nature methods

Article

https://doi.org/10.1038/s41592-024-02191-z

Toward universal cell embeddings: integrating single-cell RNA-seq datasets across species with SATURN

Received: 3 February 2023

Accepted: 22 January 2024

Published online: 16 February 2024

Check for updates

Yanay Rosen[1,5], Maria Brbić[2,5], Yusuf Roohani[3,5], Kyle Swanson[1], Ziang Li[4] & Jure Leskovec[1]

Analysis of single-cell datasets generated from diverse organisms offers unprecedented opportunities to unravel fundamental evolutionary processes of conservation and diversification of cell types. However, interspecies genomic differences limit the joint analysis of cross-species datasets to homologous genes. Here we present SATURN, a deep learning method for learning universal cell embeddings that encodes genes' biological properties using protein language models. By coupling protein embeddings from language models with RNA expression, SATURN integrates datasets profiled from different species regardless of their genomic similarity. SATURN can detect functionally related genes coexpressed across species, redefining differential expression for cross-species analysis. Applying SATURN to three species whole-organism atlases and frog and zebrafish embryogenesis datasets, we show that SATURN can effectively transfer annotations across species, even when they are evolutionarily remote. We also demonstrate that SATURN can be used to find potentially divergent gene functions between glaucoma-associated genes in humans and four other species.

Cell mapping consortia efforts have generated large-scale single-cell datasets comprising hundreds of thousands of cells with the goal of uncovering underlying cellular processes. In-depth analysis of diverse datasets generated across different species through global efforts such as the Human Cell Atlas[1,2], the Mouse Cell Atlas[3] and the Fly Cell Atlas[4,5] has broadened our understanding of cell biology characterizing many cell types for the first time. However, current analyses remain limited in their ability to jointly analyze datasets generated across different species. Such joint analysis offers great potential for understanding fundamental evolutionary processes such as identifying cell types that are conserved across species and identifying the corresponding gene programs that drive similarities and differences of such cell types.

A variety of linear[6,7] and, more recently, deep learning approaches[8–10] have been developed to learn low-dimensional representations of single-cell RNA expression data (cell embeddings). However, existing methods represent genes only as columns of an RNA expression matrix and thus do not account for the biological properties of genes. This severely limits their usability when analyzing datasets generated from different species in which only a subset of genes can be matched as one-to-one homologs. While sequence alignment methods have been explored to incorporate weighted relationships between genes across species[11], they are dependent on arbitrary alignment thresholds and do not capture remote homology. Recent advances in protein language models trained on hundreds of millions protein sequences[12–14] suggest strong potential in addressing these issues by

[1]Department of Computer Science, Stanford University, Stanford, CA, USA. [2]School of Computer and Communication Sciences, Swiss Federal Institute of Technology (EPFL), Lausanne, Switzerland. [3]Department of Biomedical Data Science, Stanford University, Stanford, CA, USA. [4]Department of Computer Science and Technology, Tsinghua University, Beijing, China. [5]These authors contributed equally: Yanay Rosen, Maria Brbić, Yusuf Roohani. e-mail: jure@cs.stanford.edu

learning informative representations of the proteins a gene encodes. This is evidenced through the remarkable ability of protein representations to encode protein structure, function, molecular properties[12] and homology[15]. However, so far, the representational power of these models has not been exploited to learn cell representations that capture functional similarity of genes.

We present SATURN (Species Alignment Through Unification of Rna and proteiNs), a deep learning approach that integrates cross-species single-cell RNA-sequencing (scRNA-seq) datasets by coupling gene expression with protein embeddings generated by large protein language models. SATURN introduces a concept of macrogenes defined as groups of genes that share similar protein embeddings. The strength of associations of genes to macrogenes is learned to reflect this similarity, thereby allowing functionally related genes as captured by the protein embeddings to group together.

SATURN is uniquely able to perform multispecies differential expression analysis revealing functionally related groups of genes coexpressed across species. By mapping single-cell datasets generated with different genes to a joint embedding space, SATURN takes important steps toward universal cell embeddings.

We apply these embeddings to diverse tasks such as integration of cross-species cell atlas datasets, discovery of species-specific cell types, reannotation and cross-species label transfer, as well as identification of protein differences across species. In particular, we apply SATURN to integrate Tabula Sapiens[2], Tabula Microcebus[16] and Tabula Muris[3] cell atlas datasets, creating a mammalian cell atlas of 335,000 cells across nine common tissues. We further apply SATURN to integrate frog and zebrafish embryogenesis datasets[17]. Our results show that SATURN successfully transfers annotations even across evolutionarily remote species and finds homologous and species-specific cell types, outperforming existing cross-species integration methods. Finally, we apply SATURN to reannotate the five species of the Cell Atlas of Human Trabecular Meshwork and Aqueous Outflow Structures (AH atlas)[18]. We find that SATURN identifies glaucoma-associated macrogenes that have potentially divergent functions across species.

Results

Overview of SATURN

The major challenge of cross-species integration is that different datasets have different genes that may not have common one-to-one homologs. Subsetting each species' set of genes to the common set of one-to-one homologs leads to losing a large portion of biologically relevant genes. Increasing the number of species exacerbates this problem, as a gene must have a homolog in each species to be considered for integration. SATURN overcomes this problem by using large protein language models to learn cell embeddings that encode the biological meaning of genes. SATURN maps cross-species datasets in the space of functionally related genes determined by protein embeddings. SATURN's use of protein language models allows it to represent functional similarities even between remotely homologous genes that are missed by integration methods that rely on sequence-based similarity[11].

In particular, SATURN integrates scRNA-seq datasets generated from different species with different genes by mapping them to a joint low-dimensional embedding space using gene expression and protein representations. SATURN takes as input: (i) scRNA-seq count data from one or multiple species, (ii) protein embeddings generated by a large protein embedding language model like ESM2 (ref. 14), and (iii) initial within-species cell annotations (from cell-type assignments if available or obtained by running a clustering algorithm). The language model takes a sequence of amino acids and produces a protein representation vector (Fig. 1a). Given gene expression and protein embeddings, SATURN learns an interpretable feature space shared between multiple species. We refer to this space as a macrogene space and it represents a joint space composed of genes inferred to be functionally related based on the similarity of their protein embeddings. The importance of a gene to a macrogene is defined by a neural network weight—the stronger the importance, the higher the value of the weight that connects the gene to the macrogene.

Given the shared macrogene expression space across different species, SATURN then learns to represent cells across multiple species as nonlinear combinations of macrogenes. The neural network in SATURN is first pretrained with an autoencoder with zero inflated negative binomial (ZINB) loss, regularized to reconstruct protein embedding similarities using gene-to-macrogene weights (Methods). Using the pretrained network as initialization, SATURN then learns a mapping of all cells to the shared embedding space with a weakly supervised metric learning objective. This allows SATURN to calibrate distances in the embedding space to reflect cell label similarity. In particular, the objective function in SATURN consists of two main components: (i) forcing different cells within the same dataset far apart using weak supervision; and (ii) forcing similar cells across datasets close to each other in an unsupervised manner (Methods). This objective enables SATURN to integrate cells across different species, while preserving cell-type information within each species' dataset.

SATURN creates multispecies cell atlases

We applied SATURN to integrate large-scale single-cell atlas datasets generated from human (Tabula Sapiens), mouse lemur (Tabula Microcebus) and mouse (Tabula Muris), creating the mammalian cell atlas of 335,000 cells (Fig. 1b and Supplementary Fig. 1a). We found that major cell types aligned well across three species such as T cells, B cells and muscle cells, and then we analyzed the alignment on a per-tissue level. For example, in muscle, we found a small subcluster of cells labeled as mouse macrophages that grouped with human and lemur granulocytes, while the rest of cells labeled as mouse macrophages aligned with human and lemur macrophages (Extended Data Figs. 1 and 2). To investigate whether this alignment is indeed correct, we checked the expression of known granulocyte marker *Cd55* (refs. 19,20) and known macrophage marker *Cd74* (refs. 19,20). Interestingly, we found that this small subcluster labeled as mouse macrophages indeed expresses *Cd55* and does not express *Cd74*, indicating that this small cluster was wrongly annotated as macrophages, while it should be annotated as granulocytes (Extended Data Fig. 2).

In spleen, SATURN separated out human naive B cells from human memory B cells, but aligned human memory B cells with cells annotated as B cells in mouse and lemur (Extended Data Figs. 3 and 4). To investigate whether this alignment is meaningful, we checked the marker genes and found that indeed mouse and lemur B cells express *Cd19*, a B cell marker known to be preferentially expressed in memory B cells, which was only weakly expressed in human naive B cells (Extended Data Fig. 4)[21]. This indicates that mouse and lemur B cells are correctly clustered with human memory B cells, which is additionally confirmed by strong expression of *Cd19*. Thus, SATURN can be used to obtain fine-grained-level annotations when cell atlases have been annotated with different granularity levels. Additionally, we found that SATURN correctly identified cell types specific to a single species within the integrated datasets. For instance, in muscle tissue, SATURN separated human epithelial and mesothelial cells from all other cell types (Extended Data Fig. 1). These cell types are indeed absent in mouse and lemur datasets. In spleen, SATURN separated human erythrocytes (Extended Data Fig. 3).

We next applied SATURN to a multispecies dataset of frog (97,000 cells) and zebrafish (63,000 cells) embryogenesis[17]. SATURN aligned evolutionarily related cell types between these two remote species (Fig. 1c and Supplementary Fig. 1b). We further inspected small clusters that are aligned by SATURN, but their ground-truth cell-type annotations differ. We find that these clusters indeed correspond to related cell types. For example, SATURN integrated zebrafish early-stage macrophages and frog myeloid progenitors, which can differentiate into

https://doi.org/10.1038/s41592-024-02191-z

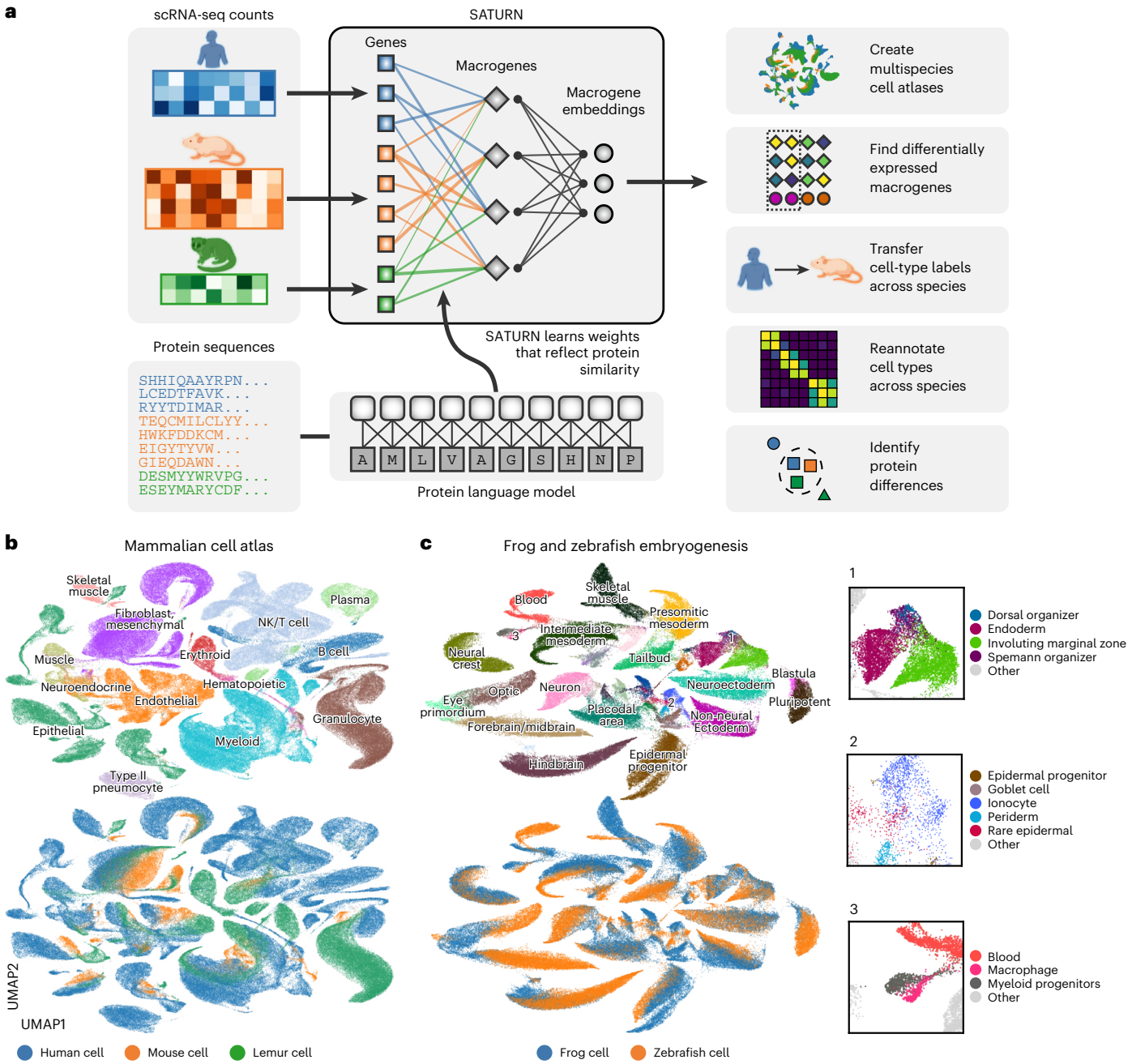


Fig. 1 | SATURN incorporates protein sequences and gene expression to embed single cells. a, Overview of SATURN. SATURN takes as input scRNA-seq datasets generated from one or more species and the amino acid sequences of proteins present in these species. SATURN then maps each species' genes to a joint feature space by learning 'macrogenes', that is, groups of functionally related intraspecies and interspecies genes. Finally, in the shared macrogene space, SATURN integrates datasets across species by learning a joint cell embedding space in which cell types conserved across species are aligned with each other. **b**, UMAP visualization of a joint embedding space across three distinct species. We applied SATURN to integrate cell atlas datasets of 335,000 cells from Tabula Sapiens (human), Tabula Microcebus (mouse lemur) and Tabula Muris (mouse), creating a mammalian cell atlas. Colors denote coarse-grained cell-type annotations (top) and species annotations (bottom). Only cell types with more than 350 cells were included. **c**, UMAP visualization of SATURN's integration of datasets from frog (97,000 cells) and zebrafish (63,000 cells) embryogenesis. Colors denote different major cell types (top) and different species (bottom). In SATURN's embedding space, cell types conserved across species aligned well (for example, frog/zebrafish neural crest), while species-specific cell types formed separate single-species clusters (for example, frog goblet cells). Cell types not directly mapped between both species shared similar ontology, for example, the zebrafish dorsal organizer and frog Spemann organizer (inset 1). Epidermal cell types including periderm, epidermal progenitor and rare epidermal cell types were also aligned, as were specialized epithelial cells such as goblet cells and ionocytes (inset 2). Finally, myeloid cell types including macrophages and myeloid progenitors clustered together (inset 3).

macrophages. Terminal differentiation in both cell types involves activation of a number of conserved master regulatory genes, such as *Cybb*, *Cyba*, *Spib* and *Cepba*[17]. These cell types are embedded close to blood cells, which further demonstrates that local distances in SATURN's embedding space are meaningful.

SATURN performs differential expression on macrogenes

SATURN extends differential expression analysis to a multispecies setting. Instead of performing differential expression analysis on individual genes, which is highly limited when datasets do not share genes, SATURN performs differential expression on macrogenes, which enables

characterization of cell-type-specific macrogenes across different datasets. To perform differential expression on macrogenes, SATURN first aggregates the contributions of individual genes to macrogenes using gene–macrogene neural network weights (Fig. 2a). The aggregated values can be seen as macrogene expression for each individual cell. Like in conventional differential expression analysis, SATURN then performs differential expression on cell clusters, such as those determined by cell-type label. The difference compared to conventional differential expression is that in SATURN the statistical test is performed on the macrogenes. Finally, to interpret the biological meaning of a macrogene, SATURN considers genes with the highest weight to the macrogene. We note that mean expression of a gene does not affect its macrogene weight. In particular, in the frog and zebrafish embryogenesis datasets, the correlation between a gene's expression and its maximum weight is 0.08 and 0.05 in the frog and zebrafish datasets, respectively.

By performing macrogene differential expression SATURN has two major advantages over existing integration methods. First, SATURN can identify differentially expressed genes that lack a one-to-one homolog. This is in contrast to existing methods that rely on one-to-one homologs and, therefore, ignore unmapped genes. Second, differentially expressed macrogenes provide natural gene modules that aid in interpretation, as they rely on groups of related genes instead of individual genes. This can lead to the identification of shared gene programs across species.

We conduct macrogene differential expression analysis in frog and zebrafish embryogenesis datasets. We demonstrate examples for the macrophage/myeloid progenitor cluster (Fig. 2b) and for the ionocytes cluster (Fig. 2b). In particular, we show the top five differentially expressed macrogenes and their corresponding highly weighted genes that characterize them, and we name each macrogene according to the gene with the highest weight to that macrogene. We focus on genes with known annotations. Gene-to-macrogene weights are listed in Supplementary Table 1.

For both macrophage/myeloid progenitors and ionocyte cell types, we find that highly expressed macrogenes indeed capture groups of related genes that are known to have the function associated with these cell types. In particular, for macrophage/myeloid progenitors, the top differentially expressed macrogenes include *Arhgdi*, *Cebp*, *Ptp*, *Cybb* and *Lcp1* (Fig. 2b). All these macrogenes contain genes associated with functions in blood cells. For example, the *Arhgdi* macrogene contains frog and zebrafish homologs of *Arhgdig*, as well as frog-specific paralogs such as *Arhgdib* and *Arhgdia*, which encode proteins involved in Rho protein signal transduction and RacGTPase binding activity[22,23]. RhoGTPases play an important role in hematopoietic stem cell functions[24]. Similarly, the *Cebp* macrogene contains frog and zebrafish homologs of *Cebpd*, *Cebpb* and *Cebpa*. *Cebpa* is associated with zebrafish hemopoiesis, and *Cebpb* is known to be expressed in zebrafish macrophages[22,23].

For ionocytes, SATURN ranks *Foxi*, *Dmrt2*, *Cldn*, *Ubp1* and *Atp6v0* as the top five differentially expressed macrogenes (Fig. 2b). Indeed, we find that all these macrogenes contain genes that are known to be associated with ionocytes. *Foxi* consists of Fox transcription factors that are known ionocyte markers[25]. The *Dmrt2* macrogene contains *Dmrt2* and *Dmrt2a*. *Dmrt2* is a known ionocyte marker in human pulmonary ionocytes[26]. The *Cldn* macrogene contains various claudins, which are found in gill ionocytes of teleost fish like zebrafish[27]. SATURN's identification of a claudin marker macrogene for ionocytes is notable because the set of genes that can be mapped as one-to-one homologs does not contain any of these genes. Additionally, claudins that can be mapped as one-to-one homologs (*Cldn1*, *Cldn12*, *Cldn18*, *Cldn19* and *Cldn2*) are not differentially expressed within the top 200 differentially expressed ionocyte genes in the individual datasets, nor in the shared one-to-one homolog space.

Moreover, macrogene differential expression can also be used to find species-level differences between cell types conserved across species. For example, when comparing zebrafish and frog ionocytes, a macrogene represented by *Gnpda1*, *Apip* and *Paics* and a macrogene represented by *Ppp1r14b* and *Fosab* are specific to zebrafish, while a macrogene represented by *Gadd45g*, *Aen*, and *Msgn1* is highly expressed in frog ionocytes but not in zebrafish (Fig. 2c). To analyze the proportion of macrogenes in a single species versus the proportion of shared macrogenes accross species, we found the top 20 differentially expressed macrogenes and then calculated the proportion of macrogenes that only had weights above 0.5 to genes in one species. Across all cell types, 35% of macrogenes were represented by genes in a single species.

Macrogenes capture homology

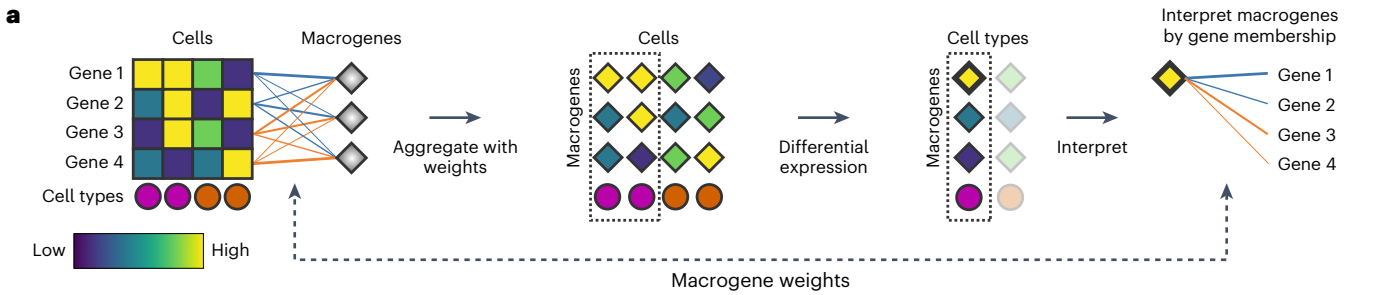
We find that macrogenes generated by SATURN recapture sequence-based gene homologs. In particular, we computed the proportion of macrogenes with a homologous gene pair between zebrafish and frog among their top-ranked genes. To assess gene homology, we use BLASTP, which determines gene homologs based on protein amino acid sequence similarity[28]. We find that even with only the top-ranked genes of each species, 56% of macrogenes in SATURN recapture gene homology information, while by considering ten top-ranked genes from each species, 91.2% of macrogenes recapture gene homology information (Fig. 2d). In comparison, random assignment of genes to macrogenes results in homologous pairs in only 0.25% of macrogenes when considering two top-ranked genes and in only 18.8% macrogenes when considering ten top-ranked genes. Altogether, these results indicate not only that macrogenes in SATURN recapture homology information, but also that they can also be used to reveal functional similarities between genes even when these genes are not considered as homologs by sequence-based similarity tools such as BLASTP. To further demonstrate that macrogenes capture functional similarities of genes, we performed Gene Ontology (GO)[29] analysis between the human and mouse genes in the mammalian cell atlas datasets. The analysis revealed significantly enriched GO terms within the gene sets of the same macrogene (Supplementary Note 5).

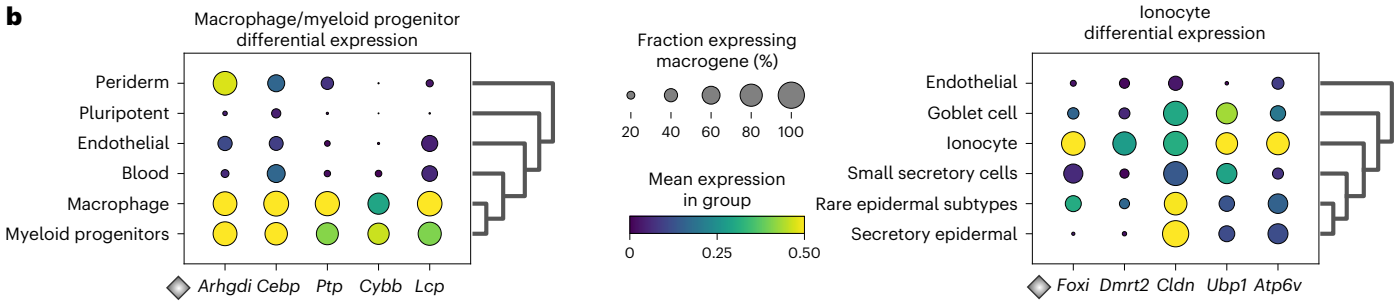
SATURN outperforms other methods by a large margin

We quantitatively assess the performance of SATURN on the alignment of frog and zebrafish embryogenesis datasets. We evaluate performance by measuring how well labels can be transferred from zebrafish to frog. In particular, we first integrated the datasets using SATURN and then used the cell-type annotations of cells from a reference species, zebrafish, to train a logistic classifier to predict cell types[30] (Supplementary Note 3). The classifier's performance was then tested on the embeddings of the query species, frog (Fig. 3a). Predictions are assessed as correct if they match the known frog cell type, based on a predetermined mapping of cell types between species (Supplementary Table 2). Because not all frog cells can be mapped to zebrafish cells, the maximum possible accuracy of such a model is 93%.

We compare the performance of SATURN to another single-cell multispecies integration method, SAMap[11], and unsupervised integration methods Harmony[6], scVI[8] and Scanorama[7]. SAMap is run in a weakly supervised mode in which cell neighborhoods are determined by cell type, which involves using the prior cell-type label information within each species but not across species, which is the same setting we used for running SATURN. SAMap is initialized with a gene graph based on protein sequence similarity as determined by BLASTP. For the unsupervised methods, the input genes for each species are taken as the one-to-one homologs as determined by ENSEMBL[31]. We found that SATURN achieves 85.8% median accuracy in cell label transfer from zebrafish to frog, achieving remarkable 119% performance gain over the next best-performing method, SAMap (Fig. 3b). We obtained similar performance gains when transferring labels from frog to zebrafish (Extended Data Fig. 5). Performance gains of SATURN are retained using other evaluation metrics, such as F1-score, precision and recall

https://doi.org/10.1038/s41592-024-02191-z





Macrogene	Genes
Arhgdi	*Arhgdib*, *Arhgdig*, *Arhgdia*
Cebp	*Cebpd*, *Cebpb*, *Cebpa*
Ptp	*Ptprc*, *Iqcd*, *Ptpn6*, *Ptpreb*, *Ptpn22*
Cybb	*Cybb*, *Nox4*, *Nox1*
Lcp1	*Lcp1*, *Parvg*, *Parvb*, *Parva*, *Parvab*

Macrogene	Genes
Foxi	*Foxi1*, *Foxi3a*, *Foxg1a*, *Foxg1*
Dmrt2	*Dmrt2*, *Dmrt2a*, *Kank1*
Cldn	*Cldna*, *Cldnh*, *Cldn4*, *Cldnb*, *Cldne*
Ubp1	*Ubp1*, *Grhl3*, *Grhl1*, *Grlh2a*, *Tp63*
Atp6v0	*Atp6v0c*, *Atp6v0ca*, *Atp6v0b*

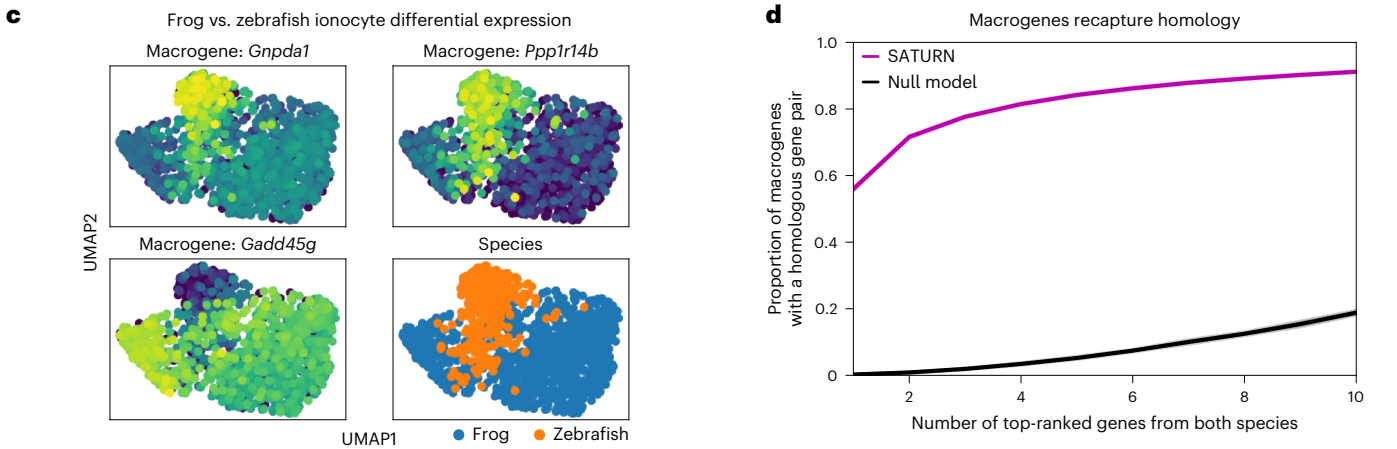


Fig. 2 | SATURN enables multispecies differential expression analysis in the macrogene space. a, Overview of SATURN's differential expression analysis on macrogenes. Every gene is connected to a macrogene with a corresponding weight that represents the importance of that gene to the given macrogene. Thus, each cell has corresponding macrogene values calculated as the weighted and normalized sum of its gene expression values. Because SATURN operates in the macrogene space, differential expression for resulting cell clusters gives the set of differentially expressed macrogenes of a given cell type. Finally, the genes with the highest weights to a macrogene are used to interpret the macrogene. **b**, Differentially expressed macrogenes on frog and zebrafish embryogenesis datasets for macrophage and myeloid progenitors (left) and ionocytes (right). Differential expression is performed by comparing these cell types with all other cell types. We show only cell types that are similar to target cell types determined as expressing a subset of the top differentially expressed macrogenes. We assigned names to macrogenes based on the set of genes with the highest weight in the given macrogene. The tables show the top five differentially expressed macrogenes and the top weighted genes in each macrogene. Genes are shown in black if a gene is included in the top genes for both species in a given macrogene, and blue or orange if the gene is frog or zebrafish specific, respectively. **c**, Macrogene differential expression can also be used to find species-level differences between cell types conserved across species. Example of differentially expressed macrogenes between frog and zebrafish ionocytes. **d**, SATURN macrogenes contained a far higher proportion of homolog gene pairs than what would be expected by chance, demonstrating that SATURN recaptures sequence-based homology. The purple curve shows the proportion of SATURN macrogenes that contain, within their top-ranked frog and top-ranked zebrafish genes, at least one homolog gene pair, versus the top number of genes. Homology was determined according to BLASTP results. The black curve shows the proportion obtained by a null model in which the same number of genes are randomly selected without replacement from both species.

(Extended Data Fig. 6), as well as data integration metrics[32] (Extended Data Fig. 7). We additionally visualized embeddings obtained by using the dimensionality reduction techniques principal component analysis and uniform manifold approximation and projection (UMAP)[33] on the one-to-one homolog expression space, demonstrating the gap between the species (Supplementary Fig. 2).

To test whether choice of protein language model for obtaining protein embeddings affects SATURN's performance, we compared

https://doi.org/10.1038/s41592-024-02191-z

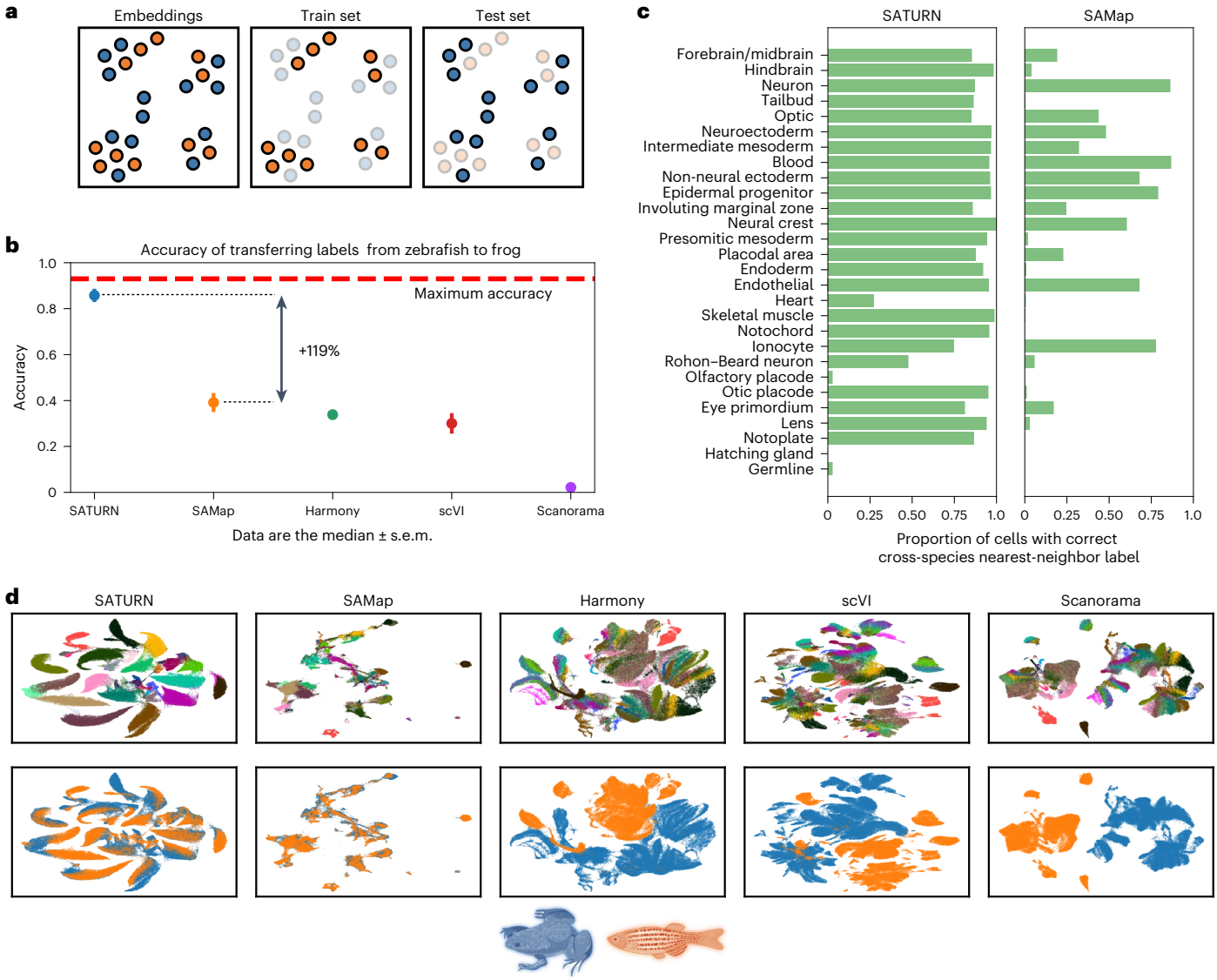


Fig. 3 | SATURN embeddings capture shared cell-type identity in frog and zebrafish embryogenesis. a, Explanation of how multispecies embeddings are scored. A joint embedding space, containing cells from multiple species, is split by species into a training set and a test set. A classification model to predict cell types is trained on a single-species training set, and evaluated on the test set of another species. The maximum test set accuracy achievable will be lower than 100% if the test set species contains specific cell types that cannot be predicted by a classifier trained on the training species. Blue denotes frog, while orange denotes zebrafish. **b**, Median performance of SATURN compared to alternative methods. The performance is evaluated using the prediction accuracy of a logistic classifier model trained to differentiate zebrafish cell types and tested on predicting the cell-type annotations of frog cells. Higher values indicate better performance, and 0.93 is the maximum accuracy that can be reached by label transfer on this dataset. SAMap represents a version of the SAMap method in which cell-type annotations are used to integrate datasets. Vertical position of scatterplot points represents the median accuracy score across 30 runs for each method. Error bars represent standard error. For batch correction methods (Harmony, scVI and Scanorama), the input genes are selected as the one-to-one homologs determined by ENSEMBL. **c**, SATURN produces more homogeneous clusters than SAMap, and these clusters contain accurate multispecies cell types. Bars represent the percentage of cells from zebrafish that are nearest neighbors of frog cells of the given cell type conserved across these two species. Cell types are ordered by frequency. **d**, Comparison of UMAP visualizations of integrated frog and zebrafish embryogenesis datasets generated by SATURN and alternative methods. In SATURN's embedding space, different cell types naturally form clusters and cells from different species align well. On the other hand, alternative baselines either do not preserve cell-type information (SAMap) or cannot integrate two species (Harmony, scVI and Scanorama).

ESM2 embeddings[14] to ESM1b[12] and ProtXL[13]. The results show that SATURN is highly robust to the choice of protein language model (Extended Data Fig. 8), as well as to the number of macrogenes (Extended Data Fig. 9). SATURN also outperforms the best baseline approach on the mammalian cell atlas dataset (Supplementary Fig. 3).

We further compare SATURN's ability to generate cell clusters that reflect conserved cell types across species, to the best baseline approach (SAMap). For each frog cell type, we analyzed its cross-species neighborhood by computing the cell-type frequency of its nearest cross-species neighbors in the embedding space. We found that SATURN generates cell clusters that are both species heterogeneous and cell-type homogeneous (Fig. 3c). For the most commonly occurring cell types, SATURN's neighborhoods were consistently highly homogeneous. On the other hand, this was not the case for SAMap where the neighborhoods were typically cell-type heterogeneous. For example, forebrain/midbrain, hindbrain, optic and eye primordium clusters were intermixed using SAMap but clearly distinguished using SATURN. SATURN aligned rare cell types such as notoplate, which only has 339 frog cells and 115 zebrafish cells. For a few very rare cell types, such as germline, which has only 33 frog cells and 53 zebrafish cells,

https://doi.org/10.1038/s41592-024-02191-z

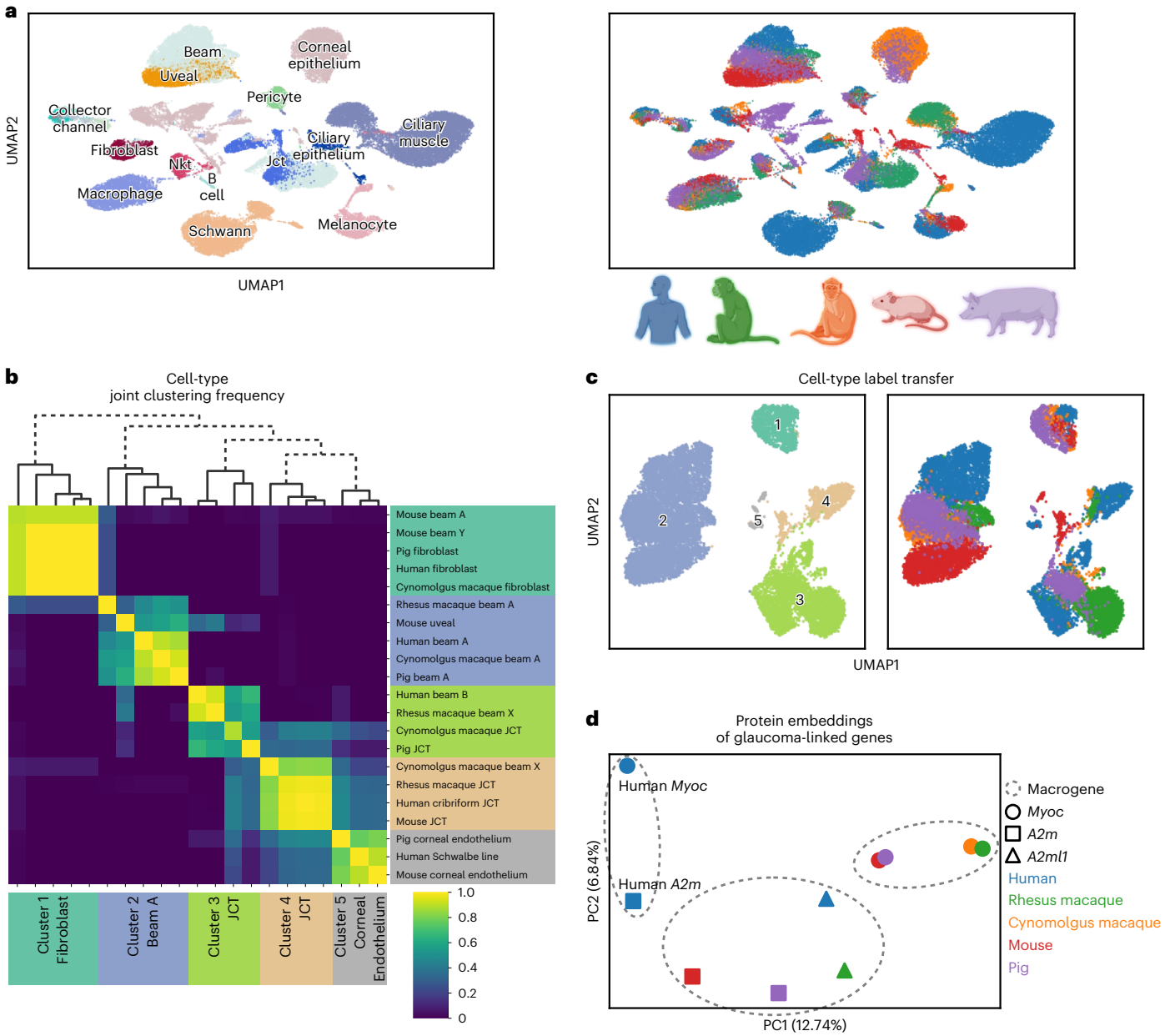


Fig. 4 | SATURN discovers new cell types and facilitates the analysis of protein embeddings for the AH cell atlas. a, SATURN successfully aligned 50,000 cells from the AH cell atlas consisting of five species: human, cynomolgus macaque, rhesus macaque, mouse and pig. UMAP visualization of SATURN's embeddings where colors denote cell types (left) and species (right). **b**,**c**, We applied SATURN to regroup cell types in a multispecies context. By clustering SATURN's embeddings, we found five broad cell types. **b**, Heat map and dendrogram of reannotated cell types using SATURN. Labels on the right side show original cell-type annotations, while on the bottom we show reannotations obtained using SATURN. These clusters include cell types originally labeled as fibroblast and beam A/Y cells (cluster 1), beam A and uveal cells (cluster 2), JCT and beam cells (cluster 3 and cluster 4) and corneal endothelium cells (cluster 5). Across 30 independent experiments, we regrouped cluster 1 as fibroblast cells, cluster 2 as beam A cells, clusters 3 and 4 as JCT cells, and cluster 5 as corneal endothelium cells. We specifically reannotated mouse beam A and beam Y cells, which have high expression of fibroblast markers such as *Pi16*, *Fbn1* and *Mfap5* as originally noted[18]. We additionally regrouped human beam B cells, which were not found in other species, as JCT cells. Finally, we mapped beam X cells, which were unique to rhesus and cynomolgus macaque, to two JCT clusters. **c**, UMAP visualizations of reannotated cell types. Cells are colored according to annotations inferred by SATURN (left) and species information (right). **d**, SATURN facilitates the analysis of protein embeddings by creation of multispecies macrogenes. Human *MYOC* had the highest weight to a different macrogene than the other four species' *Myoc* variants. The human gene *A2M* also had the highest weight to the human *MYOC* macrogene. We can investigate this discrepancy by visualizing the protein embeddings of *Myoc* and *A2m* from all five species using principal component analysis. This analysis offers potential to point to similar function in *A2m* as *Myoc*, which would otherwise not be identified by sequence-based homology, as well as potential differences in human *MYOC* and *Myoc* in the other four species.

SATURN and SAMap both failed to align. SATURN and SAMap failed to directly align additional rare cell types such as olfactory placode and hatching gland. However, SATURN aligns these cell types to functionally related cell types: 77% of olfactory placode cells were mapped to placodal area for SATURN (37% for SAMap) and 66% of hatching gland cells were mapped to another component of the EVL, the periderm, which was not case with SAMap (36% epidermal progenitor, 33% blastula).

We visually inspected low-dimensional embeddings produced by SATURN and other baselines by projecting them into a two-dimensional UMAP space[33]. We found that in SATURN's embedding space different cell types formed separate clusters, while cell types conserved across species were mixed (Fig. 3d and Supplementary Fig. 1b). On the other hand, existing methods were not able to produce biologically meaningful cell embeddings that reflect evolutionary signatures. In particular, Harmony, scVI and Scanorama failed to integrate datasets across remote species. While SAMap is able to integrate datasets across species, the cell-type information in its embedding space is no longer preserved and different cell types intermingle.

SATURN integrates five species from the AH atlas

SATURN scales to large datasets and it can handle multiple datasets at once. We applied SATURN to integrate five species of the AH atlas[18]. The AH atlas contains 50,000 cells from human, cynomolgus macaque, rhesus macaque, mouse and pig. SATURN jointly aligns different species in the embedding space, identifying many conserved cell types between these species (Fig. 4a and Supplementary Fig. 1c). SATURN embeddings suggest that cell types including melanocytes, macrophages and ciliary muscle align in all species, as do cell types that are present only in a subset of species like fibroblasts and collector channel.

SATURN can be used to reannotate cell types and correct for incomplete annotations by aligning datasets across multiple species. To demonstrate that, we use SATURN to regroup cell types from the original AH atlas in a multispecies context. We focus on beam cells (beam A/B/X/Y), fibroblasts, juxtacanalicular tissue (JCT) cells and corneal endothelium cells, due to their differential conservation across the five species profiled in the atlas.

Among these 21 cell types, SATURN found five broad clusters (Fig. 4b,c). The first cluster contained mouse beam cells and fibroblasts from pig, human and cynomolgus macaque, which we relabeled as fibroblasts. The reannotated mouse beam cells are indeed characterized as having high expression of fibroblast marker genes (Extended Data Fig. 10 and Supplementary Table 3). The second cluster contained beam A cells from pig, human, macaque and a mouse uveal cluster, which we reannotated as beam A cells. The third and fourth clusters contained beam X, beam B and JCT cells, which we reannotated as JCT cells, as beam X cells were only found in the two macaque species and beam B cells were only found in human. The fifth cluster contained the human Schwalbe line cells, and pig and mouse corneal endothelium cells. Within these new cell-type groupings, we found differentially expressed macrogenes that recapture specific cell-type marker genes (Extended Data Fig. 10 and Supplementary Table 3).

SATURN predicts different function among homologous genes

We investigate the macrogenes corresponding to glaucoma-associated genes from each species in the AH atlas. While pig, mouse, cynomolgus and rhesus macaque *Myoc* gene were expectedly linked to the same macrogene, we found that the human *MYOC* gene was not linked to that macrogene. We next visualized protein embeddings of glaucoma-associated genes and found that the human *MYOC* gene is embedded further away from the *Myoc* genes of the other species (Fig. 4d). Interestingly, the human *MYOC* gene has the highest weight to a macrogene containing human *A2M*, which is a nonhomologous gene that has also been associated with glaucoma[34], and a number of different nonhuman species' genes such as mouse *Folr1*, mouse *Fbln2*, mouse *Srgn* and pig *SCP2D1*. *A2m* genes from nonhuman species had the highest weights to the same macrogene. This analysis demonstrates that protein embeddings in SATURN and their association to macrogenes can be used to search for sequence-based gene homologs with potentially different functions across species and that SATURN can facilitate the analysis of protein embeddings through the creation of macrogenes.

Discussion

SATURN combines protein embeddings generated using large protein language models with gene expression from scRNA-seq datasets. By coupling protein embeddings with gene expression, SATURN learns universal cell embeddings that bridge differences between individual single-cell experiments even when they have different genes.

SATURN has a unique ability to map heterogeneous datasets to an interpretable space of macrogenes that can group together functionally related genes across species. In SATURN, every gene has a weight to a macrogene, which defines the importance of that gene to the macrogene. This enables SATURN to perform differential expression in the macrogene space and identify gene programs shared across different datasets. However, explicitly associating each macrogene with an interpretable function is not always possible due to the varied definitions of biological function across different contexts and scales, coupled with insufficient existing gene annotations.

SATURN represents cells as nonlinear combinations of macrogenes. To integrate datasets, the objective function introduced in SATURN learns distance metrics from weakly supervised data, which forces cells to cluster according to their cell types. SATURN allows integration of datasets generated across multiple different species. SATURN is a scalable approach, making it applicable to large-scale cross-species cell atlas datasets. Our approach also has important implications for the creation of new multi-omic machine learning methods, including those that incorporate protein assay information (for example, CITE-seq[35]), genotype or those that assay a limited section of the transcriptome (for example, MERFISH[36]). For example, to improve machine learning methods that incorporate protein assay information, proteins could be represented using protein embeddings, rather than as indices. Protein embeddings could also be modified or personalized using jointly measured genotype information. For integration of spatial datasets that profile only a subset of a transcriptome, SATURN does not require subsetting them to a set of common genes, which is required by current methods.

On the other hand, the limitation of SATURN is the requirement of a reference proteome, which may be missing for some species of interest. Reference proteomes and genomes can under-represent the genetic diversity of species, even for well-studied species such as humans[37]. Moreover, to generate the protein embeddings used by SATURN, we averaged over the embeddings produced for each gene's available protein products, ignoring various RNA splicing dynamics that affect the final translational products of genes. SATURN also requires cell clusters as an input for each dataset. These cell clusters could be created at various resolutions, which could limit the transferability of labels. Finally, smaller cell clusters, such as the germline cells in frog and zebrafish embryogenesis, are difficult to faithfully integrate.

SATURN generates cell embeddings that can be used for many downstream tasks. These tasks include but are not limited to dataset integration, discovery of conserved and species-specific cell types, differential macrogene expression analysis, cell-type reannotation, signature set enrichment, gene module determination[38] or trajectory inference[39]. As single-cell transcriptomics is applied to an increasing number of species, we expect SATURN will be an important tool for comprehending conservation and diversification of cell types across species and revealing fundamental evolutionary processes.

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/s41592-024-02191-z.

References

1. Regev, A. et al. The Human Cell Atlas. *eLife* **6**, e27041 (2017).
2. Tabula Sapiens Consortium. The Tabula Sapiens: a multiple-organ, single-cell transcriptomic atlas of humans. *Science* **376**, eabl4896 (2022).
3. Tabula Muris Consortium. Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. *Nature* **562**, 367–372 (2018).
4. Li, H. et al. Fly Cell Atlas: a single-nucleus transcriptomic atlas of the adult fruit fly. *Science* **375**, eabk2432 (2022).
5. Lu, T.-C. et al. Aging Fly Cell Atlas identifies exhaustive aging features at cellular resolution. *Science* **380**, eadg0934 (2022).
6. Korsunsky, I. et al. Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat. Methods* **16**, 1289–1296 (2019).
7. Hie, B., Bryson, B. & Berger, B. Efficient integration of heterogeneous single-cell transcriptomes using Scanorama. *Nat. Biotechnol.* **37**, 685–691 (2019).
8. 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).
9. Amodio, M. et al. Exploring single-cell data with deep multitasking neural networks. *Nat. Methods* **16**, 1139–1145 (2019).
10. Brbić, M. et al. MARS: discovering novel cell types across heterogeneous single-cell experiments. *Nat. Methods* **17**, 1200–1206 (2020).
11. Tarashansky, A. J. et al. Mapping single-cell atlases throughout metazoa unravels cell type evolution. *eLife* **10**, e66747 (2021).
12. Rives, A. et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc. Natl Acad. Sci. USA* **118**, e2016239118 (2021).
13. Elnaggar, A. et al. ProtTrans: Toward understanding the language of life through self- supervised learning. *IEEE Trans. Pattern Anal. Mach. Intell.* **44**, 7112–7127 (2022).
14. Lin, Z. et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123–1130 (2023).
15. Kilinc, M., Jia, K., & Jernigan, R. L. Improved global protein homolog detection with major gains in function identification. *Proc. Natl Acad. Sci. USA* **120**, e2211823120 (2023).
16. The Tabula Microcebus Consortium et al. Tabula Microcebus: a transcriptomic cell atlas of mouse lemur, an emerging primate model organism. Preprint at *BioRxiv* https://doi.org/10.1101/2021.12.12.469460 (2021).
17. Briggs, J. A. et al. The dynamics of gene expression in vertebrate embryogenesis at single-cell resolution. *Science* **360**, eaar5780 (2018).
18. van Zyl, T. et al. Cell atlas of aqueous humor outflow pathways in eyes of humans and four model species provides insight into glaucoma pathogenesis. *Proc. Natl Acad. Sci. USA* **117**, 10339–10349 (2020).
19. Uhlén, M. et al. Tissue-based map of the human proteome. *Science* **347**, 1260419 (2015).
20. The Human Protein Atlas. https://www.proteinatlas.org/
21. Weisel, N. M. et al. Surface phenotypes of naive and memory B cells in mouse and human tissues. *Nat. Immunol.* **23**, 135–145 (2022).
22. Sprague, J. et al. The zebrafish information network (ZFIN): the zebrafish model organism database. *Nucleic Acids Research* **31**, 241–243 (2003).
23. Bradford, Y. M. et al. Zebrafish information network, the knowledgebase for *Danio rerio* research. *Genetics* **220**, iyac016 (2022).
24. Cancelas, J. A. & Williams, D. A. Rho GTPases in hematopoietic stem cell functions. *Curr. Opin. Hematol.* **16**, 249–254 (2009).
25. Montoro, D. T. et al. A revised airway epithelial hierarchy includes CFTR-expressing ionocytes. *Nature* **560**, 319–324 (2018).
26. Deprez, M. et al. A single-cell atlas of the human healthy airways. *Am. J. Respir. Crit. Care Med.* **202**, 1636–1645 (2020).
27. Kolosov, D., Bui, P., Chasiotis, H. & Kelly, S. P. Claudins in teleost fishes. *Tissue Barriers* **1**, e25391 (2013).
28. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
29. Ashburner, M. et al. Gene Ontology: tool for the unification of biology. *Nature Genet.* **25**, 25–29 (2000).
30. Song, Y., Miao, Z., Brazma, A., & Papatheodorou, I., Benchmarking strategies for cross-species integration of single-cell RNA sequencing data. *Nat. Commun.* **14**, 6495 (2023).
31. Yates, A. et al. The ensembl REST API: ensembl data for any language. *Bioinformatics* **31**, 143–145 (2015).
32. Luecken, M. D. et al. Benchmarking atlas-level data integration in single-cell genomics. *Nat. Methods* **19**, 41–50 (2022).
33. McInnes, L., Healy, J. & Melville, J. UMAP: uniform manifold approximation and projection for dimension reduction. *J. Open Source Softw.* **3**, 861 (2018).
34. Bai, Y. et al. During glaucoma, alpha2-macroglobulin accumulates in aqueous humor and binds to nerve growth factor, neutralizing neuroprotection. *Invest. Ophthalmol. Vis. Sci.* **52**, 5260–5265 (2011).
35. Stoeckius, M. et al. Simultaneous epitope and transcriptome measurement in single cells. *Nat. Methods* **14**, 865–868 (2017).
36. Xia, C., Fan, J., Emanuel, G., Hao, J. & Zhuang, X. Spatial transcriptome profiling by MERFISH reveals subcellular RNA compartmentalization and cell cycle-dependent gene expression. *Proc. Natl Acad. Sci. USA* **116**, 19490–19499 (2019).
37. Liao, W.-W. et al. A draft human pangenome reference. *Nature* **617**, 312–324 (2023).
38. Jones, M. G., Rosen, Y. & Yosef, N. Interactive, integrated analysis of single-cell transcriptomic and phylogenetic data with PhyloVision. *Cell Rep. Methods* **2**, 100200 (2022).
39. Saelens, W., Cannoodt, R., Todorov, H. & Saeys, Y. A comparison of single-cell trajectory inference methods. *Nat. Biotechnol.* **37**, 547–554 (2019).

Methods

Overview of SATURN

SATURN takes multiple annotated single-cell RNA expression count datasets generated from S species $X^{s_1}, X^{s_2} \ldots X^{s_S}$ where $X^{s_i} \in \mathbb{N}+^{C_{s_i} \times |\mathcal{G}_{s_i}|}$ where C_{s_i} is the number of cells in species s_i and $\mathcal{G}_{s_i}$ is the set of genes in species s_i. The initial cell annotations can be obtained either from cell-type assignments if available or by running a clustering algorithm. In all experiments in the paper, we run SATURN with initial cell-type assignments within the individual species but never matched across species. In addition to count matrices and cell-type labels, SATURN also takes as input p-dimensional protein embeddings $P \in \mathbb{R}^{|\mathcal{G}| \times p}$ generated from large protein language models where $\mathcal{G} = \cup_{i=1}^{S} \mathcal{G}_{s_i}$.

SATURN maps multispecies expression data to a joint low-dimensional macrogene expression space by learning a set of macrogenes $\mathcal{M}$ with weights $W \in \mathbb{R}+^{|\mathcal{G}| \times \mathcal{M}}$ where $W_{g,m} \in \mathbb{R}+$ is a weight from a macrogene $m \in \mathcal{M}$ to a gene $g \in \mathcal{G}$. SATURN generates final k-dimensional latent cell embeddings by combining macrogenes using an encoder neural network $f : \mathbb{R}^{|\mathcal{M}|} \to \mathbb{R}^k$. SATURN consists of two main steps: (i) pretraining using an autoencoder, and (ii) fine-tuning using metric learning approach. Both steps are performed jointly on the datasets from all species.

Macrogene initialization

SATURN initializes macrogenes by soft-clustering protein embeddings. In particular, SATURN first clusters protein embeddings using the k-means algorithm[40]. Given a matrix that stores protein embeddings for all genes $P \in \mathbb{R}^{|\mathcal{G}| \times p}$, SATURN applies k-means to learn a set of centroids $\mathcal{M} = \{\mathbf{m}_i \in \mathbb{R}^p\}_{i=1}^{N_M}$ where N_M defines the number of centroids/macrogenes. k-means minimizes the within-cluster sum of squares given by equation (1):

$$\sum_{g \in \mathcal{G}} min_{\mathbf{m} \in \mathcal{M}}(||P_g - \mathbf{m}||^2), \tag{1}$$

where P_g denotes a row protein embedding vector of matrix P. Here, each centroid $\mathbf{m}$ represents a different macrogene. SATURN then defines an initial set of weights $\left\{\{W_{g,m} \in \mathbb{R}+\}_{g=1}^{|\mathcal{G}|}\right\}_{m=1}^{|\mathcal{M}|}$ from each gene g to each macrogene m as given by equation (2):

$$W_{g,m} = 2 \times \left(\log\left(\frac{1}{\mathrm{rd}_{m,g}} + 1\right)\right)^2, \tag{2}$$

where $\mathrm{rd}_{m,g} : \mathbb{N} \to \mathbb{N}$ represents the ranked Euclidean distance from gene g to a macrogene m and $\mathrm{rd}_{m,g} = 1$ for the nearest gene to a macrogene. This initialization function is arbitrarily chosen so that genes have the highest weights to the macrogenes they are closest to. Gene-to-macrogene weights are strictly positive, differentiable and updated during pretraining. We also explore different weight initialization strategies and show robustness of SATURN to different initialization functions (Supplementary Fig. 4 and Supplementary Note 6). We multiply by two so that the highest weights are close to 1.

Pretraining with an autoencoder

Following macrogene initialization, SATURN pretrains a network using an autoencoder with ZINB loss[8]. The autoencoder is composed of encoder and decoder modules. The encoder module first aggregates expression values using macrogene weights. In particular, for a cell c from species s with count values $X_c^s \in \mathbb{N}+^{|\mathcal{G}_s|}$, genes $g \in \mathcal{G}_s$ and macrogenes $m \in \mathcal{M}$, SATURN defines macrogene expression values $\mathbf{e}_c \in \mathbb{R}+^{|\mathcal{M}|}$ as given by equations (3) and (4):

$$\mathbf{e}_c = \mathrm{ReLU}(\mathrm{LayerNorm}(W_s^T \log(X_c^s + 1))) \tag{3}$$

$$W_s^T = \begin{bmatrix} W_{1,1} & \ldots & W_{1,|\mathcal{G}_s|} \\ \ldots & \ldots & .. \\ W_{|\mathcal{M}|,1} & \ldots & W_{|\mathcal{M}|,|\mathcal{G}_s|}, \end{bmatrix}, \tag{4}$$

where ReLU denotes the rectified linear unit used as the activation function and defined as 'ReLU($\cdot$) = max(0, $\cdot$)'. Macrogene expression values are always positive to ensure that each gene positively contributes to a macrogene or does not contribute at all. LayerNorm is layer normalization[41] defined according to equation (5):

$$\mathrm{LayerNorm}(\mathbf{X}) = \frac{\mathbf{X} - E[\mathbf{X}]}{\sqrt{\mathrm{Var}(\mathbf{X}) + \epsilon}} \times \boldsymbol{\gamma} + \boldsymbol{\beta}. \tag{5}$$

The encoder module f consists of two fully connected neural network layers with ReLU activation, layer normalization and dropout, and takes as an input $\mathbf{e}_c \in \mathbb{R}+$ and outputs a low-dimensional embedding $\mathbf{z}_c \in \mathbb{R}^k$ given by equation (6):

$$\mathbf{z}_c = f(\mathbf{e}_c). \tag{6}$$

The decoding module outputs three distinct heads, parameterizing $|\mathcal{G}|$ ZINB distributions as given by equations (7–9): $\boldsymbol{\mu}_c \in \mathbb{R}+^{|\mathcal{G}|}, \mathbf{O}_c \in \mathbb{R}^{|\mathcal{G}|}, \boldsymbol{\theta} \in \mathbb{R}+^{|\mathcal{G}|}$.

$$\boldsymbol{\mu}_c = \mathrm{Softmax}(W_s D_\mu(D_S(\mathbf{z}_c))) \sum X_c^s \tag{7}$$

$$\mathbf{O}_c = D_O(D_S(\mathbf{z}_c)) \tag{8}$$

$$\boldsymbol{\theta}, \tag{9}$$

where D_S, D_μ and D_O represent fully connected neural network layers. D_S and D_μ have ReLU activation, dropout and layer normalization. $\boldsymbol{\theta}$ is a differentiable parameter of the model. SATURN provides the ability to concatenate a one-hot representation of the species s to the embedding $\mathbf{z}_c$ in equation (6) during pretraining of the autoencoder. However, we find that this does not improve the performance and set the species conditional variable to a constant value in all experiments (Supplementary Fig. 5). That including the species as a conditional variable does not improve performance may be of consideration for the development of other autoencoder-based methods for single-cell expression data. However, while performance was not helped in this case, for other settings, or datasets, a conditional autoencoder (CAE) might be the correct choice, and we include the ability to pretrain with a CAE in the SATURN codebase.

The autoencoder reconstruction loss $\mathcal{L}_{rc}$, is calculated as the negative log likelihood of a ZINB distribution[8] parameterized according to equations (10) and (11):

$$\mathrm{ZINB}_{c,g} \approx \begin{cases} \mathrm{Poisson}(\mathrm{gamma}(\boldsymbol{\theta}_g, \boldsymbol{\theta}_g/\boldsymbol{\mu}_{cg})), & \text{if Bernoulli}\left(\frac{\exp \mathbf{O}_{cg}}{1+\exp \mathbf{O}_{cg}}\right) = 0 \\ 0, & \text{otherwise} \end{cases} \tag{10}$$

$$\mathcal{L}_{rc} = \sum_{g \in \mathcal{G}_s} -\log(\mathbb{P}(\mathrm{ZINB}_{c,g} = X_{cg})), \tag{11}$$

where $\mathbb{P}$ denotes probability. To ensure that gene-to-macrogene weights reflect similarity in protein embedding space, we add an additional loss term $\mathcal{L}_s$ defined according to equation (12):

$$\mathcal{L}_s = \mathrm{MSE}(sim(B, B_{\mathrm{shuffled}}), sim(P, P_{\mathrm{shuffled}})), \tag{12}$$

where $B = Q(W)$ and $Q : \mathbb{N}+^{|\mathcal{M}|} \to \mathbb{N}^n$ is a fully connected neural network layer with ReLU activation, layer normalization and dropout, which encodes macrogene weights. MSE denotes mean squared error and sim is the cosine similarity. The encoded macrogene weights $B \in \mathbb{R}^{|\mathcal{G}| \times n}$ and protein embeddings P are jointly shuffled row-wise (gene-wise).

The final pretraining loss $\mathcal{L}_p$ that SATURN optimizes is defined according to equation (13):

$$\mathcal{L}_p = \tau\mathcal{L}_s + \frac{1}{|c \in \text{mini} - \text{batch}|}\sum_{c\in\text{mini}-\text{batch}}\mathcal{L}_{rc}, \quad (13)$$

where τ is a regularization parameter and it is set to 1 in all experiments and mini-batch is a training mini-batch.

Metric learning across species

To automatically learn a distance metric across species, SATURN fine-tunes pretrained cell embeddings with a weakly supervised metric learning objective. In particular, SATURN relies on the triplet margin loss function given by equation (14):

$$\mathcal{L}_t = \max(D(\mathbf{z}_a, \mathbf{z}_p) - D(\mathbf{z}_a, \mathbf{z}_n) + m, 0), \quad (14)$$

where D is a cosine distance, a, p and n denote an anchor cell, a positive cell and a negative cell, respectively, and the margin m is a tunable hyperparameter that we set to 0.2 in all experiments. Triplets are mined using semihard online mining in a weakly supervised fashion. To mine triplets, SATURN iterates over the species-specific cell-type annotations, but no cross-species annotations are ever used. These within-species annotations can be predetermined or generated in an unsupervised manner with clustering techniques like Leiden clustering[42]. For each annotation, SATURN selects all cells with that annotation from the same species as candidate anchor cells. Then, for each anchor cell, SATURN selects candidate positive cells as mutual 1-nearest neighbors measured using cosine distance in the embedding space. Here, mutual means that if cell x from species s_1 selected as its cross-species nearest neighbor cell y from species s_2, SATURN finds the nearest neighbor x' of cell y in species s_1. If cells x and x' from species s_1 have the same annotation, then positive pairs are generated. The anchor cells and positive cells are pooled, and then matched such that each anchor cell candidate has a corresponding randomly selected positive cell candidate from a different species. Finally, negative cells are randomly selected such that they have a different label than either the anchor label or the positive label. Triplets are semihard filtered such that (equation (15)):

$$D(\mathbf{z}_a, \mathbf{z}_p) < D(\mathbf{z}_a, \mathbf{z}_n) < D(\mathbf{z}_a, \mathbf{z}_p) + m. \quad (15)$$

During the fine-tuning stage, macrogene weights are not updated.

Generation of protein embeddings

Protein embeddings are generated by applying a pretrained protein embedding language model on each species' reference proteome. Protein embeddings generated by the ESM2 model[14] were used for all experiments. The ESM2 protein embedding model accepts a sequence of amino acids as an input and outputs a $p = 5120$ dimensional vector representing the embedding of the protein. To obtain a protein embedding for a gene, the protein embeddings of all proteins available for the gene are averaged. Any protein embedding model, or any model that outputs numerical representations of genes, can be used as an input to SATURN (Extended Data Fig. 8).

Differential macrogene expression

Differential expression on macrogene values is performed using a Wilcoxon rank-sum test as implemented in SCANPY[43]. For a cell-type annotation t, with cells $c \in t$ (from any species), the test statistic U_m for macrogene m is calculated according to equations (16) and (17):

$$U_m = R_m - \frac{|c \in t|(|c \in t| + 1)}{2} \quad (16)$$

$$R_m = \sum_{c\in t}\text{Rank}(m)[c], \quad (17)$$

where $R(m)$ is the rank sum of cells with label t for macrogene m.

Determining gene homologs

BLASTP (v2.9.0) with default settings was applied to publicly available reference proteomes from ENSEMBL. BLASTP homolog results were used to find homolog gene pairs within the genes with highest weight to each macrogene (Fig. 2d). BLASTP results are also used for SAMap alignment (Fig. 3). The ENSEMBL homology API was queried to determine one-to-one gene homologs.

Reporting summary

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

Data availability

All analyzed datasets are publicly available. Tabula Sapiens is available at https://cellxgene.cziscience.com/collections/e5f58829-1a66-40b5-a624-9046778e74f5. Tabula Microcebus is available at https://figshare.com/articles/dataset/Tabula_Microcebus_v1_0/14468196?file=31777475. Tabula Muris is available at https://figshare.com/articles/dataset/Single-cell_RNA-seq_data_from_microfluidic_emulsion_v2_/5968960/2. For embryogenesis datasets, frog is available under accession code GSE113074 and zebrafish is available in h5ad format at https://kleintools.hms.harvard.edu/paper_websites/wagner_zebrafish_timecourse2018/WagnerScience2018.h5ad. The five species AH atlas datasets are available under accession code GSE146188.

Code availability

SATURN was written in Python using the PyTorch (v1.13.1) library. The source code is available on GitHub at https://github.com/snap-stanford/saturn/. The repository used in the paper is deposited under https://doi.org/10.5281/zenodo.10258201 in Zenodo[44].

References

40. Lloyd, S. Least squares quantization in PCM. *IEEE Trans. Inform. Theory* **28**, 129–137 (1982).
41. Ba, J. L., Kiros, J. R., & Hinton, G. E., Layer normalization. Preprint at https://arxiv.org/abs/1607.06450 (2016).
42. Traag, V. A., Waltman, L. & Van Eck, N. J. From Louvain to Leiden: guaranteeing well-connected communities. *Scientific Rep.* **9**, 5233 (2019).
43. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.* **19**, 15 (2018).
44. Rosen, Y. et al. Towards universal cell embeddings: integrating single-cell RNA-seq datasets across species with SATURN. Preprint at *BioRxiv* https://doi.org/10.1101/2023.02.03.526939 (2023).
45. Stelzer, G. et al. The genecards suite: from gene data mining to disease genome sequence analyses. *Curr. Protoc. Bioinformatics* **54**, 1.30.1–1.30.33 (2016).
46. Safran, M. et al. The GeneCards suite. in *Practical Guide to Life Science Databases* 27–56 (Springer, 2021).

Acknowledgements

We gratefully acknowledge the support of DARPA under nos. HR00112190039 (TAMI), N660011924033 (MCS); ARO under nos. W911NF-16-1-0342 (MURI), W911NF-16-1-0171 (DURIP); NSF under nos. OAC-1835598 (CINES), OAC-1934578 (HDR), CCF-1918940 (Expeditions); National Institutes of Health under no. 3U54HG010426-04S1 (HuBMAP), Stanford Data Science Initiative, Wu Tsai Neurosciences Institute, Amazon, Docomo, GSK, Hitachi, Intel, JPMorgan Chase, Juniper Networks, KDDI, NEC and Toshiba. M.B. acknowledges the EPFL support. Figure elements, including icons of species, were created with BioRender.com.

Author contributions

M.B., Y. Roohani and J.L. conceived the study. Y. Rosen, M.B., Y. Roohani and J.L. performed research, contributed new analytical tools, designed algorithmic framework, analyzed data and wrote the paper. Y. Rosen performed experiments and developed the software. K.S. and Z.L. contributed to code.

Competing interests

The authors declare no competing interests.

Additional information

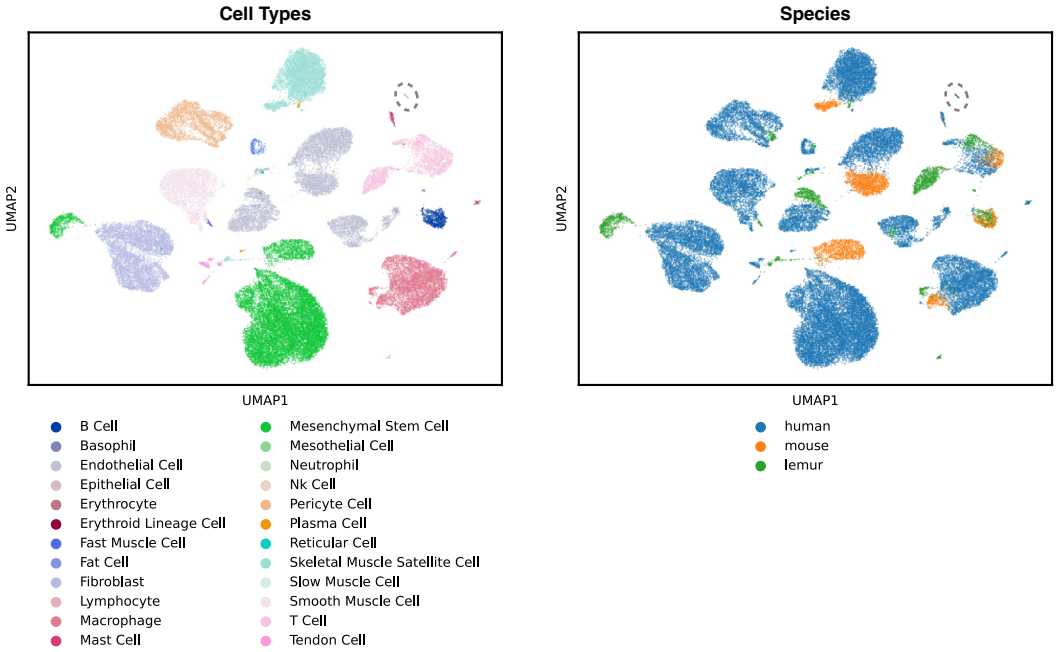
Extended data is available for this paper at https://doi.org/10.1038/s41592-024-02191-z.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41592-024-02191-z.

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

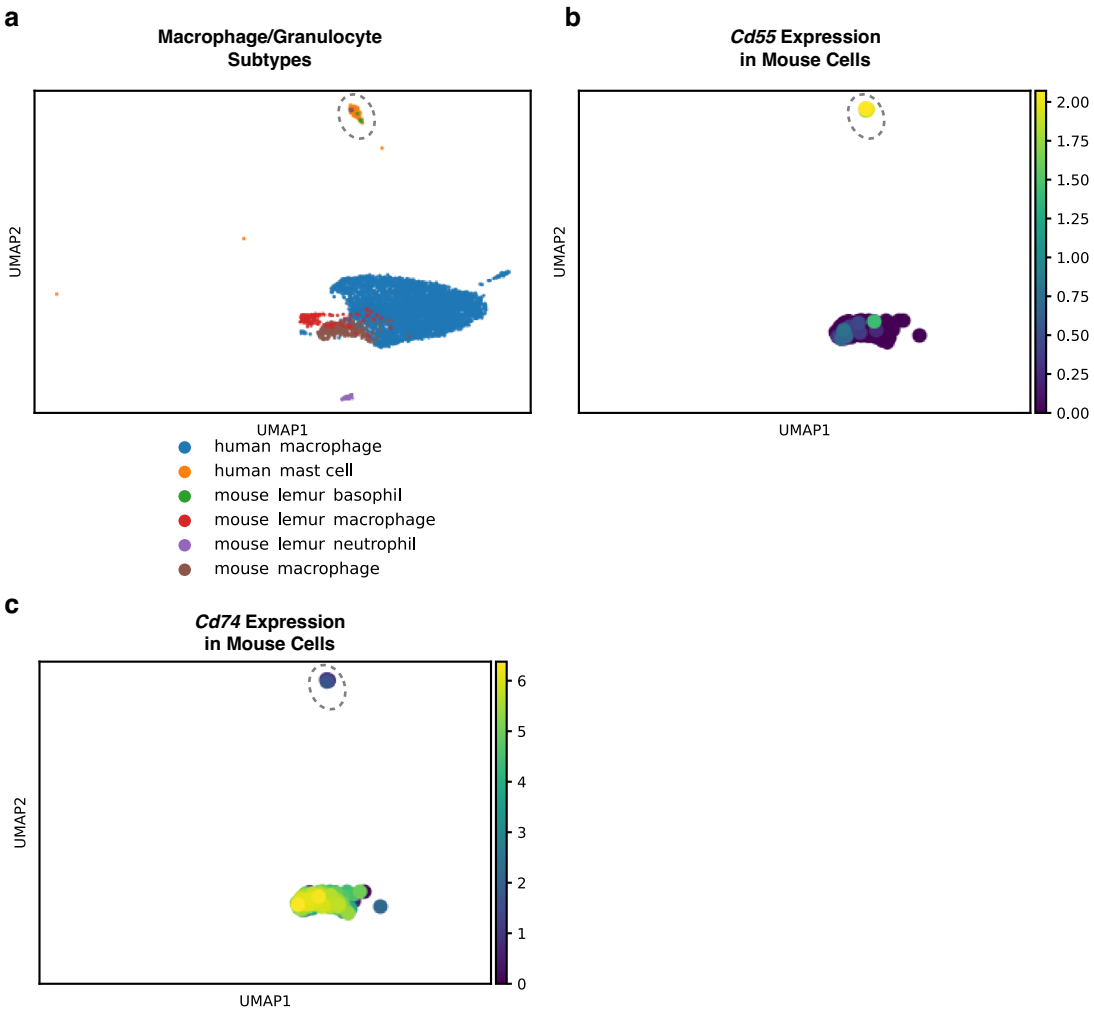
Peer review information *Nature Methods* thanks Xin Gao, Malte Luecken and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Primary Handling Editor: Lin Tang, in collaboration with the *Nature Methods* team.

https://doi.org/10.1038/s41592-024-02191-z



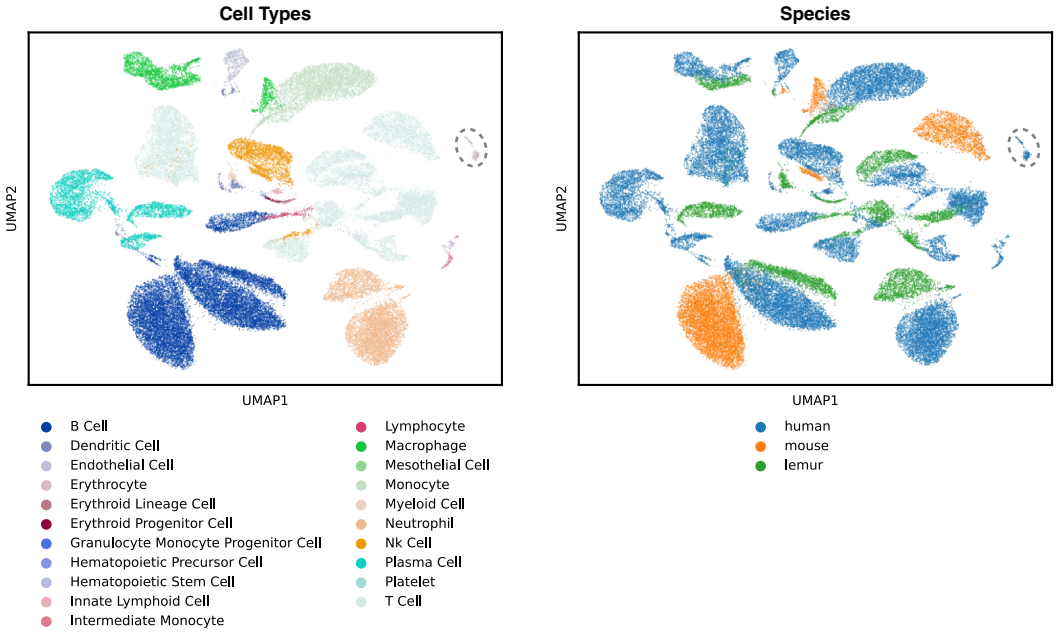
Extended Data Fig. 1 | SATURN integrates muscle cell types across three mammalian species. UMAP visualization of SATURN's embeddings obtained by integrating muscle tissues from human, mouse and lemur. Cells are colored based on broad-level cell types (left) and based on the species they come from (right). Epithelial and mesothelial cells types, which were only found within human, form a unique cluster (circled). To create each dataset, the larger Tabula datasets were subsetted. The human subset included cells labeled as muscle and vasculature. For mouse, limb muscle was chosen. For lemur, limb muscle and diaphragm were chosen. Human mesothelial cells belong to the tissue labeled muscle, and epithelial cells belong to the tissue labeled vasculature.

https://doi.org/10.1038/s41592-024-02191-z



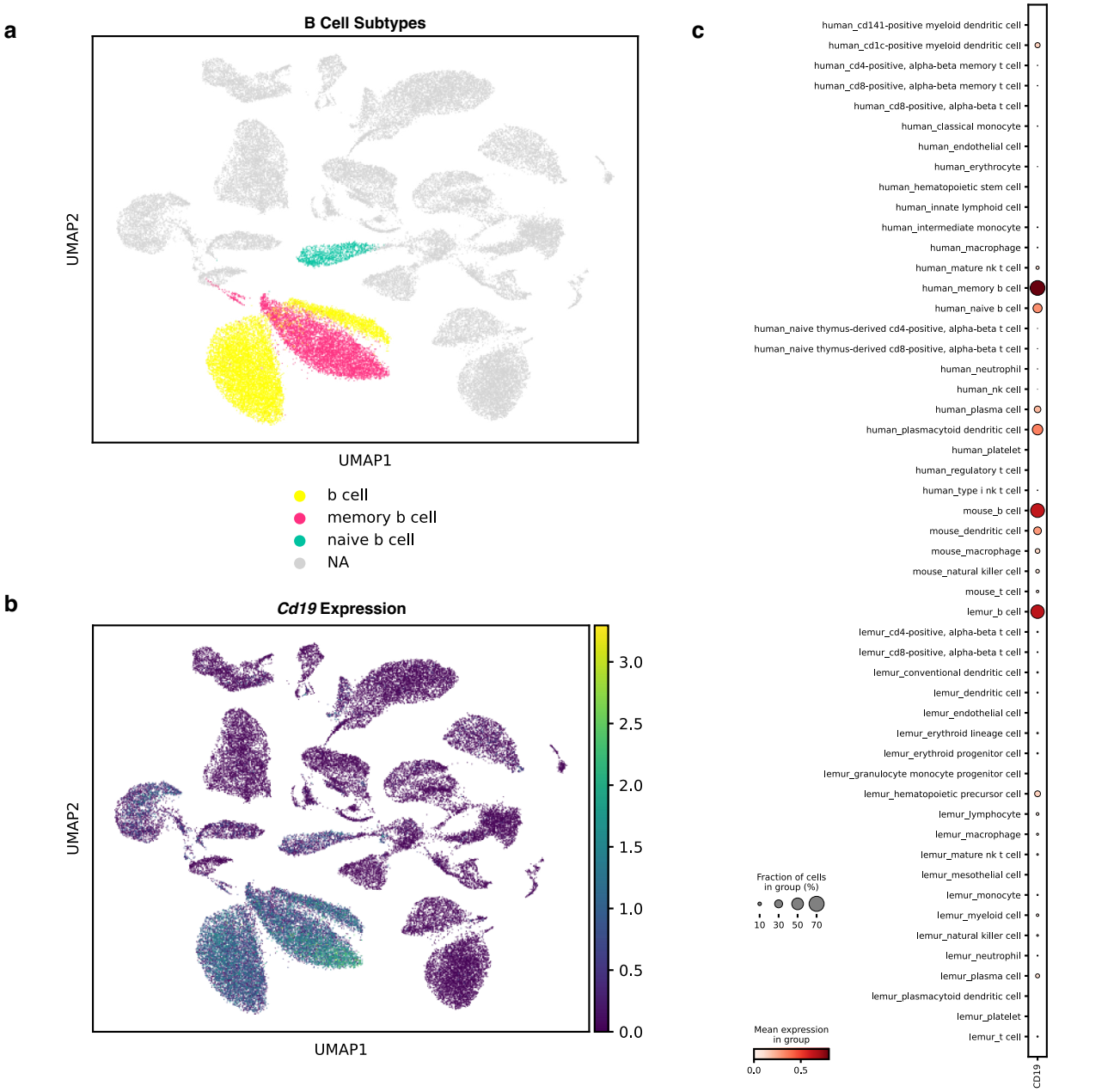
Extended Data Fig. 2 | Reannotation of cells labeled as mouse macrophage in mouse muscle. UMAP visualization of macrophage and granulocyte cell types obtained by integrating cells from human, mouse and lemur. **(a)** A small group of mouse macrophages cluster with granulocyte cell types from human (mast cells) and lemur (basophil) (circled), while other mouse macrophages cluster with human and lemur macrophages. These mouse cells **(b)** express *Cd55*, which has been shown to be preferentially expressed in granulocytes[19,20], and **(c)** do not express *Cd74*, which has been shown to be preferentially expressed in macrophages and not expressed in granulocytes[19,20]. **(b)**, **(c)** are colored by log-normalized expression.

https://doi.org/10.1038/s41592-024-02191-z



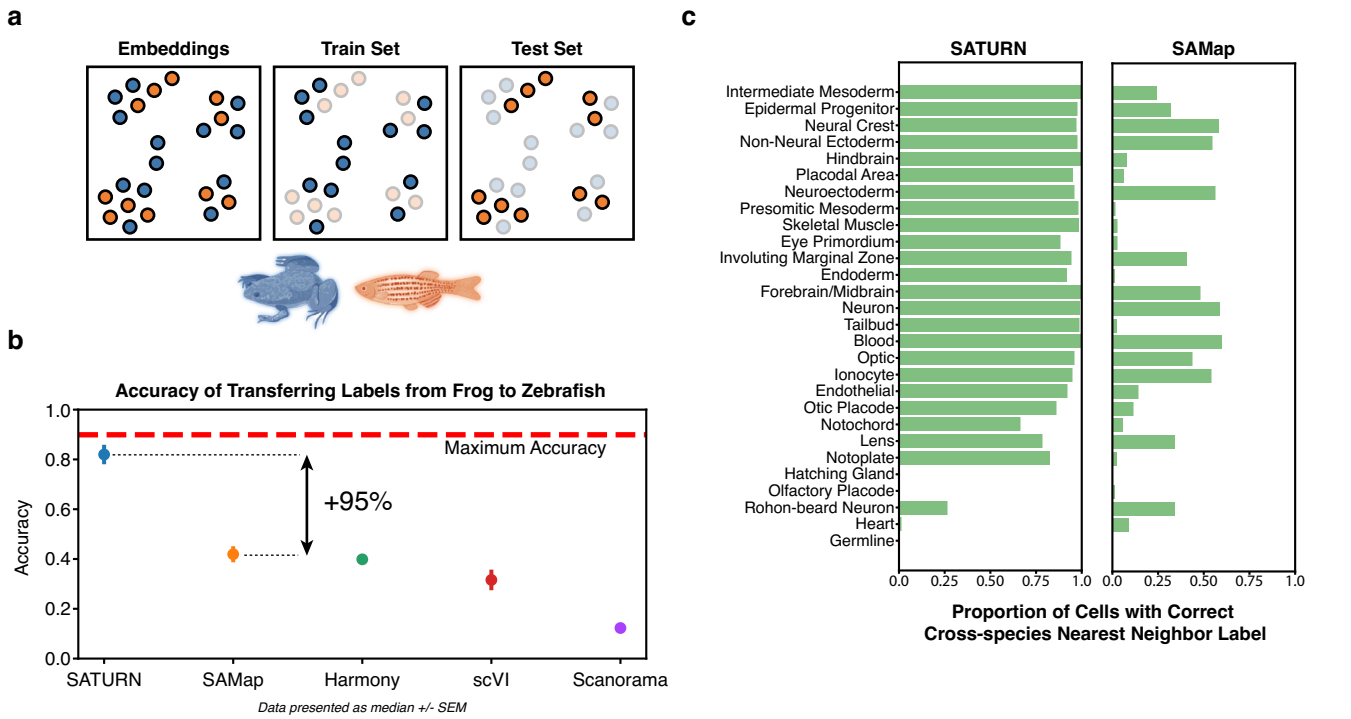
Extended Data Fig. 3 | Saturn integrates spleen cell types across three mammalian species. UMAP visualization of SATURN's embeddings obtained by integrating spleen tissues from the human, mouse and lemur. Cells are colored based on broad-level cell types (left) and based on the species they come from (right). Erythrocytes, which were only found within human, form a unique cluster (circled).

https://doi.org/10.1038/s41592-024-02191-z



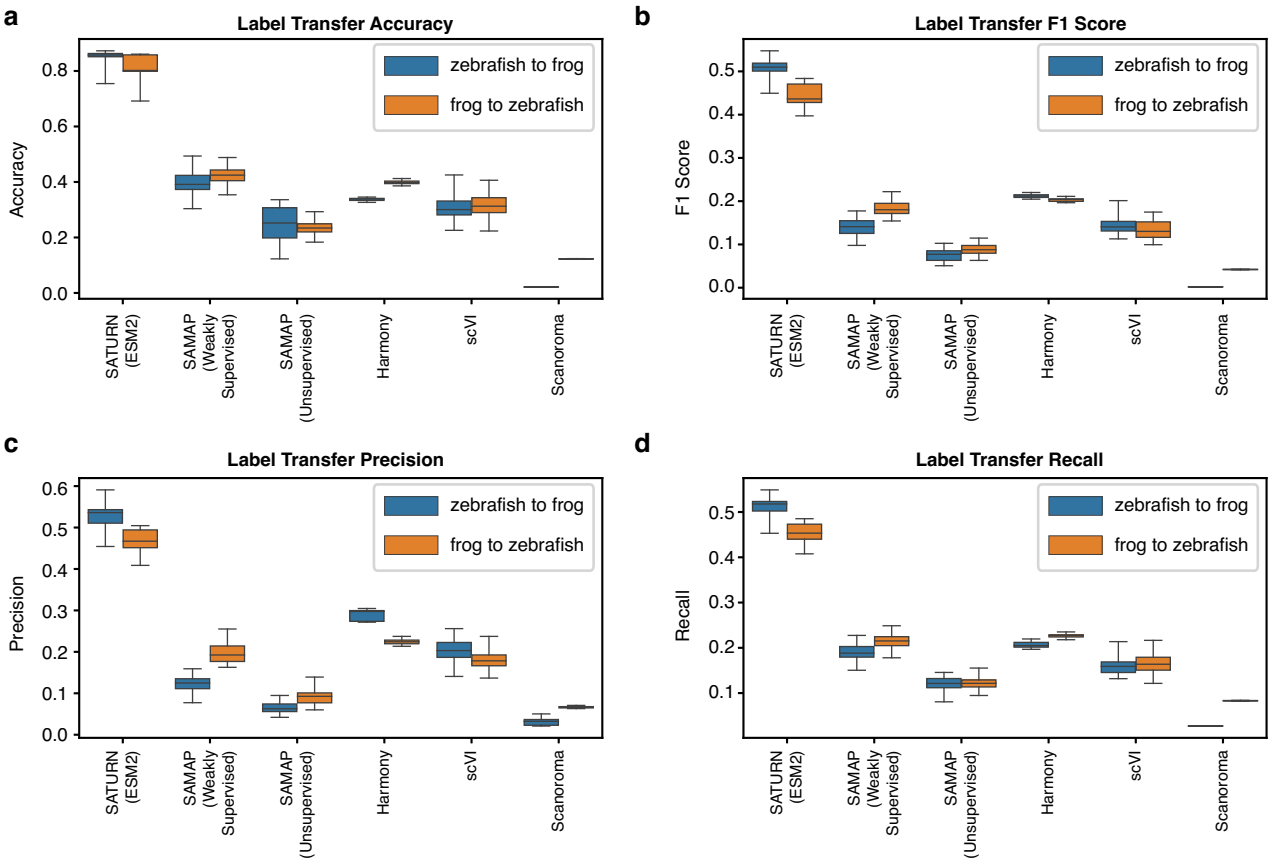
Extended Data Fig. 4 | SATURN annotates B cells in mouse and lemur spleen on a fine-grained level. (a) UMAP visualization of SATURN's embeddings obtained by integrating spleen cells from the human, mouse and lemur. B cells are shown in different colors based on ground-truth annotations while other cells are in grey. **(b)** UMAP visualization of expression of B cell marker *Cd19*. **(c)** Dotplot of *Cd19* expression vs species and cell type. *Cd19* is expressed in human memory B cells, mouse and lemur B cells, and only weakly in human naive B cells. This indicates that mouse and lemur B cells are correctly clustered with memory B cells.

https://doi.org/10.1038/s41592-024-02191-z



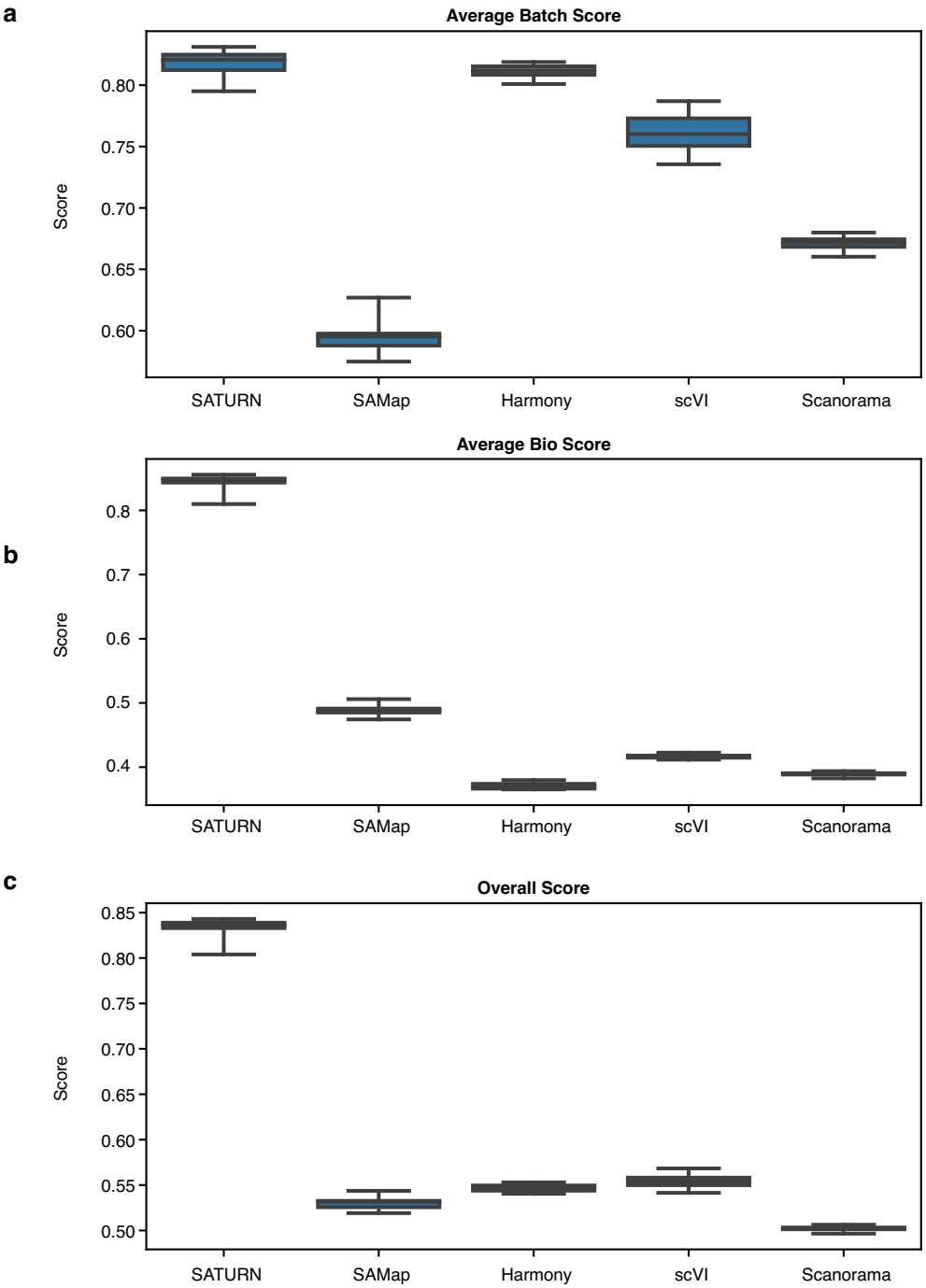
Extended Data Fig. 5 | Label transfer from from frog to zebrafish embryogenesis datasets. (a) Explanation of how multi-species embeddings are scored. A joint embedding space, containing cells from multiple species, is split by species into a training set and a test set. A classification model to predict cell types is trained on the frog training set cells, and evaluated on the zebrafish test set cells. The maximum test set accuracy achievable will be lower than 100% if the test set species contains specific cell types that can not be predicted by a classifier trained on the training species. Blue color denotes frog, while orange denotes zebrafish. **(B)** Median performance of SATURN compared to alternative methods. The performance is evaluated using the prediction accuracy of a logistic classifier model trained to differentiate frog cell types and tested on predicting the cell type annotations of zebrafish cells. Higher values indicate better performance, and 90% is the maximum accuracy that can be reached by label transfer on this dataset. SAMap represents a version of the SAMap method in which cell-type annotations are used to integrate datasets. Vertical position of scatter plot points represents the median accuracy score across 30 runs for each method. Error bars represent standard error. For batch correction methods (Harmony, scVI and Scanorama), the input genes are selected as the one to one homologs determined by ENSEMBL. **(c)** SATURN produces more homogeneous clusters than SAMap, and these clusters contain accurate multi species cell types. Bars represent the percentage of cells from frog that are nearest neighbors of zebrafish cells of the given cell type conserved across these two species. Cell types are ordered by frequency.

https://doi.org/10.1038/s41592-024-02191-z



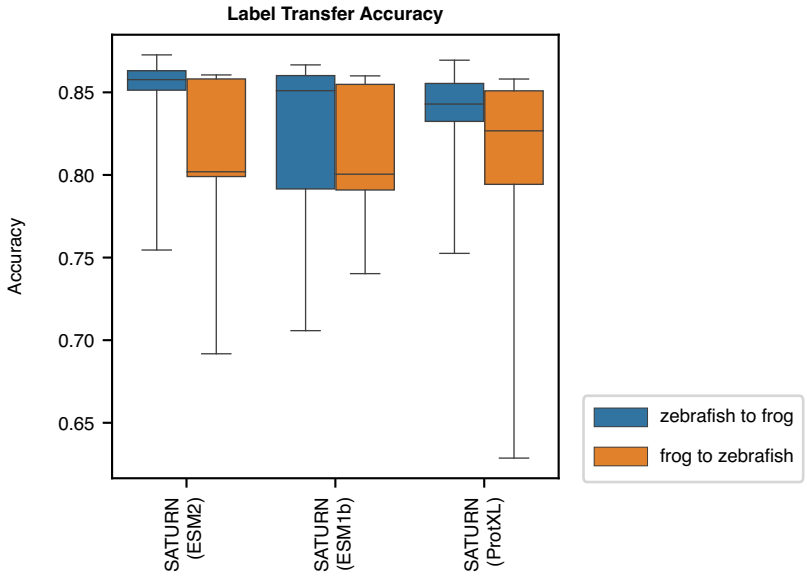
Extended Data Fig. 6 | Performance comparison using different evaluation metrics. Median performance of SATURN and baseline methods on label transfer between frog and zebrafish embryogenesis datasets evaluated using **(a)** accuracy, **(b)** macro-F1-score, **(c)** macro-precision, and **(d)** macro-recall. Blue boxplots show zebrafish to frog label transfer performance, while orange boxplots show frog to zebrafish label transfer performance. Distribution is estimated with n = 30 runs of each method.

https://doi.org/10.1038/s41592-024-02191-z



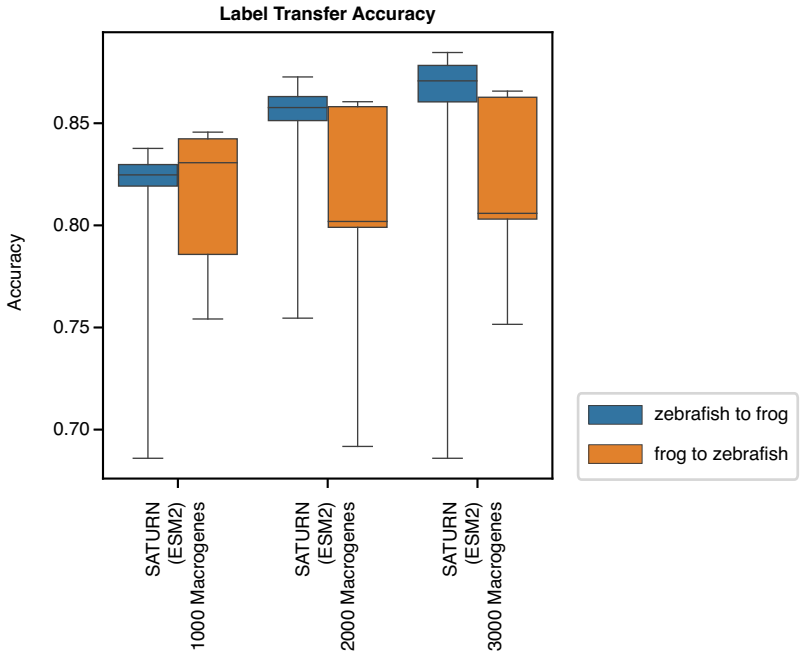
Extended Data Fig. 7 | Performance comparison using batch integration evaluation metrics. Median performance of SATURN and baseline methods evaluated using **(a)** weighted mean of the batch removal score (Avg Batch), **(b)** bio-conservation score (Avg Bio), and **(c)** overall score. Distribution is estimated with n = 30 runs of each method. Overall score is calculated as (0.6 * Avg Bio) + (0.4 * Avg Batch). Avg Bio is calculated as the average of NMI cell type score, ARI cell type score, and ASW cell type scores. Avg Batch is calculated as the average of ASW batch score, and graph connectivity scores, where species is taken as the batch variable. Neighbor calculation is done using default Scanpy settings, using each methods' embeddings. Score calculation is done using default SCIB settings[32].

https://doi.org/10.1038/s41592-024-02191-z



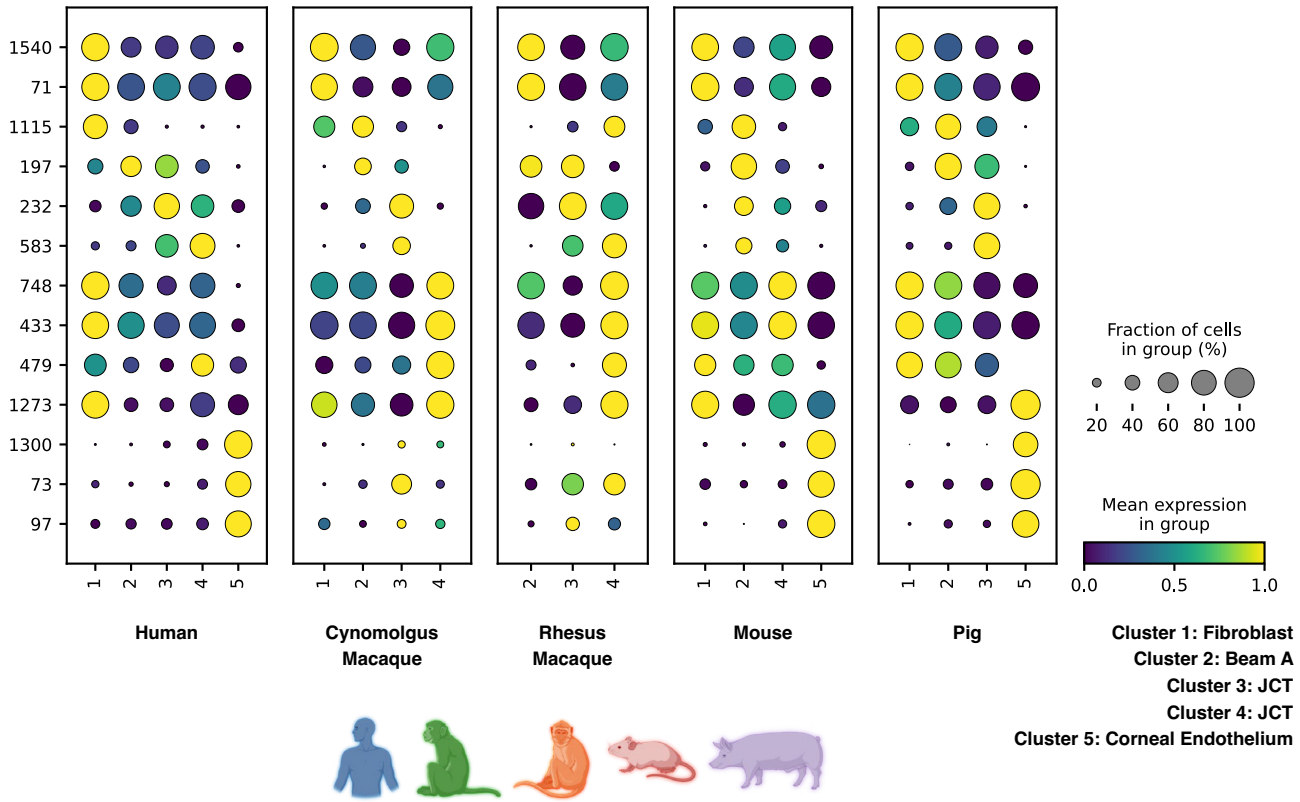
Extended Data Fig. 8 | SATURN is robust to protein language model choice. Median performance of SATURN with different protein language model embeddings on label transfer between frog and zebrafish embryogenesis datasets evaluated using accuracy. Blue boxplots show zebrafish to frog label transfer performance, while orange boxplots show frog to zebrafish label transfer performance. Distribution is estimated with n = 30 runs. ESM2 refers to the *esm2 t48 15B UR50D* model[14]. ESM1b refers to the *esm1b t33 650M UR50S* model[12]. ProtXL refers to the *ProtT5 XL U50* model[13].

https://doi.org/10.1038/s41592-024-02191-z



Extended Data Fig. 9 | SATURN is robust to choice of number of macrogenes. Median performance of SATURN with different number of macrogenes on label transfer between frog and zebrafish embryogenesis datasets evaluated using accuracy. Blue boxplots show zebrafish to frog label transfer performance, while orange boxplots show frog to zebrafish label transfer performance. Distribution is estimated with n = 30 runs.

https://doi.org/10.1038/s41592-024-02191-z



Extended Data Fig. 10 | Differentially expressed macrogenes in regrouped AH Atlas cell types. Rows correspond to macrogene numbers, and columns correspond to cluster numbers. Genes composing each macrogene are listed in Supplementary Data Table 3. Cluster 1 expresses collagen genes like *Col6a2* (macrogene 1540) which are known fibroblast markers[18]. Cluster 2 expresses *Nr2f1* (macrogene 197) which was identified as a trabecular meshwork marker[18]. Cluster 3 expresses Rspo genes (macrogene 583). *Rspo4* was identified as a marker in human JCT[18]. Cluster 4 expresses *Angptl7* (macrogene 479) which was identified as a JCT marker[18]. Cluster 5 expresses corneal endothelium markers including *Ca3* (macrogene 1300). Additionally, Cluster 5 contains a macrogene composed of *Slc4a* genes. Another solute carrier family (SLC) gene[45,46], *Slc11a2* was identified as differentially expressed in human corneal endothelium[18].

nature portfolio

Corresponding author(s): Jure Leskovec

Last updated by author(s): 2023/12/05

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	We used code for proccessing scRNA-seq count data available here: https://github.com/snap-stanford/SATURN
Data analysis	SATURN was written in Python 3.8.1 using the PyTorch library (v1.13.1). The source code is available at https://github.com/snap-stanford/SATURN including code to download and process raw data. Additionally, SATURN outputs are available at http://snap.stanford.edu/saturn/data/. Homologs were determined using BLASTP version 2.9.0

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 analyzed datasets are publicly available and listed in the data availability section. Tabula Sapiens is available at CellXGene https://cellxgene.cziscience.com/collections/e5f58829-1a66-40b5-a624-9046778e74f5. Tabula Microcebus is available at FigShare https://figshare.com/articles/dataset/

Tabula_Microcebus_v1_0/14468196?file=31777475. Tabula Muris is available at FigShare https://figshare.com/articles/dataset/Single-cell_RNA-seq_data_from_microfluidic_emulsion_v2_/5968960/2. For embryogenesis datasets, frog is available with accession code GSE113074 and zebrafish is available in h5ad format at KleinTools. The five species aqueous humor outflow pathway atlas datasets are available with accession code GSE146188.

Human research participants

Policy information about studies involving human research participants and Sex and Gender in Research.

Reporting on sex and gender	N/A
Population characteristics	N/A
Recruitment	N/A
Ethics oversight	N/A

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	No samples were taken and the full dataset was used in each case for alignment.
Data exclusions	For aligning mammalian cell atlases, the cell atlases were subset to the list of common tissues in order to provide a more meaningful alignment.
Replication	Alignment benchmarks and reannotation results were verified by replicating SATURN results with 30 independent runs, each paramtetrized with a different random seed, and provided consistent results.
Randomization	For all methods, results for benchmarks and reannotation were calculated for 30 independent runs, each parameterized with a different random seed set from 0-29. For PCA calculation, data was shuffled with 30 different random seeds.
Blinding	No blinding was performed in the study because no animals or patients were treated in the study and computational algorithms were trained on preestablished datasets.

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

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging