nature biotechnology

Brief Communication

https://doi.org/10.1038/s41587-025-02702-z

Limitations of cell embedding metrics assessed using drifting islands

Received: 2 April 2024

Accepted: 8 May 2025

Published online: 11 June 2025

Check for updates

Hanchen Wang [1,2], Jure Leskovec [2] ✉ & Aviv Regev [1] ✉

Biological studies rely on embeddings of single-cell profiles but assessing the quality of these embeddings is challenging. Here we show that current evaluation metrics are incomplete by training a three-layer perceptron, Islander. Islander outperforms all leading embedding methods on a diverse set of cell atlases but it distorts biological structures, limiting its use for biological discovery. We then present a new metric, scGraph, to help flag such distortions.

Embeddings of single-cell profiles are now routinely used as a research tool in biological investigation to characterize cell types and states, their changes over time and their distinction between conditions, including diseases, organs or drug treatments[1,2]. With a dramatic growth in single-cell data, including the Human Cell Atlas[3,4], multiple efforts have focused on learning universal embeddings for diverse single-cell data, with different integration methods or foundation models[5–10]. Given their broad use, it is crucial to scrutinize the quality of embeddings to evaluate the performance of the underlying integration methods[11–13] and zero-shot capabilities of the resulting foundation models[14,15]. Thus, development of new successful methods and models also relies on good evaluation metrics.

A critical aspect in deriving helpful cell embeddings is the correction of nonbiological batch effects that stem from technical variations, such as sample handling and sequencing protocols. These unwanted variations can mask biological signals and lead to misleading interpretations. Integration methods, thus, aim to mitigate batch-specific discrepancies while preserving essential biological variation. The effectiveness of these integrated cell embeddings is typically assessed through two evaluation lenses: how well the cells from various batches mix together and how closely cells of the same type group together.

Here, we identified an overlooked challenge in the evaluation metrics used to assess embeddings. To demonstrate the limitations of current gold-standard metrics for cell profile embeddings[11], we developed Islander (Fig. 1a), a model that scores best on established evaluation metrics but generates biologically problematic embeddings. Islander is a three-layer perceptron, directly trained on cell type annotations with mixup augmentations[16]. We tested Islander across a diverse set of 11 different human tissue cell atlases (brain[17], spanning breast[18], eye[19], fetal gut[20], heart[21], fetal lung[22], pancreas[11] and skin[23]), which together cover different strengths of batch effects and diverse biological systems, overall comprising more than 3.5 million cells from ten human organ systems (Extended Data Table 1). For each atlas, we trained an Islander model and then compared it with another 13 embedding baselines: three dimension reduction methods (principal component analysis (PCA), uniform manifold approximation and projection (UMAP) and *t*-distributed stochastic neighbor embedding (tSNE))[24], eight batch integration methods (Harmony[25], Scanorama[26], BBKNN[27], fastMNN[28], scVI[29], scANVI[30], scGen[31] and scPoli[32]) and two foundation models (Geneformer[6] and scGPT[9]) (Methods). In addition, for each atlas, we compared to the performance of the original authors' integration, if available.

Across all datasets, Islander consistently outperformed all baseline strategies across all 12 metrics[11] (Fig. 1b,c, Extended Data Table 2 and Supplementary Tables 3–12). This is largely because of the principles underlying the evaluation metrics[11], which focus on assessing the efficiency of cell embeddings in terms of the coherence of cell clustering structures with cell type labels and the blending of batches within clusters. When Islander explicitly aligns these cell embeddings with cell type annotations, it forms well-separated cell 'islands' (Fig. 1d, right), with each island comprising cells annotated as the same type. This alignment greatly boosts the biological variance conservation metrics, leading to top-tier overall performance by these evaluation criteria (Supplementary Tables 3–12). To explore the impact of supervision signals, we also trained Islander's model using two semisupervised losses, contrastive[33] and triplet[34]. Both achieved high scIB scores but resulted in distinctly different structures (Supplementary Information).

While such structure is driven by (and complies well) with the most granular annotation level, it comes at the cost of ignoring any higher level relationships between cells and, thus, distorts biological structures, potentially obstructing downstream analyses and future discoveries; therefore, this would not be advisable for an actual integration method. In particular, when annotated cell subsets follow a continuum, as is the case for fibroblasts, Islander separates its constituent parts

[1]Genentech Research and Early Development, Genentech, South San Francisco, CA, USA. [2]Department of Computer Science, Stanford University, Palo Alto, CA, USA. ✉e-mail: jure@cs.stanford.edu; regev.aviv@gene.com

https://doi.org/10.1038/s41587-025-02702-z

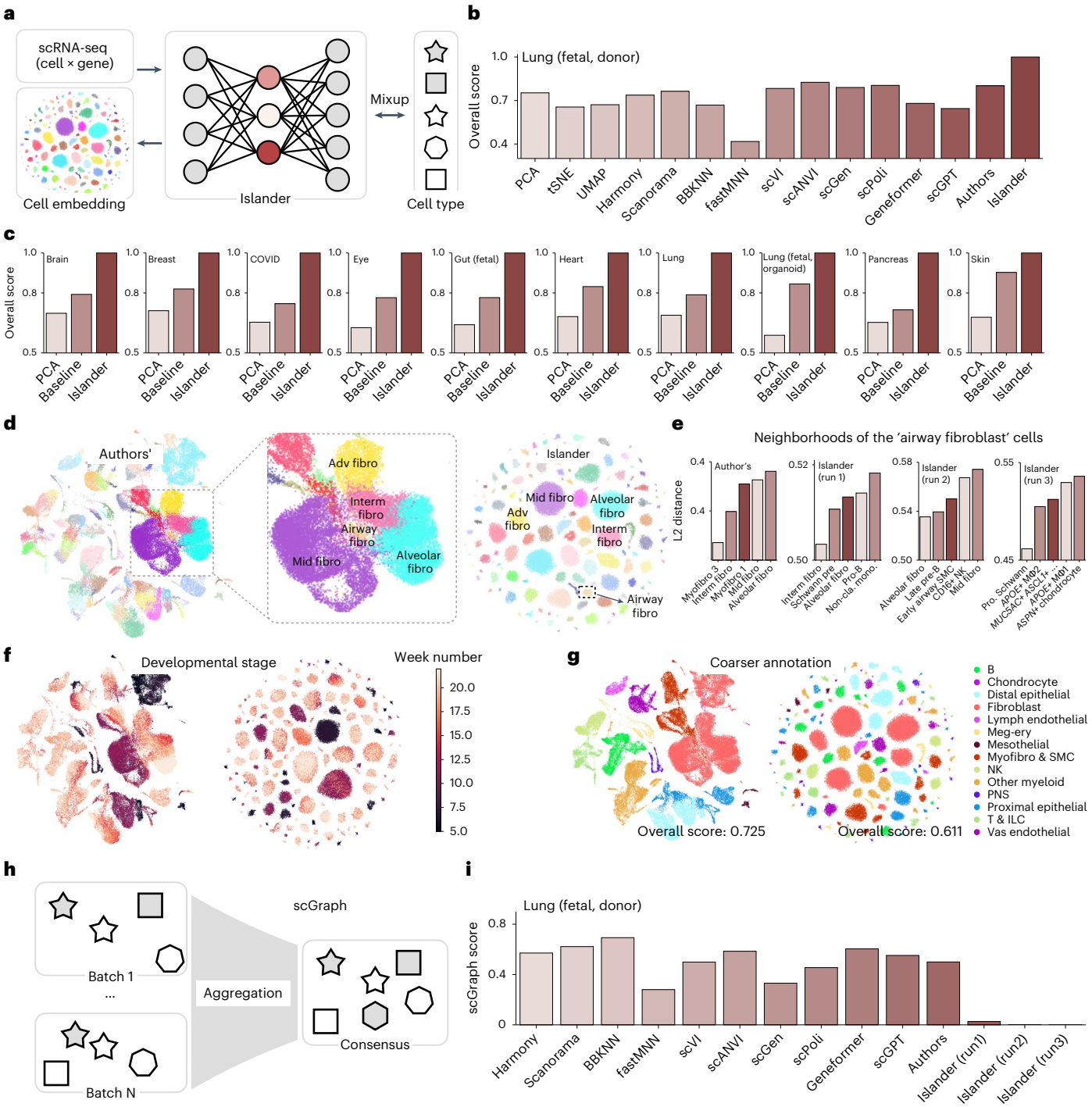


Fig. 1 | Drifting cell islands highlight limitations of current evaluation metrics. a, Islander overview. **b**,**c**, Evaluation of cell embeddings. Normalized overall score (*y* axis; Methods) over 12 metrics for each integration method (*x* axis) assessed using the Fetal Lung Cell Atlas (**b**) or ten other cell atlases (**c**). Baseline refers to the best baseline results (Methods). **d**, Fetal Lung Cell Atlas embedding space. Single-cell profiles (dots; color-coded by cell type annotation) from the Fetal Lung Cell Atlas embedded by the authors' integration method (left; with zoomed-in view in inset) or Islander (right). Annotations denote the fibroblast subsets. Adv fibro, adventitial fibroblasts. **e**, 'Airway fibroblast' cell neighborhood changes across Islander runs. Normalized Euclidean distance (*y* axis) between the centroids of airway fibroblast profiles and those of its five nearest neighbor clusters (*x* axis) in the 50-dimensional (authors' integration) and 16-dimensional (Islander's integration) embedding space. Non-cla. mono., non-classical monocytes; SMC, smooth muscle cells; NK, natural killer cells; Pro. Schwann, progenitor Schwann cells. **f**,**g**, Cell islands distort developmental stage structure and cell–cell relationships. Cell embeddings are as in **d**, colored by developmental week (**f**; color bar) or coarse cell type annotations (**g**). PNS, peripheral nervous system cells; T & ILC, T lymphocytes and innate lymphoid cells. **h**,**i**, scGraph, an evaluation metric for cell embedding using learned cell similarity graphs: method overview (**h**) and scGraph score (*y* axis, where higher is better) for each method (*x* axis) assessed using the Fetal Lung Cell Atlas (**i**).

(Fig. 1d). In the developing human lung, the original analysis[22] identified multiple subtypes of fibroblasts, each distinguished by different marker genes and spatial locations. While the original embedding preserves a continuum between these fibroblasts (Fig. 1d, left), they are fully separated by Islander into separate islands (Fig. 1d, right). Similarly, the Islander embedding disrupted the developmental continuum, clearly observed in the original study (Fig. 1f, left) but obscured by Islander (Fig. 1f, right).

Moreover, the cell islands drifted in different ways across distinct runs, especially for smaller cell subsets. For example, in three separate runs with overall similar scores, the composition of the neighborhood of airway fibroblasts varied substantially, involving as many as 14 distinct cell types within the five nearest neighbors (Fig. 1e, Extended Data Fig. 1 and Extended Data Table 2). Thus, aside from cluster identity, the embedding may be largely arbitrary in all other relationships and this arbitrariness would carry into downstream analysis or the biologist's interpretation.

Prompted by these limitations of established quality evaluation criteria, we reasoned that focusing solely on the most granular cell relationships in evaluation can pose substantial limitations, whereas preserving relationships between broader cell types (coarser annotations) is an important additional criterion and may also be more robust to noise. Indeed, when evaluating the same set of embeddings using broader cell type annotations provided by the authors, Islander now achieved an overall score of 0.523, inferior to PCA (0.557) or the top-performing scVI[29] (0.701) (Extended Data Table 3).

Because hierarchical Cell Ontology annotations are often unavailable[7], we next developed scGraph (Methods) as a new evaluation metric to complement and augment exiting metrics for quality assessment. In this framework, for each set of cell embeddings, an affinity graph is defined to elucidate the similarities between various cell types. scGraph compares each affinity graph to a consensus graph, derived by aggregating individual graphs from different batches on the basis of batch-wise PCA loadings. We used ablation studies, including the trimming of outlier cells, to design scGraph (Methods and Extended Data Fig. 2). The framework effectively highlights inherent biological structures and emphasizes cell type similarities while minimizing technical variations across batches. Notably, scGraph does not require any single batch to contain cells of all types, making it suitable for real datasets with diverse constraints.

Embedding methods differed in their performance according to the scGraph metric versus established scIB metrics (Extended Data Table 4). In comparing cell embeddings on this particular fetal lung collection using scGraph (Fig. 1i), BBKNN achieved the highest scores, as it was the authors' chosen method for integration and annotation. While the final authors' annotations showed midlevel performance overall, they accurately captured fine details, such as the fibroblast family (Extended Data Fig. 3 and Extended Data Table 5). In this case, interbatch variance is biologically meaningful, as batches correspond to developmental stages (with multiple batches potentially representing the same time point) and scGraph effectively preserves key biological features, including trajectories and subpopulations. In contrast, scIB promotes interbatch mixing and distinct cell type separation, favoring models such as Islander, scGen and scANVI but overlooking developmental dynamics. scGraph provided more biologically meaningful rankings, where PCA and the authors' annotations excelled in capturing developmental trajectories. However, when interbatch variance primarily reflects unwanted technical noise (for example, single-cell RNA sequencing (scRNA-seq) versus single-nucleus RNA-seq), scIB is advantageous in minimizing these effects. Thus, scGraph and scIB complement each other and using both is crucial for a comprehensive assessment of embeddings. Moreover, because most evaluation metrics require harmonized annotations, resources such as the Human Lung Cell Atlas[35] and CellHint[36] tutorials can be particularly useful.

While designing an alternative 'null' algorithm (similar to Islander) to optimize scGraph is possible, it is more challenging, highlighting its robustness and reliability as an evaluation metric. However, scGraph has its biases. Like scIB, it tends to favor higher-dimensional embeddings such as PCA over PCA-derived UMAP and it is not based on single-cell-level calculations. Additionally, its assumption that functionally similar cells should be proximal in the embedding space may not always hold true. Despite these limitations, scGraph represents a step toward developing more robust frameworks for evaluating embeddings by capturing diverse aspects of biological relevance and structural integrity.

In conclusion, we demonstrated the limitation of current quality metrics by introducing Islander, a three-layer perceptron, as a null algorithm for an integration approach that outperforms all major methods across diverse cell atlases, even though it introduces island-like distortions in the biological structures in cell embedding spaces. Islander serves as a touchstone for evaluating and refining future evaluation metrics. To address the limitations Islander highlighted in current evaluation frameworks, we further propose augmenting those with scGraph as an additional, complementary metric, designed to assess how well embeddings preserve cell–cell relationships at multiple levels of granularity. scGraph specifically focuses on the consistency of cell relationships before and after integration, rather than on the closeness of similar types and the mixing of different batches. As a result, it is effective at detecting artifacts such as drifting cell islands and 'zig-zag' structures (Supplementary Fig. 1), offering a unique perspective on the preservation of biological structure across batches. Both scIB and scGraph rely on the assumption of a Euclidean distance within the embedding space, which typically favors higher-dimensional representations. Similarly, their use is constrained to datasets with harmonized annotations. Furthermore, because scGraph constructs its reference using PCA-based loadings, it may preferentially favor embeddings that resemble PCA or those influenced by the authors' annotations. While scGraph is useful for benchmarking large-scale atlas integration, it is especially informative for smaller, focused datasets with finer annotations, where interbatch variance reflects meaningful biological differences rather than technical noise (for example, subtypes of fibroblasts in the human fetal lung atlas; Extended Data Fig. 3). In such cases, scGraph provides deeper insight into the preservation of continuous cell state transitions, lineage hierarchies and rare subpopulations that may be masked by overcorrections among batches. Unlike traditional batch-mixing metrics that emphasize the uniformity of cell type distributions across batches, scGraph enables the identification of biologically meaningful structure, such as lineage-specific divergence patterns or subtle state transitions that remain distinct across conditions. This makes it a valuable complement to existing evaluation frameworks, ensuring that integration methods do not inadvertently distort biologically relevant variation while correcting for technical effects.

Overall, by capturing aspects of biological structure that scIB might overlook (and vice versa), scGraph and scIB together form a more comprehensive evaluation framework for integration methods and their resulting cell embeddings.

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/s41587-025-02702-z.

References

1. de Sande, B. V. et al. Applications of single-cell RNA sequencing in drug discovery and development. *Nat. Rev. Drug Discov.* **22**, 496–520 (2023).
2. Zhang, M. J. et al. Polygenic enrichment distinguishes disease associations of individual cells in single-cell RNA-seq data. *Nat. Genet.* **54**, 1572–1580 (2022).

3. Rood, J. E. et al. Impact of the Human Cell Atlas on medicine. *Nat. Med.* **28**, 2486–2496 (2022).
4. Rood, J. E. et al. The Human Cell Atlas from a cell census to a unified foundation model. *Nature* **637**, 1065–1071 (2025).
5. Hao, Y. et al. Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573–3587.e29 (2021).
6. Theodoris, C. V. et al. Transfer learning enables predictions in network biology. *Nature* **618**, 616–624 (2023).
7. Heimberg, G. et al. A cell atlas foundation model for scalable search of similar human cells. *Nature* **638**, 1085–1094 (2025).
8. Rosen, Y. et al. Universal cell embeddings: a foundation model for cell biology. Preprint at *bioRxiv* https://doi.org/10.1101/2023.11.28.568918 (2023).
9. Cui, H. et al. scGPT: toward building a foundation model for single-cell multi-omics using generative AI. *Nat. Methods* **21**, 1470–1480 (2024).
10. Hao, M. et al. Large-scale foundation model on single-cell transcriptomics. *Nat. Methods* **21**, 1481–1491 (2024).
11. Luecken, M. D. et al. Benchmarking atlas-level data integration in single-cell genomics. *Nat. Methods* **19**, 41–50 (2022).
12. Tran, H. T. N. et al. A benchmark of batch-effect correction methods for single-cell RNA sequencing data. *Genome Biol.* **21**, 12 (2020).
13. Wang, H. et al. Scientific discovery in the age of artificial intelligence. *Nature* **620**, 47–60 (2023).
14. Liu, T., Li, K., Wang, Y., Li, H. & Zhao, H. Evaluating the utilities of foundation models in single-cell data analysis. Preprint at *bioRxiv* https://doi.org/10.1101/2023.09.08.555192 (2023).
15. Kedzierska, K. Z., Crawford, L., Amini, A. P. & Lu, A. X. Zero-shot evaluation reveals limitations of single-cell foundation models. *Genome Biol.* **26**, 101 (2025).
16. Zhang, H., Cisse, M., Dauphin, Y. N. & Lopez-Paz, D. mixup: beyond empirical risk minimization. Preprint at https://arxiv.org/abs/1710.09412 (2018).
17. Siletti, K. et al. Transcriptomic diversity of cell types across the adult human brain. *Science* **382**, eadd7046 (2023).
18. Kumar, T. et al. A spatially resolved single-cell genomic atlas of the adult human breast. *Nature* **620**, 181–191 (2023).
19. Wang, S. K. et al. Single-cell multiome of the human retina and deep learning nominate causal variants in complex eye diseases. *Cell Genom.* **2**, 100164 (2022).
20. Elmentaite, R. et al. Single-cell sequencing of developing human gut reveals transcriptional links to childhood Crohn's disease. *Dev. Cell* **55**, 771–783.e5 (2020).
21. Knight-Schrijver, V. R. et al. A single-cell comparison of adult and fetal human epicardium defines the age-associated changes in epicardial activity. *Nat. Cardiovasc. Res.* **1**, 1215–1229 (2022).
22. He, P. et al. A human fetal lung cell atlas uncovers proximal–distal gradients of differentiation and key regulators of epithelial fates. *Cell* **185**, 4841–4860.e25 (2022).
23. Solé-Boldo, L. et al. Single-cell transcriptomes of the human skin reveal age-related loss of fibroblast priming. *Commun. Biol.* **3**, 188 (2020).
24. Heumos, L. et al. Best practices for single-cell analysis across modalities. *Nat. Rev. Genet.* **24**, 550–572 (2023).
25. Korsunsky, I. et al. Fast, sensitive and accurate integration of single-cell data with harmony. *Nat. Methods* **16**, 1289–1296 (2019).
26. Hie, B., Bryson, B. & Berger, B. Efficient integration of heterogeneous single-cell transcriptomes using scanorama. *Nat. Biotechnol.* **37**, 685–691 (2019).
27. Polański, K. et al. BBKNN: fast batch alignment of single cell transcriptomes. *Bioinformatics* **36**, 964–965 (2020).
28. Haghverdi, L. et al. Batch effects in single-cell rna-sequencing data are corrected by matching mutual nearest neighbors. *Nat. Biotechnol.* **36**, 421–427 (2018).
29. Lopez, R. et al. Deep generative modeling for single-cell transcriptomics. *Nat. Methods* **15**, 1053–1058 (2018).
30. Xu, C. et al. Probabilistic harmonization and annotation of single-cell transcriptomics data with deep generative models. *Mol. Syst. Biol.* **17**, e9620 (2021).
31. Lotfollahi, M., Wolf, F. A. & Theis, F. J. scGen predicts single-cell perturbation responses. *Nat. Methods* **16**, 715–721 (2019).
32. De Donno, C. et al. Population-level integration of single-cell datasets enables multi-scale analysis across samples. *Nat. Methods* **20**, 1683–1692 (2023).
33. Khosla, P. et al. Supervised contrastive learning. In *Advances in Neural Information Processing Systems 33* (eds Larochelle, H. et al.) 18661–18673 (NeurIPS, 2020).
34. Hoffer, E. & Ailon, N. Deep metric learning using triplet network. In *Similarity-Based Pattern Recognition: SIMBAD 2015* (eds Feragen, A. et al.) 84–92 (Springer, 2015).
35. Sikkema, L. et al. An integrated cell atlas of the human lung in health and disease. *Nat. Med.* **29**, 1563–1577 (2023).
36. Xu, C. et al. Automatic cell-type harmonization and integration across Human Cell Atlas datasets. *Cell* **186**, 5876–5891.e20 (2023).

Methods

Datasets and preprocessing

Raw sequencing data were downloaded from the respective data providers on October 1, 2023, as detailed in Extended Data Table 1. A total of 11 cell atlases were analyzed, totaling 3,510,450 cell profiles. A uniform preprocessing protocol was applied across the datasets. Cell profiles with fewer than 1,000 reads or fewer than 500 detected genes were filtered out and genes present in fewer than five cells were also excluded. Normalization was performed using Scanpy[37], scaling each cell's read counts to a total of 10,000 and subsequently applying a log1p transformation.

Baselines

A total of 13 methods were used for comparison, three dimensionality reduction baselines (PCA, tSNE[38] and UMAP[39]), eight integration methods (Harmony[25], BBKNN[27], Scanorama[26], fastMNN[28], scVI[29], scANVI[30], scGen[31] and scPoli[32]) and two pretrained foundation models (Geneformer[6] and scGPT[9]), for both zero-shot and fine-tuned embedding extraction. For dimensionality reduction methods, the log1p-transformed raw counts from gene-by-cell matrices were provided as input. For each integration method, an independent grid search was conducted around the default recommended hyperparameter settings (Supplementary Note). For Geneformer and scGPT, the largest pretrained model weights provided by the authors[9] (https://huggingface.co/ctheodoris/Geneformer/tree/main/) were used. While scANVI, scGen, scPoli and fine-tuned scGPT use cell type as parts of their computational pipelines, other methods do not require such information. The top 1,000 highly variable genes were identified.

Assessment metrics

Cell embeddings were assessed using established evaluation metrics as previously described[11] and implemented in scib-metrics (https://scib-metrics.readthedocs.io/en/stable/). The following evaluation metrics were used (abbreviations noted are used in Extended Data Tables 1–5 and Supplementary Tables 3–12): isolated labels (I-label), Leiden normalized mutual information (L-NMI), Leiden averaged Rand index (L-ARI), *k*-means NMI and ARI (K-NMI and K-ARI), silhouette label and batch (S-label and S-batch), batch mixing (iLISI) and cell type separation (cLISI), graph connectivity (G-Con) and principal component regression (PCR). Consistent with previous studies, selection of highly variable genes enhanced the performance of data integration methods.

Islander design

Islander is a three-layer perceptron with two hidden layers of sizes 128 and 16 and an output layer matching the total number of cell types as annotated. The first hidden layer incorporates rectified linear unit (ReLU) activation and batch normalization. Cell embeddings are derived from the last hidden layer, which, while trained using one-hot encoding in the final layer, does not rely on one-hot encoding for the embedding itself. The output layer uses a softmax normalization function. Each layer in this extended setup uses ReLU activation and batch normalization, except for the final linear layer.

Training setup

The model was trained in a manner aligned with scvi-tools[40], with minibatches of 256 randomly sampled cells from all batches, along with their cell type annotations. Islander was trained using cross-entropy loss with mixup[16] augmentations (default setting). The Adam optimizer was used with an initial learning rate of 0.001 over ten epochs and a cosine annealing scheduler for learning rate decay. All cells were used for training to maximize overfitting.

Impact of semisupervised loss formats with Islander

Two Islander variants were explored using the same neural architecture but different loss functions: triplet loss (Tri)[34] and supervised contrastive loss (SCL)[33] (Supplementary Note). While both adhered to semisupervised learning principles and achieved high scIB scores, they displayed contrasting behaviors. The SCL variant tended to produce problematic cell embeddings, detectable by scGraph, while the Tri variant fostered a more biologically valid embedding space. Strategies such as encoder regularization with additional losses (for example, reconstruction or unsupervised large-scale pretraining) emphasize the importance of careful metric development and offer guidance for future computational biology research.

Neighborhood calculation

Neighborhoods of each cell type were identified by the Euclidean distance between the centroids of cell profiles of each type in the embedding space. To mitigate the effects of batch variation and measurement noise, a trimming strategy was applied. The outlying 5% of cells on both sides were excluded before calculating the centroid coordinates. This ensures a more accurate representation of cell type proximity by focusing on the most representative data points. Evidence supporting the rationale behind these design choices is provided in Extended Data Fig. 2.

Design of scGraph

scGraph quantifies the similarity between two graphs that each represent the closeness between cell types. In these graphs, each entry (x, y) denotes the proximity of cell type x to cell type y. The goal is to align the neighborhood graphs from the embeddings with the reference graph derived from the data, indicating that cells with similar profiles are appropriately clustered in the embedding space. The first graph is derived from the provided embeddings, while the second, serving as a reference, is based on batch-wise PCA loadings from each batch. For the reference, proximity graphs are initially computed from each batch using normalized Euclidean distances between centroids of the cell type profiles. These batch-specific graphs are then amalgamated into a single consensus graph through averaging. The similarity of neighborhoods for each cell type is assessed using weighted Pearson correlation, where the weights are inversely proportional to the distances. This modification provides a ranking similar to simple correlation but with greater emphasis on closer neighbors (capturing finer structure) while reducing the influence of more distant cell types. The final score, reflecting the overall similarity and ranging from −1 to 1 (with higher values indicating greater similarity), is the average across all cell types. Notably, computing scGraph scores does not require every cell type to be present in each batch nor does it require the graphs to be fully connected. Lastly, as scGraph constructs its reference using PCA-based loadings, it may favor embeddings that resemble PCA or those where the authors made annotations.

Practical usage considerations of scIB and scGraph

scGraph evaluates how well cell type relationships are preserved across batches by computing Euclidean distances between PCA-based centroids and averaging them into a consensus matrix. To assess stability, the variance-to-mean ratio of these distances across batches can be measured. Low variance suggests stable biological relationships, while high variance indicates batch-specific shifts, which could stem from either technical artifacts or meaningful biological variation. When batch effects are purely technical, scIB's emphasis on mixing is more appropriate, whereas scGraph is more useful when interbatch variation carries biological importance.

Because scGraph scores depend on dataset context, comparing integration methods relative to each other is often more informative than interpreting absolute values. Differences of 0.05–0.1 in scores can signal meaningful improvements or artifacts but no universal threshold defines 'good' performance. For example, in datasets with expected biological heterogeneity across batches, lower scGraph scores may reflect true biological differences rather than poor integration.

Conversely, in cases where batch effects are purely technical, higher scGraph scores indicate better correction. Thus, scGraph should be used alongside other metrics to ensure biologically meaningful evaluation of integration quality.

Reporting summary

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

Data availability

All data in this study are publicly available. Statistics, resources and corresponding studies are listed in Extended Data Table 1.

Code availability

The implementation code for Islander, as well as tutorial notebooks to reproduce the results in this paper, can be accessed from GitHub (https://github.com/Genentech/Islander). The standalone scgraph evaluation toolkit can be installed using pip (https://pypi.org/project/scgraph-eval/). For scIB evaluation pipelines, the implementations by Gayso et al. were obtained from GitHub (https://github.com/yoseflab/scib-metrics).

References

37. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.* **19**, 15 (2018).
38. van der Maaten, L. & Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **9**, 2579–2605 (2008).
39. Becht, E. et al. Dimensionality reduction for visualizing single-cell data using UMAP. *Nat. Biotechnol.* **37**, 38–44 (2019).
40. Gayoso, A. et al. A Python library for probabilistic analysis of single-cell omics data. *Nat. Biotechnol.* **40**, 163–166 (2022).
41. Su, Y. et al. Multi-omics resolves a sharp disease-state shift between mild and moderate COVID-19. *Cell* **183**, 1479–1495.e20 (2020).
42. Luecken, M. et al. Benchmarking atlas-level data integration in single-cell genomics—integration task datasets. *figshare* https://doi.org/10.6084/m9.figshare.12420968 (2022).

Acknowledgements

We thank R. Lopez, R. Sosic, P. He, M. Bereket, L. Dony, S.-J. Dunn, G. Eraslan, A. Gayoso, G. Heimberg, K. Huang, J. Marioni, D. Pe'er, L. Peng, Y. Roohani, Y. Rosen, A. Whitehead and J. Zhang for invaluable insights, along with all the members from the J.L. and A.R. labs and colleagues at the Human Cell Atlas, Chan Zuckerberg Initiative and Google DeepMind, for constructive and insightful discussions. J.L. was supported by the National Science Foundation through grants OAC-1835598 (CINES), CCF-1918940 (Expeditions) and DMS-2327709 (IHBEM), the Stanford Data Applications Initiative, the Wu Tsai Neurosciences Institute, the Stanford Institute for Human-Centered Artificial Intelligence, the Chan Zuckerberg Initiative, Amazon, Genentech, GSK, Hitachi, SAP and UCB.

Author contributions

H.W. and A.R. conceptualized the study. H.W. performed the experiments. H.W., J.L. and A.R. wrote the paper.

Competing interests

H.W. and A.R. are employees of Genentech, a member of the Roche Group. A.R. has equity in Roche. A.R. is a cofounder and equity holder of Celsius Therapeutics and is an equity holder in Immunitas. Until 31 July 2020, A.R. was a scientific advisory board member of Thermo Fisher Scientific, Syros Pharmaceuticals, Neogene Therapeutics and Asimov. A.R. is a named inventor on multiple filed patents related to single-cell and spatial genomics, including for scRNA-seq, spatial transcriptomics, Perturb-Seq, compressed experiments and PerturbView.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41587-025-02702-z.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41587-025-02702-z.

Correspondence and requests for materials should be addressed to Jure Leskovec or Aviv Regev.

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

 https://doi.org/10.1038/s41587-025-02702-z

Extended Data Table 1 | Statistics of cell atlases

Atlas	# Gene	# Cell	# Class	# Batch	Reference
Brain	59,357	888,263	11	4	Paper[17], Data source[17]
Breast	33,234	703,512	39	126	Paper[18], Data source[18]
COVID	33,537	559,517	31	10	Paper[41], Data source[41]
Eye	36,484	51,645	11	8	Paper[19], Data source[19]
Gut (Fetal)	26,328	62,849	21	9	Paper[20], Data source[20]
Heart	33,234	486,134	27	14	Paper[21], Data source[21]
Lung	28,024	584,444	53	166	Paper[35], Data source[35]
Lung (Fetal, Donor)	26,354	71,752	144	29	Paper[22], Data source[22]
Lung (Fetal, Organoid)	24,653	70,495	28	37	Paper[22], Data source[22]
Pancreas	19,093	16,382	14	9	Paper[11], Data source[42]
Skin	30,933	15,457	13	5	Paper[23], Data source[23]

This table provides an overview of each dataset used in the study. For each dataset, we report the total number of unique genes ("# Gene"), cells ("# Cell"), cell types ("# Class"), and batches ("# Batch"). Additionally, the table includes links to relevant literature and dataset associated with each atlas.

 https://doi.org/10.1038/s41587-025-02702-z

Extended Data Table 2 | Benchmarking cell embeddings using scIB with default annotations for 144 cell types on the Human Fetal Lung Cell Atlas, the donor split

Method		Bio conservation							Batch correction					Aggregate score		
	HVG	I-label	L-NMI	L-ARI	K-NMI	K-ARI	S-label	cLISI	S-batch	iLISI	KBET	G-Con	PCR	Batch	Bio	Total
PCA		0.571	0.809	0.581	0.723	0.237	0.532	**1.000**	0.834	0.043	0.620	0.821	0.000	0.464	0.636	0.567
PCA	✓	0.581	0.799	0.619	0.716	0.191	0.535	0.999	**0.868**	0.059	0.640	0.797	0.000	0.473	0.634	0.570
TSNE		0.583	0.762	0.318	0.720	0.160	0.499	**1.000**	0.542	0.042	0.480	0.663	0.000	0.345	0.577	0.484
TSNE	✓	0.585	0.767	0.350	0.716	0.157	0.504	**1.000**	0.568	0.059	0.509	0.693	0.000	0.366	0.582	0.496
UMAP		0.580	0.765	0.387	0.713	0.162	0.489	0.999	0.576	0.068	0.572	0.708	0.000	0.385	0.585	0.505
UMAP	✓	0.548	0.771	0.404	0.713	0.162	0.524	0.999	0.586	0.085	0.560	0.707	0.000	0.388	0.589	0.508
Harmony		0.544	0.720	0.339	0.648	0.203	0.490	0.929	0.831	0.144	0.808	0.774	0.291	0.570	0.553	0.560
Harmony	✓	0.487	0.641	0.334	0.554	0.115	0.488	0.927	0.831	0.150	0.675	0.671	0.597	0.585	0.507	0.538
Scanorama		0.487	0.821	0.692	0.723	0.236	0.538	**1.000**	0.860	0.096	0.739	0.818	0.000	0.503	0.643	0.587
Scanorama	✓	0.564	0.816	0.703	0.725	0.237	0.536	**1.000**	0.865	0.091	0.744	0.821	0.000	0.504	0.654	0.594
BBKNN		0.413	0.753	0.340	0.703	0.150	0.541	0.927	0.590	0.160	0.785	0.725	0.000	0.452	0.547	0.509
BBKNN	✓	0.573	0.754	0.392	0.697	0.158	0.487	0.931	0.580	0.133	0.669	0.649	0.000	0.406	0.571	0.505
fastMNN	✓	0.415	0.251	0.060	0.228	0.040	0.397	0.984	0.778	0.162	0.101	0.067	0.616	0.345	0.339	0.341
scVI		0.552	0.709	0.369	0.636	0.136	0.522	0.927	0.834	0.139	0.840	0.860	0.415	0.618	0.550	0.577
scVI	✓	0.606	0.724	0.407	0.663	0.142	0.521	0.923	0.838	0.142	0.818	0.849	**0.681**	**0.666**	0.569	0.608
scANVI		0.532	0.785	0.559	0.682	0.174	0.540	**1.000**	0.818	0.137	0.850	0.862	0.154	0.564	0.610	0.592
scANVI	✓	0.597	0.856	0.738	0.736	0.232	0.554	**1.000**	0.829	0.121	0.834	0.861	0.521	0.633	0.673	0.657
scGen	✓	0.603	0.902	0.756	0.789	0.285	0.609	0.931	0.695	0.144	0.846	0.906	0.138	0.546	0.697	0.636
scPoli		0.462	0.876	0.663	0.802	0.290	0.624	**1.000**	0.745	0.145	0.869	0.903	0.000	0.532	0.674	0.617
scPoli	✓	0.661	0.879	0.700	0.802	0.313	0.629	**1.000**	0.739	0.143	0.867	0.899	0.230	0.575	0.712	0.657
Geneformer		0.492	0.640	0.304	0.520	0.107	0.475	0.996	0.829	0.114	0.672	0.624	0.410	0.530	0.505	0.515
scGPT	✓	0.486	0.583	0.225	0.467	0.063	0.445	0.991	0.770	**0.165**	0.636	0.549	0.482	0.521	0.466	0.488
scGPT (FT)	✓	0.517	0.717	0.342	0.658	0.156	0.518	0.998	0.762	0.130	0.819	0.826	0.585	0.624	0.558	0.584
Author's		0.575	0.844	0.561	0.774	0.347	0.567	**1.000**	0.834	0.070	0.780	0.897	0.000	0.516	0.667	0.607
Islander (Tri)		0.624	0.923	0.932	0.822	0.315	0.724	**1.000**	0.815	0.114	0.825	0.838	0.000	0.518	0.763	0.665
Islander (SCL)		0.625	0.854	0.380	0.852	0.399	0.785	**1.000**	0.748	0.145	0.792	0.822	0.000	0.501	0.699	0.620
Islander (Run1)		0.818	**0.999**	**1.000**	**0.901**	**0.449**	0.793	**1.000**	0.854	0.124	**0.889**	**0.972**	0.240	0.616	**0.851**	**0.757**
Islander (Run2)		**0.824**	**0.999**	**1.000**	0.891	0.406	0.793	**1.000**	0.853	0.123	0.883	0.970	0.217	0.609	0.845	0.751
Islander (Run3)		0.817	**0.999**	**1.000**	0.894	0.440	**0.794**	**1.000**	0.854	0.123	0.888	0.970	0.249	0.617	0.849	0.756

The highest scores for each metric are highlighted in **bold**. All subsequent tables adhere to the same annotation scheme.

 https://doi.org/10.1038/s41587-025-02702-z

Extended Data Table 3 | Benchmarking cell embeddings using the scIB framework with a broad annotation of 14 cell types on the Human Fetal Lung Atlas

Method	Bio conservation							Batch correction					Aggregate score		
	I-label	L-NMI	L-ARI	K-NMI	K-ARI	S-label	cLISI	S-batch	iLISI	KBET	G-Con	PCR	Batch	Bio	Total
PCA	0.575	0.770	0.480	0.743	0.453	0.594	**1.000**	0.849	0.043	0.258	0.872	0.000	0.404	0.659	0.557
TSNE	0.291	0.594	0.099	0.636	0.270	0.510	**1.000**	0.625	0.042	0.205	0.633	0.000	0.301	0.486	0.412
UMAP	0.399	0.650	0.167	0.684	0.327	0.531	**1.000**	0.627	0.068	0.305	0.798	0.000	0.360	0.537	0.466
Harmony	0.588	0.783	0.556	0.778	0.695	0.613	**1.000**	0.748	0.142	**0.615**	0.794	0.603	0.581	0.716	0.662
Scanorama	0.569	0.790	0.477	0.732	0.464	0.589	**1.000**	**0.869**	0.083	0.396	**0.942**	0.132	0.484	0.660	0.590
BBKNN	**0.901**	0.682	0.186	0.756	0.436	0.569	**1.000**	0.570	**0.155**	0.513	0.895	0.079	0.442	0.647	0.565
scVI	0.620	**0.886**	**0.896**	0.780	0.567	0.571	**1.000**	0.851	0.133	0.478	0.934	**0.668**	**0.613**	**0.760**	**0.701**
scANVI	0.639	0.848	0.618	0.778	0.581	0.606	**1.000**	0.830	0.123	0.470	0.909	0.527	0.572	0.724	0.663
scPoli	0.691	0.751	0.390	**0.882**	**0.841**	**0.733**	**1.000**	0.674	0.137	0.495	0.775	0.331	0.482	0.755	0.646
Geneformer	0.506	0.759	0.550	0.547	0.319	0.527	**1.000**	0.848	0.113	0.403	0.834	0.405	0.521	0.601	0.569
Author's	0.713	0.663	0.174	0.673	0.358	0.553	**1.000**	0.590	0.107	0.392	0.829	0.000	0.384	0.591	0.508
Islander (Tri)	0.614	0.758	0.377	0.679	0.362	0.600	**1.000**	0.766	0.114	0.434	0.823	0.000	0.427	0.627	0.547
Islander (SCL)	0.743	0.572	0.091	0.383	0.074	0.411	**1.000**	0.640	0.145	0.477	0.442	0.000	0.341	0.468	0.417
Islander (Run1)	0.650	0.686	0.321	0.586	0.329	0.582	**1.000**	0.748	0.124	0.474	0.365	0.240	0.390	0.593	0.512
Islander (Run2)	0.678	0.686	0.321	0.616	0.369	0.589	**1.000**	0.746	0.123	0.469	0.365	0.217	0.384	0.608	0.519
Islander (Run3)	0.695	0.687	0.321	0.616	0.365	0.585	**1.000**	0.747	0.123	0.485	0.364	0.249	0.394	0.610	0.523
Islander (UMAP1)	0.328	0.637	0.151	0.445	0.271	0.442	**1.000**	0.485	0.149	0.462	0.385	0.360	0.368	0.468	0.428
Islander (UMAP2)	0.700	0.635	0.147	0.430	0.294	0.515	**1.000**	0.475	0.149	0.461	0.375	0.538	0.400	0.532	0.479
Islander (UMAP3)	0.354	0.634	0.147	0.405	0.259	0.495	**1.000**	0.494	0.152	0.473	0.375	0.312	0.361	0.471	0.427

https://doi.org/10.1038/s41587-025-02702-z

Extended Data Table 4 | Benchmarking cell embeddings, using scGraph

Method	HVG	Brain	Breast	COVID	Eye	Gut (F)	Heart	Lung (F,D)	Lung (F,O)	Lung	Pancreas	Skin
Harmony		0.168	**0.739**	0.770	0.405	0.538	**0.763**	0.511	0.284	0.700	**0.520**	0.465
Harmony	✓	**0.427**	0.736	**0.804**	0.515	0.696	0.552	0.570	0.356	**0.781**	0.431	**0.694**
Scanorama		0.239	0.645	0.776	0.522	0.706	0.628	0.594	0.263	0.351	0.439	0.559
Scanorama	✓	0.250	0.694	0.760	**0.534**	0.635	0.554	0.622	0.201	0.309	0.291	0.465
BBKNN		0.091	0.644	0.775	0.524	0.596	0.684	0.579	0.314	0.685	0.563	0.626
BBKNN	✓	0.166	0.658	0.771	0.456	0.736	0.627	**0.693**	0.550	0.689	0.445	0.690
scVI		0.065	0.632	0.719	0.393	0.650	0.316	0.493	0.478	0.704	0.378	0.387
scVI	✓	0.254	0.690	0.752	0.314	0.649	0.588	0.499	0.453	0.674	0.506	0.567
scANVI		0.116	0.647	0.757	0.408	0.626	0.350	0.567	0.552	0.672	0.390	0.386
scANVI	✓	0.396	0.735	0.763	0.517	0.600	0.569	0.585	0.509	0.678	0.394	0.436
scGen	✓	0.217	0.600	0.779	0.436	0.526	0.606	0.331	0.161	0.337	0.354	0.692
scPoli		-	-	0.573	0.431	0.588	0.679	0.394	0.572	0.588	0.311	0.462
scPoli	✓	0.295	0.519	0.672	0.360	0.594	0.706	0.455	**0.590**	0.518	0.401	0.422
Geneformer		-	-	-	0.524	**0.747**	0.449	0.604	0.265	0.479	-	0.540
scGPT	✓	-	-	0.535	0.256	0.447	0.487	0.552	0.388	0.390	-	0.378
Author's		0.295	0.689	-	0.284	0.641	0.702	0.500	-	0.640	-	0.472
Islander		-0.071	-0.032	0.361	-0.335	0.098	0.013	-0.011	0.022	-0.061	0.234	-0.093

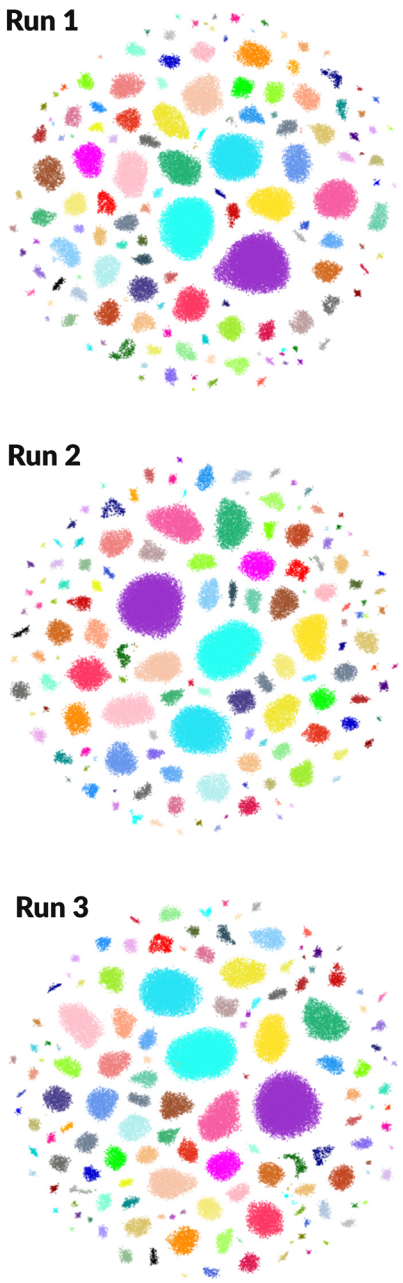
"F", "D" and "O" represents fetal, donor and organoid, respectively. "-" means the embeddings are not available, due to memory limitations (>500 G in RAM) or unavailability of raw counts or ensembl ids (used in Geneformer and scGPT). We **bold** the highest and underline the lowest scores for each dataset.

 https://doi.org/10.1038/s41587-025-02702-z

Extended Data Table 5 | Benchmarking cell embeddings using scIB and scGraph with default annotations for 9 cell subtypes of fibroblasts, applied to the fibroblast subset of the Human Fetal Lung Cell Atlas. All methods are re-trained on this subset

Method	Bio conservation							Batch correction					scIB aggregate score			scGraph -
	I-label	L-NMI	L-ARI	K-NMI	K-ARI	S-label	cLISI	S-batch	iLISI	KBET	G-Con	PCR	Batch	Bio	Total	
PCA	0.584	0.707	0.618	0.638	0.581	0.554	0.990	0.910	0.004	0.189	0.952	0.000	0.411	0.667	0.565	**0.452**
TSNE	0.653	0.510	0.160	0.567	0.445	0.597	1.000	0.641	0.021	0.182	0.748	0.000	0.318	0.562	0.464	0.372
UMAP	0.757	0.542	0.203	0.635	0.490	0.633	1.000	0.663	0.035	0.262	0.904	0.000	0.373	0.609	0.514	0.408
Harmony	0.452	0.232	0.154	0.185	0.101	0.491	0.815	0.863	0.139	0.770	0.820	0.887	**0.696**	0.347	0.487	0.396
Scanorama	0.542	0.670	0.586	0.535	0.471	0.539	0.994	**0.914**	0.066	0.517	0.920	0.123	0.508	0.620	0.575	0.352
BBKNN	0.659	0.470	0.178	0.527	0.408	0.580	0.994	0.669	0.076	0.388	0.897	0.000	0.406	0.545	0.489	0.075
scVI	0.493	0.251	0.171	0.265	0.180	0.503	0.811	0.860	0.145	0.677	0.813	**0.936**	0.686	0.382	0.504	0.339
scANVI	0.541	0.846	0.824	0.663	0.629	0.549	0.998	0.858	0.082	0.701	0.926	0.636	0.640	0.721	0.689	0.356
scPoli	0.569	0.440	0.245	0.475	0.332	0.549	0.951	0.800	0.108	0.623	**0.970**	0.521	0.604	0.509	0.547	0.273
Geneformer	0.486	0.340	0.149	0.121	0.082	0.504	0.940	0.892	0.081	0.542	0.677	0.558	0.550	0.375	0.445	0.378
scGPT	0.455	0.147	0.046	0.110	0.051	0.473	0.773	0.764	**0.161**	0.551	0.710	0.589	0.555	0.293	0.398	0.259
scGPT (FT)	0.487	0.338	0.138	0.309	0.195	0.520	0.948	0.739	0.104	0.615	0.921	0.723	0.621	0.419	0.500	0.283
Islander	**0.822**	**1.000**	**1.000**	**1.000**	**1.000**	**0.803**	**1.000**	0.896	0.073	**0.811**	0.943	0.000	0.545	**0.946**	**0.786**	-0.036
Author's	0.753	0.569	0.220	0.723	0.615	0.647	**1.000**	0.683	0.069	0.474	0.946	0.000	0.435	0.647	0.562	0.421

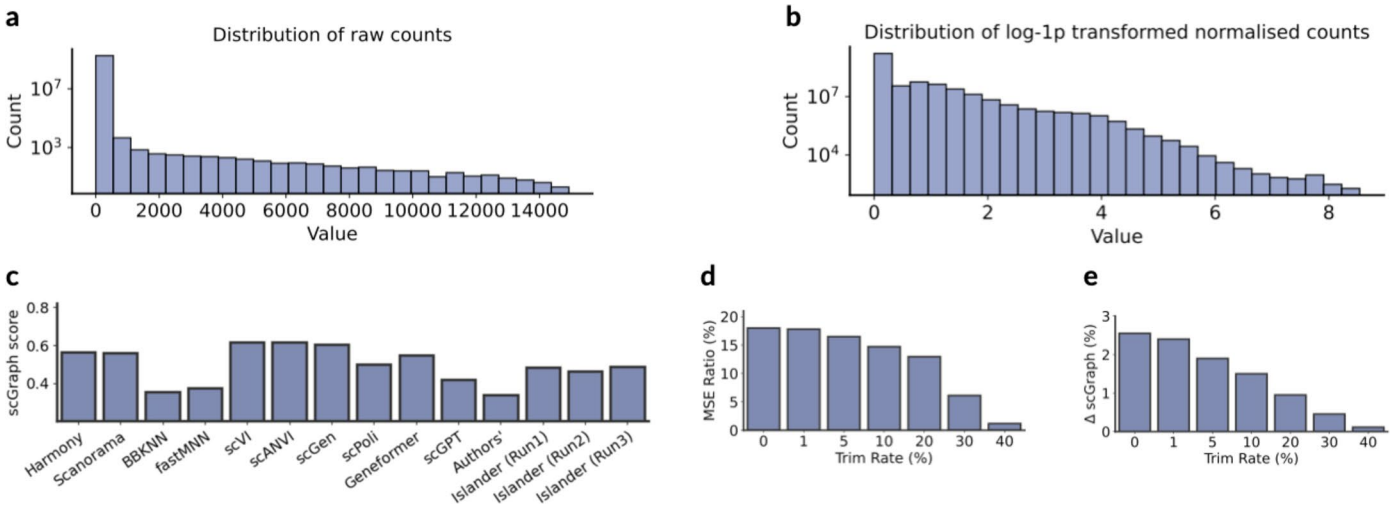
https://doi.org/10.1038/s41587-025-02702-z



ACTC+ SMC
APOE+ MΦ1
APOE+ MΦ2
ASPN+ chondrocyte
AT1
AT2
Activated NK
Adventitial fibro
Aerocyte
Airway fibro
Alveolar fibro
Arterial endo
Basophil
CD4 T
CD5+ CCL22+ mature B
CD5+ CCL22- mature B
CD5- Mature B
CD8 T
CD16+ NK
CD56bright NK
CMP
COL20A1+ Schwann
CX3CR1+ MΦ
CXCL9+ MΦ
Ciliated
Club
Cycling DC
Cycling NK
Cycling T
Cycling definitive erythroblast
DC1
DC2
DC3
Definitive erythroblast
Definitive erythrocyte
Definitive reticulocyte
Deuterosomal
Early Schwann
Early airway SMC 1
Early airway SMC 2
Early airway progenitor
Early cap
Early fibro
Early mesothelial
Early stalk
Early tip
Eosinophil
FGFBP2+ Neural progenitor
GHRL+ NE precursor
GHRL+ neuroendocrine
GMP
GRIA2+ arterial endo
HMOX1+ primitive erythroblast
HSC
HSC/ELP
ILC2
ILC3
ILCP
Immature B
Interm chondrocyte
Interm fibro
Interm neuroendocrine
Intermediate NK
Intermediate lymphatic endo
KCNIP4+ neuron
Large pre-B
Late Schwann
Late airway SMC
Late airway progenitor
Late basal
Late cap
Late mesothelial
Late pre-B
Late pro-B
Late stalk
Late tip
Lymphatic endo
MEP
MFNG+ DBH+ neuron
MUC5AC+ ASCL1+ progenitor
MUC16+ ciliated
MYL4+ SMC
Mast
Megakaryocyte
Mesenchymal 1
Mesenchymal 2
Mesenchymal 3
Mid Schwann
Mid airway progenitor
Mid basal
Mid cap
Mid fibro
Mid mesothelial
Mid stalk
Mid tip
Myelocyte-like
Myofibro 1
Myofibro 2
Myofibro 3
NKT1
NKT2
Neutrophil
Non-cla. mono.
OMD+ endo
PCP4+ neuron
Pericyte
Platelet
Primitive erythrocyte
Pro-B
Pro-B/Pre-B transition
Proliferating Schwann
Promonocyte-like
Promyelocyte-like
Proximal basal
Proximal secretory 1
Proximal secretory 2
Proximal secretory 3
Proximal secretory progenitors
Pulmonary NE precursor
Pulmonary neuroendocrine
Resting chondrocyte
S100A12-hi cla. mono.
S100A12-lo cla. mono.
SCG3+ lymphatic endothelial
SMG
SMG basal
SPP1+ MΦ
SST+ neuron
Schwann precursor
Squamous
TM4SF4+ CHODL+ neuron
TM4SF4+ PENK+ neuron
Th17
Treg
Tαβ_Entry
Vascular SMC 1
Vascular SMC 2
Venous endo
aDC 1
aDC 2
pDC
pre-pDC/DC5
κ small pre-B
λ small pre-B

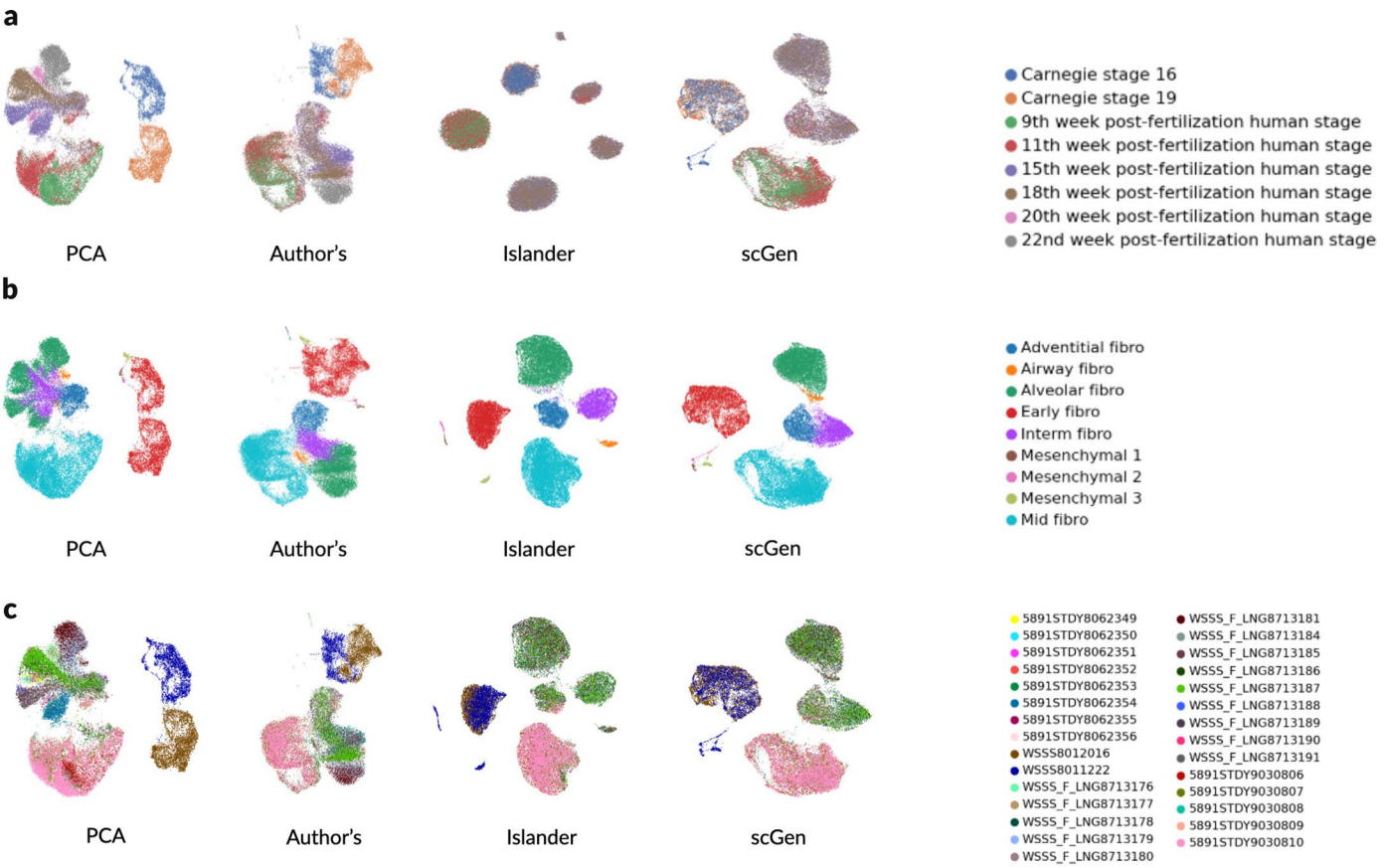
Extended Data Fig. 1 | Drifting Cell Islands, different runs of Islander on fetal lung atlas (donor).

https://doi.org/10.1038/s41587-025-02702-z



Extended Data Fig. 2 | Design optimization for scGraph using human fetal lung atlas[22]. a, b, Distribution of raw (**a**) and log1p-transformed (**b**) scRNA-seq counts. **c**, scGraph scores using log-1p counts do not effectively flag distortions caused by drifting cell islands. scGraph scores (y axis) for embeddings generated with each method (x axis) using log-1p counts. **d,e** Effect of trim rate on PCA centroid locations and scGraph scores. **d**, Normalized mean square error between centroids (MSE, y-axis) at different trimming rates (x-axis), with centroids at 49% trimming as reference. **e**, Percentage difference (y-axis) between scGraph scores at various trimming rates (x-axis) compared to the score at 49% trimming. While small trim rates lead to larger changes in centroid coordinates, the corresponding changes in scGraph scores are relatively minor. Based on these observations, we selected a trim rate of 5% per side (10% total).

https://doi.org/10.1038/s41587-025-02702-z



Extended Data Fig. 3 | Scoring human fetal lung fibroblast[22] embeddings by scIB and scGraph metrics. a-c, Embeddings of 31,020 human fetal lung fibroblast profiles from 9 fibroblast subtypes across 29 batches, generated by the top scoring methods based on scIB (scANVI and Is- lander) or scGraph (Harmony and Authors') and colored by developmental stage (**a**), cell types (**b**), or batch (**c**). Each method was trained on this subset and evaluated using both scIB and scGraph (Extended Data Table 5). **d-e**, Rankings of integration methods. scGraph (**d**, y axis) and scIB (**e**, y axis) scores for each of the 9 integration methods (x axis).

nature portfolio

Corresponding author(s): Jure Leskovec, Aviv Regev

Last updated by author(s): May 1, 2025

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
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☒	☐ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐ A description of all covariates tested
☒	☐ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ 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)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), 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	All used datasets are publicly available. We provide the code to easily download the data: https://github.com/Genentech/Islander/blob/main/scripts/_download_data.sh
Data analysis	Data analysis The analysis was performed in Python (version 3.9.18). All package versions used in the Python environment are listed in the environment YAML file: https://github.com/Genentech/Islander/blob/main/env.yml. Notably, the integration and benchmark suite includes: scib-metrics==0.4.1 (https://github.com/yoseflab/scib-metrics), pynndescent==0.5.10, pydantic==2.1.1, pydantic-core==2.4.0, pandas==2.1.3, numba==0.58.1, numpy==1.26.2, jax==0.4.20, jaxlib==0.4.20+cuda11.cudnn86, igraph==0.10.8, h5py==3.10.0, bbknn==1.6.0, anndata==0.10.3, scgen==2.1.1, scarches==0.5.9, scanorama==1.7.4, scanpy==1.9.6, harmony-pytorch==0.1.7, scvi-tools==1.0.4. Training Islander involves: wandb==0.16.0, scikit-learn==1.3.2, pyyaml==6.0.1, python-json-logger==2.0.7, python-multipart==0.0.6, pytorch-lightning==2.1.2, torch==2.1.1. Visualization modules include: umap-learn==0.5.5, matplotlib==3.8.2, matplotlib-inline==0.1.6, leidenalg==0.10.1, jupyter==1.0.0, seaborn==0.12.2.

Our own developed evaluation metric toolkit:
scgraph-eval==0.1.2 (https://github.com/Genentech/Islander).

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

All datasets used in this study are publicly available. Extended Data Table 1 and the README file in the Github (https://github.com/Genentech/Islander) contain the download links and relevant literature. All data was accessed on October 1, 2023

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	*Use the terms sex (biological attribute) and gender (shaped by social and cultural circumstances) carefully in order to avoid confusing both terms. Indicate if findings apply to only one sex or gender; describe whether sex and gender were considered in study design; whether sex and/or gender was determined based on self-reporting or assigned and methods used. Provide in the source data disaggregated sex and gender data, where this information has been collected, and if consent has been obtained for sharing of individual-level data; provide overall numbers in this Reporting Summary. Please state if this information has not been collected. Report sex- and gender-based analyses where performed, justify reasons for lack of sex- and gender-based analysis.*
Reporting on race, ethnicity, or other socially relevant groupings	*Please specify the socially constructed or socially relevant categorization variable(s) used in your manuscript and explain why they were used. Please note that such variables should not be used as proxies for other socially constructed/relevant variables (for example, race or ethnicity should not be used as a proxy for socioeconomic status). Provide clear definitions of the relevant terms used, how they were provided (by the participants/respondents, the researchers, or third parties), and the method(s) used to classify people into the different categories (e.g. self-report, census or administrative data, social media data, etc.) Please provide details about how you controlled for confounding variables in your analyses.*
Population characteristics	*Describe the covariate-relevant population characteristics of the human research participants (e.g. age, genotypic information, past and current diagnosis and treatment categories). If you filled out the behavioural & social sciences study design questions and have nothing to add here, write "See above."*
Recruitment	*Describe how participants were recruited. Outline any potential self-selection bias or other biases that may be present and how these are likely to impact results.*
Ethics oversight	*Identify the organization(s) that approved the study protocol.*

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

Life sciences study design

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

Sample size	We used full published datasets, so sample size was defined by the studies that generated the data and any relevant post-processing by the authors. We applied a uniform pre-processing protocol across the datasets. Cell profiles with fewer than 1,000 reads or less than 500 detected genes were filtered out, and genes present in fewer than five cells were also excluded. For computational experiments, we used three different random seeds to initialize the Islander models' weights and conducted three independent runs for each baseline method on each dataset. We reported the average performance of each method, as the variance across distinct runs was not significant.
Data exclusions	No data was excluded from the analyses.

Replication	All computational experiments are repeatable using the code provided. Each model training was independently replicated three times using different random seeds. Training typically takes 2–5 hours per run. Benchmarking with scIB takes approximately 0.5–5 hours per dataset, and scGraph benchmarking requires less than 10 minutes. Performance was averaged across runs, and variance was consistently low.
Randomization	Randomization was not required, as our study involved only computational analyses using publicly available datasets. All datasets were pre-collected and processed uniformly. Model training and evaluation were conducted using fixed pipelines, and variability was addressed through multiple random seed replications.
Blinding	Blinding was not required, as no new wet-lab experiments involving human or animal subjects were conducted. All analyses were computational and based on publicly available datasets with predefined labels and metadata. Model evaluation was automated and benchmarked using objective metrics.

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 & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ 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, mosiacism, off-target gene editing) were examined.*