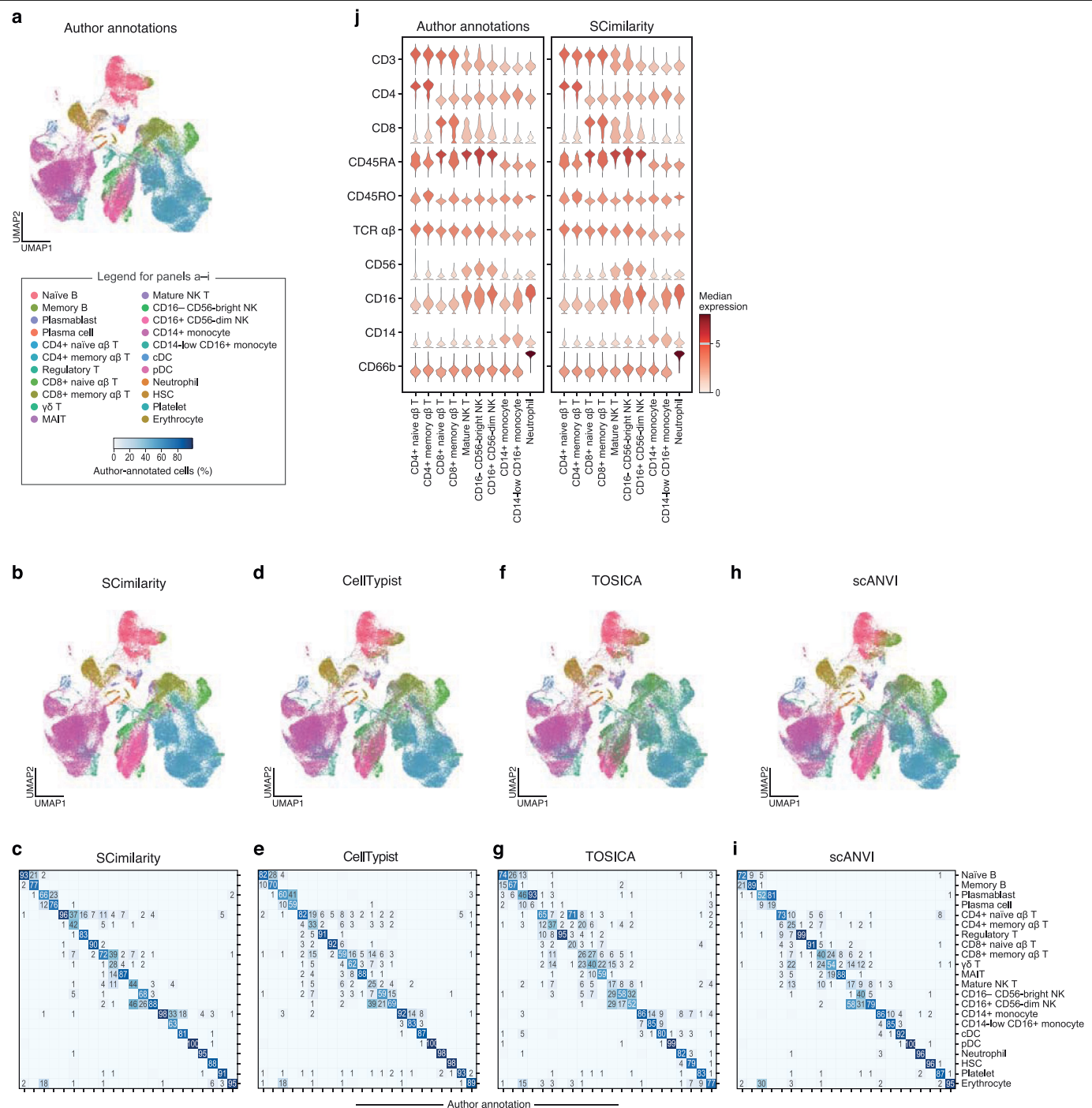


Extended Data Fig. 4 | See next page for caption.



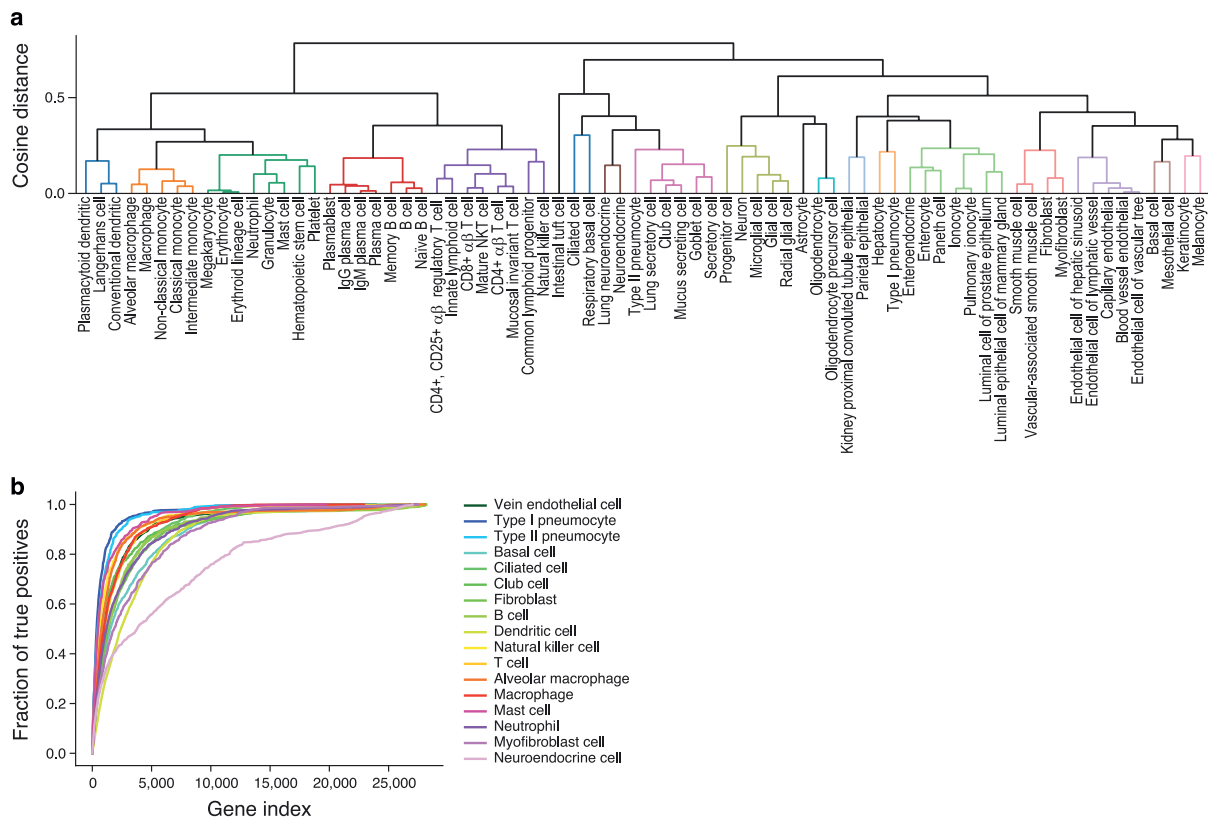
**Extended Data Fig. 4 | Validation of cell type annotation on tissue scRNA-seq.** **a**, SCimilarity unconstrained cell type annotation. UMAP embedding of single cell profiles (dots) from SCimilarity's latent representation of a test scRNA-Seq kidney data<sup>19</sup> (held out from training) (as in Fig. 3b,c), coloured by cell annotations obtained without constraining to the scope of author-provided annotations in the study. **b**, Annotation is robust to the number of nearest neighbours. Cell type classification score (y axis) at different number of nearest neighbours, *k* (x axis). **c-h**, Benchmarking of annotation by established methods. **c,e,g**, UMAP embedding of cell profiles as in **(a)** coloured by annotations predicted by CellTypist (**c**), TOSICA (**e**), or scANVI (**g**). **d,f,h**, Percentage

(colour bar) and number of author-annotated cells (columns) matching annotations predicted by CellTypist (**d**), TOSICA (**f**), and scANVI (**h**) (rows). **i,j**, Author annotated cDCs express a mixture of DC markers and markers of other cell types. Mean expression (dot colour) and percent of expressing cells (dot size) for canonical marker genes of monocytes (Mono), macrophages (Mac), and conventional dendritic cells (cDCs) (**i**) or epithelial (Epi), endothelial (Endo), or other non-myeloid lineages (Other) (**j**) in author-annotated cDCs (row 1) and the subset of those same cells predicted as different myeloid subsets (rows, i) or as non-myeloid cells (rows, j) by other annotation methods. Right bar plots and counts: number of cells per annotation.



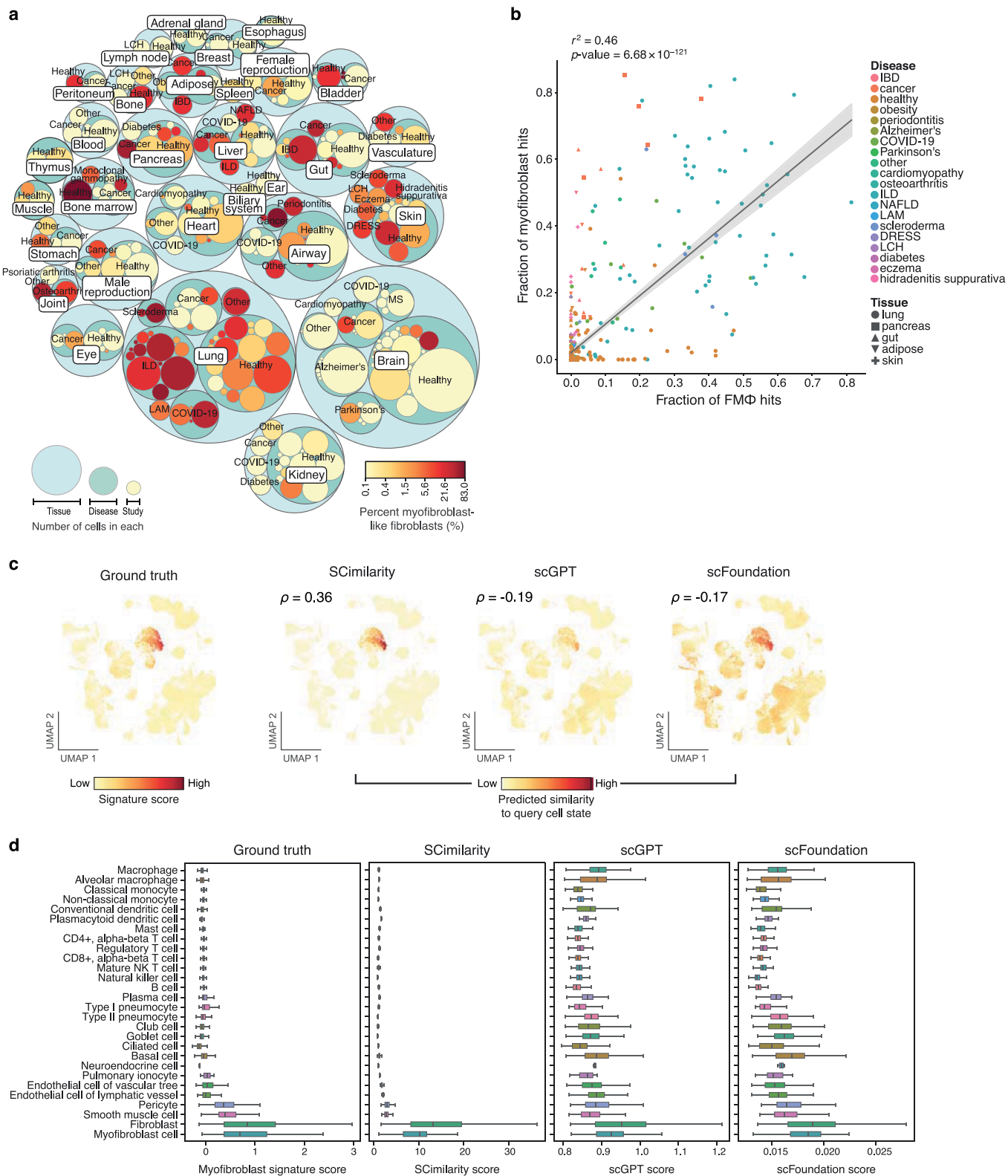
**Extended Data Fig. 5 | Validation of cell type annotation on CITE-seq of PBMCs. a.** Author annotations. UMAP embedding of single-cell profiles (dots) from SCimilarity's latent representation of PBMCs profiled by CITE-seq<sup>33</sup>. **b-i.** SCimilarity's annotation accuracy is on par or better than three other methods. Left: UMAP embedding (as in **a**) of cell profiles coloured by annotations predicted by SCimilarity (**b**), CellTypist (**d**), TOSICA (**f**), or

scANVI (**h**). Right: Percentage (colour bar) and number of author-annotated cells (columns) matching annotations predicted by SCimilarity (**c**), CellTypist (**e**), TOSICA (**g**), or scANVI (**i**) (rows). **j.** Surface marker protein levels of selected cell populations. Distribution (y axis) and median level within population (colour bar) of author-normalized protein levels for selected markers (rows) across cell types (x axis) for author (left) and SCimilarity (left) annotated cells.



**Extended Data Fig. 6 | SCimilarity annotations and gene attributions capture known biology. a**, SCimilarity annotated cell type profiles group by correct biological relations. Hierarchical clustering (average linkage with cosine distance) of centroids profiles of predicted cell types (leaves) in SCimilarity latent

space, coloured by lineage. **b**, SCimilarity cell type important genes match cell type specific signatures. Fraction of cell type-specific differentially expressed genes (from Eraslan et al.<sup>8</sup>) (y axis) captured by top-n important genes (x axis) for that cell type by SCimilarity's integrated gradients attribution analysis.

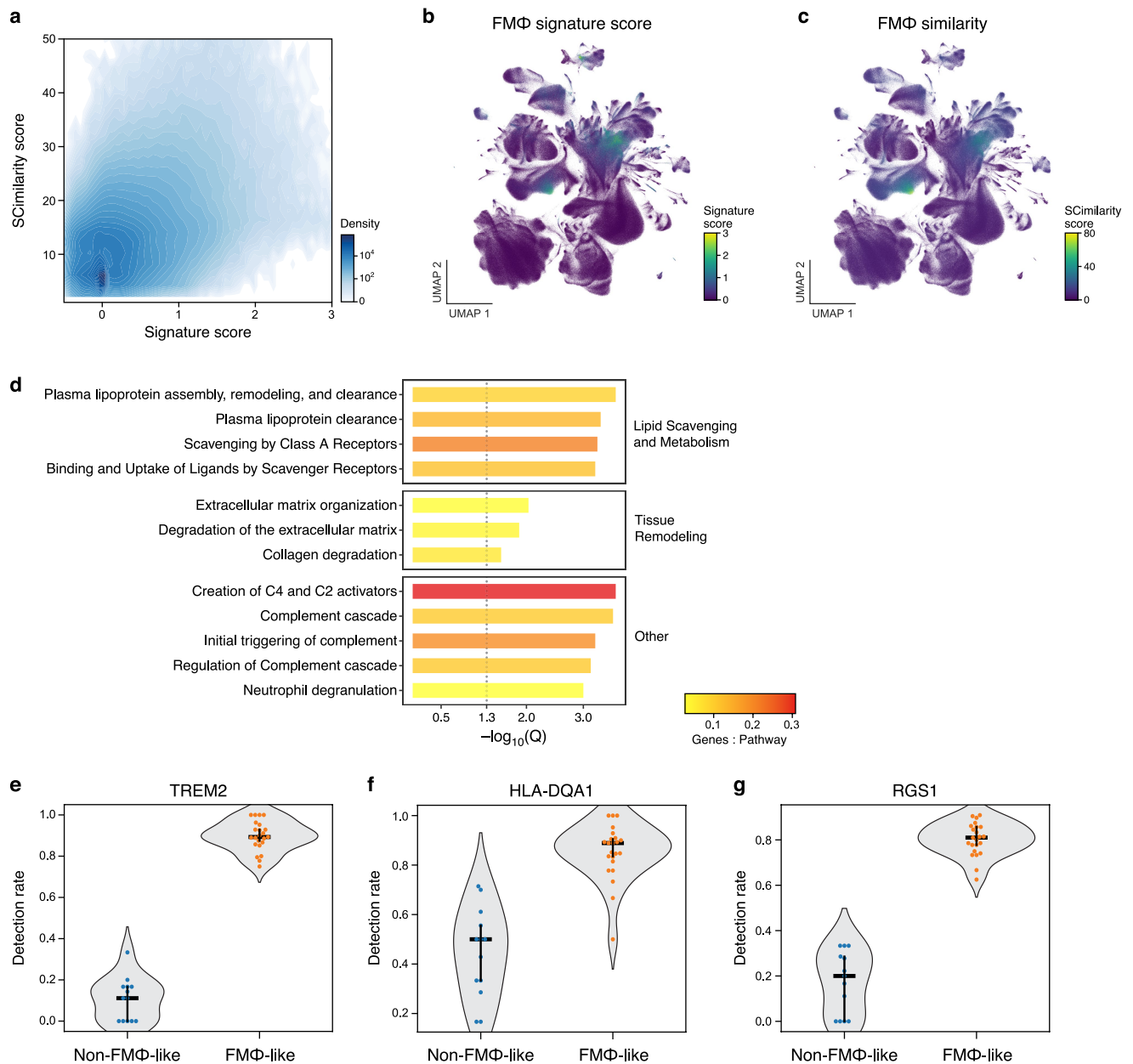


**Extended Data Fig. 7** | See next page for caption.

**Extended Data Fig. 7 | Fibrosis-associated myofibroblasts correlate with presence of fibrosis-associated macrophages across tissues and diseases.**

**a**, Myofibroblasts are prevalent across tissues and diseases. Number of cells (circle size) across tissues (outermost blue circles), disease states (middle green circles), and individual studies (innermost circles, coloured by fraction of cells annotated as fibroblasts or myofibroblasts with SCimilarity scores  $>95^{\text{th}}$  percentile of total fibrosis-associated myofibroblast query scores (log scaled colour bar)). Circle size for disease and study are scaled relative to other diseases in the same tissue or studies in the same disease. **b**, Fibrosis-associated macrophages and myofibroblasts are correlated across conditions. Fractions of FM $\Phi$ -like cells (x axis; FM $\Phi$  query hits as a fraction of total cells annotated as monocytes or macrophages) and fibrosis-associated myofibroblasts (y axis; fibrosis-associated myofibroblast query hits as a fraction of total cells annotated as fibroblasts or myofibroblasts) in each *in vivo* sample (dots, coloured by

condition) containing  $>50$  monocytes/macrophages and  $>50$  fibroblasts/myofibroblasts with a linear fit (black line) and 95% confidence interval round the fit (grey band). Inset box: Pearson correlation ( $r^2$ ) and nominal two-sided t test p-value for the correlation. **c,d**, SCimilarity better retrieves a myofibroblast query than LLM-based models. **c**, UMAP of cells from the ILD study GSE128033 with cells coloured by a myofibroblast signature score (ground truth) or similarity to the myofibroblast query state for SCimilarity (right, first), scGPT (right, second), or scFoundation (right, third). Top left: Spearman's  $\rho$  between signature score rankings and distances to the query cell. **d**, Distribution of myofibroblast signature (first), SCimilarity (second), scGPT (third), and scFoundation (fourth) scores as in (c) for  $n = 28$  SCimilarity predicted cell types across  $n = 58,530$  total cells (outliers removed). Boxplot: upper/lower quartiles (box), min/max values (whiskers), and median (center line).



#### Extended Data Fig. 8 | FMΦs among monocytes and macrophages.

**a-c**, Agreement between SCimilarity and traditional FMΦ cell scores. **a**, Scanpy FMΦ gene signature score (x axis) and FMΦ SCimilarity score using a prototypical FMΦ cellular profile defined from Adams et al.<sup>1</sup> (y axis) for each cell (density shown as colour intensity). **b,c**, UMAP embedding of  $n = 2,578,221$  monocyte and macrophage cell profiles (dots) from SCimilarity's latent space representation coloured by SCimilarity score using a prototypical FMΦ cellular profile defined from Adams et al.<sup>1</sup> (**b**) or by Scanpy's signature score for FMΦ associated genes (**c**). **d**, FMΦ important genes are enriched for relevant pathways. False Discovery Rate ( $-\log_{10}(Q)$  value), hypergeometric test, x axis) for enrichment of Reactome pathways (y axis,  $Q \leq 0.05$  and gene count  $\geq 4$ ) with the 100 genes with the top

integrated gradients attribution scores for the FMΦ query (ranked by score). Colour: ratio of important genes within a Reactome pathway to the total size of the pathway. **e-g**, Expression of known and novel genes associated with FMΦs. Distribution of the fraction of cells (y axis) in ILD tissue samples (dots) among  $n = 500$  randomly sampled FMΦ-like (top 10,000 cells by SCimilarity score) cells (orange,  $n = 23$  tissue samples) and  $n = 500$  randomly sampled non-FMΦ-like (remaining cells) macrophages and monocytes (blue,  $n = 13$  tissue sample) that express ( $>0$  UMI counts) the known FMΦ marker TREM2 (**e**) and two FMΦ-enriched genes not previously described for FMΦs (**f,g**). Crossbar: upper/lower quartiles (vertical line) and median (horizontal line).



Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our [Editorial Policies](#) and the [Editorial Policy Checklist](#).

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
<input type="checkbox"/>	<input checked="" type="checkbox"/> The exact sample size ( <i>n</i> ) for each experimental group/condition, given as a discrete number and unit of measurement
<input type="checkbox"/>	<input checked="" type="checkbox"/> A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
<input type="checkbox"/>	<input checked="" type="checkbox"/> The statistical test(s) used AND whether they are one- or two-sided <i>Only common tests should be described solely by name; describe more complex techniques in the Methods section.</i>
<input checked="" type="checkbox"/>	<input type="checkbox"/> A description of all covariates tested
<input type="checkbox"/>	<input checked="" type="checkbox"/> A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
<input type="checkbox"/>	<input checked="" type="checkbox"/> A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
<input type="checkbox"/>	<input checked="" type="checkbox"/> For null hypothesis testing, the test statistic (e.g. <i>F</i> , <i>t</i> , <i>r</i> ) with confidence intervals, effect sizes, degrees of freedom and <i>P</i> value noted <i>Give P values as exact values whenever suitable.</i>
<input checked="" type="checkbox"/>	<input type="checkbox"/> For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
<input checked="" type="checkbox"/>	<input type="checkbox"/> For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
<input type="checkbox"/>	<input checked="" type="checkbox"/> Estimates of effect sizes (e.g. Cohen's <i>d</i> , Pearson's <i>r</i> ), indicating how they were calculated

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

Software and code

Policy information about [availability of computer code](#)

Data collection	Public sc/snRNA-seq data from the Gene Expression Omnibus (GEO) was collected using a combination of the Bio.Entrez module from Biopython v1.81, GEParse v2.0.3, and custom shell scripts; data sourced from other repositories was collected by manual download or custom shell scripts.
Data analysis	Data analysis in this study was performed using scanpy v1.9.2, souporecell v2.0, harmonypy v0.0.9, Scanorama v1.7.4, scVI and scArches v1.1.1.Orc2, Hnswlib v0.8.0, CellTypist v1.6.2, TOSICA v1.0.0, scGPT v0.2.1 (June 23, 2023 model), scFoundation (December 9, 2023 model), ReactomePA v1.40.0, and custom Python and R scripts. The SCimilarity code base is available from <a href="https://github.com/Genentech/scimilarity">https://github.com/Genentech/scimilarity</a> .

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio [guidelines for submitting code & software](#) for further information.

## Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our [policy](#)

In vitro data generated in this study have been deposited in the Gene Expression Omnibus (GEO) under accession GSE280632. Model weights, single-cell data embeddings, curated metadata and k-NN graphs have been deposited on Zenodo with DOI 10.5281/zenodo.10685499 (<https://zenodo.org/records/10685499>). Source repositories and accession numbers for the public sc/snRNA-seq studies used for model training, model testing, or as part of the unlabeled referenced set are provided as supplementary tables.

## Research involving human participants, their data, or biological material

Policy information about studies with [human participants or human data](#). See also policy information about [sex, gender \(identity/presentation\), and sexual orientation](#) and [race, ethnicity and racism](#).

Reporting on sex and gender	This information was not collected or referenced.
Reporting on race, ethnicity, or other socially relevant groupings	This information was not collected or referenced.
Population characteristics	No population characteristics were relevant for this study.
Recruitment	There was no recruitment for this study.
Ethics oversight	Peripheral blood samples from healthy volunteers were provided by the Samples for Science (S4S) donor program at Genentech; donors provided written informed consent and sample collection was approved by the Western-Copernicus Group Institutional Review Board.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

## Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](https://www.nature.com/documents/nr-reporting-summary-flat.pdf)

## Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Sample size for the experiment performed as part of this study was determined based on experimental feasibility, mitigation of batch effects, and replicate consistency in prior work. No sample size calculations were performed.
Data exclusions	No data was excluded from the experiment performed as part of this study. For data collection, public sc/snRNA-seq data source from GEO was excluded if it did not match keywords denoting it was human samples from the 10x Chromium platform; data sets were further excluded if they could not be read in using loaders for .mtx, .h5ad, .tsv or .csv formats.
Replication	The experiment performed as part of this study was performed once with three biological replicates. No replicates were discarded.
Randomization	The experiment performed as part of this study had only a single experimental group, so randomization was not applicable.
Blinding	The experiment performed as part of this study had only a single experimental group, so blinding was not applicable.

## Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

## Materials &amp; experimental systems

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> Antibodies
<input checked="" type="checkbox"/>	<input type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology and archaeology
<input checked="" type="checkbox"/>	<input type="checkbox"/> Animals and other organisms
<input checked="" type="checkbox"/>	<input type="checkbox"/> Clinical data
<input checked="" type="checkbox"/>	<input type="checkbox"/> Dual use research of concern
<input checked="" type="checkbox"/>	<input type="checkbox"/> Plants

## Methods

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input checked="" type="checkbox"/>	<input type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging

## Plants

## Seed stocks

Report on the source of all seed stocks or other plant material used. If applicable, state the seed stock centre and catalogue number. If plant specimens were collected from the field, describe the collection location, date and sampling procedures.

## Novel plant genotypes

Describe the methods by which all novel plant genotypes were produced. This includes those generated by transgenic approaches, gene editing, chemical/radiation-based mutagenesis and hybridization. For transgenic lines, describe the transformation method, the number of independent lines analyzed and the generation upon which experiments were performed. For gene-edited lines, describe the editor used, the endogenous sequence targeted for editing, the targeting guide RNA sequence (if applicable) and how the editor was applied.

## Authentication

Describe any authentication procedures for each seed stock used or novel genotype generated. Describe any experiments used to assess the effect of a mutation and, where applicable, how potential secondary effects (e.g. second site T-DNA insertions, mosaicism, off-target gene editing) were examined.