Universal Cell Embeddings: A Foundation Model for Cell Biology

Yanay Rosen[1,*], Yusuf Roohani[2,*], Ayush Agrawal[1], Leon Samotorčan[1],
Tabula Sapiens Consortium[3], Stephen R. Quake[4,5,6,†], Jure Leskovec[1,†]
[1] Department of Computer Science, Stanford University, Stanford, CA, USA
[2] Department of Biomedical Data Science, Stanford University, Stanford, CA, USA
[3] Chan Zuckerberg BioHub, San Francisco, CA, USA
[4] Department of Bioengineering, Stanford University, Stanford, CA, USA
[5] Department of Applied Physics, Stanford University, Stanford, CA, USA
[6] Chan Zuckerberg Initiative, Redwood City, CA, USA
[†]Corresponding author. Email: jure@cs.stanford.edu, quake@stanford.edu
[*]These authors contributed equally

Abstract

Developing a universal representation of cells which encompasses the tremendous molecular diversity of cell types within the human body and more generally, across species, would be transformative for cell biology. Recent work using single-cell transcriptomic approaches to create molecular definitions of cell types in the form of cell atlases has provided the necessary data for such an endeavor. Here, we present the Universal Cell Embedding (UCE) foundation model. UCE was trained on a corpus of cell atlas data from human and other species in a completely self-supervised way without any data annotations. UCE offers a unified biological latent space that can represent any cell, regardless of tissue or species. This universal cell embedding captures important biological variation despite the presence of experimental noise across diverse datasets. An important aspect of UCE's universality is that any new cell from any organism can be mapped to this embedding space with no additional data labeling,

model training or fine-tuning. We applied UCE to create the Integrated Mega-scale Atlas, embedding 36 million cells, with more than 1,000 uniquely named cell types, from hundreds of experiments, dozens of tissues and eight species. We uncovered new insights about the organization of cell types and tissues within this universal cell embedding space, and leveraged it to infer function of newly discovered cell types. UCE's embedding space exhibits emergent behavior, uncovering new biology that it was never explicitly trained for, such as identifying developmental lineages and embedding data from novel species not included in the training set. Overall, by enabling a universal representation for every cell state and type, UCE provides a valuable tool for analysis, annotation and hypothesis generation as the scale and diversity of single cell datasets continues to grow.

Introduction

Cells are the fundamental unit of life and biologists have long conceptualized cells as members of different universal landscapes [1–4]. A notable example of this is the Waddington landscape, which presents a theoretical framework for the developmental lineages of cells as they transition from pluripotent stages such as stem cells to more terminally differentiated end points [5]. Broadly, the field of cell biology has sought to map the range of phenotypes that cells might exhibit, their interrelationships, and the shifts between these states during development and disease [6–10].

The substantial growth in the size of single-cell RNA sequencing (scRNA-seq) datasets presents a fresh opportunity to revisit these questions. Detailed transcriptomic snapshots of cells are now widely available from a range of timepoints, tissues, donors, and species [11–13]. These rich, high-dimensional states are typically distilled into low-dimensional vectors or embeddings to facilitate computational analysis [14, 15]. However, existing computational approaches struggle to jointly analyze these diverse datasets. The unified representations they produce are often unable to extend to new datasets due to species-specific constraints in their construction or the presence of dataset-specific artifacts (or batch effects) which can obscure the underlying biological signal [16, 17].

Some computational methods for scRNA-seq data have managed to overcome these limitations, but at the cost of requiring model tuning for each new dataset, thus rendering the representations non-universal [15, 18, 19]. As a result, whenever a new experiment is performed and new data is collected, it requires dedicated, resource-intensive data labeling and model training to perform even the most standard analyses, such as clustering or annotation. This process is both time consuming and inefficient, and results in sub-optimal analyses based on small, limited and private datasets.

Recent advances in the field of artificial intelligence have enabled general-purpose foundation models (such as ChatGPT [20, 21], PaLM [22], Llama [23] and SAM [24]) that can learn universal representations that are then applied to diverse downstream tasks and analyses. These foundation models are not specifically trained for these downstream tasks, thus presenting clear instances of emergent capabilities [25]. This foundation model strategy has also found valuable applications in biological contexts such as learning representations of protein and DNA sequences [26,27]. While some recent work has applied foundation model architectures to single-cell genomics data, the unique characteristics of these datasets necessitate a specialized modeling approach to fully realize their potential [28, 29]. Directly modeling gene expression as text in the

form of a sequence of genes is both inefficient from a learning perspective and often relies on inaccurate biological assumptions.

Here, we present Universal Cell Embedding (UCE), a foundation model for single-cell gene expression that is designed to address questions in cell and molecular biology. UCE is uniquely able to generate representations of new single-cell gene expression datasets with no model fine-tuning or retraining while still remaining robust to dataset and batch-specific artifacts. Moreover, it does so while requiring no cell type annotation and no input dataset preprocessing, such as gene selection. UCE can be applied to any set of protein-coding genes from any species, even if they are not homologs of genes seen during training. UCE learns a universal representation of cell biology that is intrinsically meaningful and can extend insights beyond the data that has been experimentally observed. The representations learned by UCE display an emergent organization of cell types that is consistent with known biology. These cell embeddings can be used to accurately predict cell types with no additional model retraining, showing improved performance in dataset integration against existing atlas-scale integration methods.

UCE presents a novel approach to analyzing cell states. It enables the mapping of new data into a universal embedding space, already populated with annotated reference states. This strategy addresses issues such as noisy measurements that limit data alignment across different experiments, and reduces reliance on small sets of marker genes to translate insights across studies [30]. UCE empowers researchers to utilize existing models on new data without needing data labeling or model retraining. This can foster novel cross-dataset discoveries and overcome the limitations currently faced when working with small, isolated datasets. For instance, a cell type classifier trained to predict specific immune cell types can be seamlessly applied to a completely new dataset. Thus, UCE offers a versatile, efficient, and broadly applicable framework for the analysis of cell states.

Results

A biologically-informed foundation model for single cell gene expression.

Integrating single-cell RNA sequencing (scRNA-seq) datasets is challenging for two primary reasons: scRNA-seq data does not always contain the same genes, or features, and those features are plagued by dataset-specific experimental artifacts or batch effects, which means models have to be built separately for each dataset. UCE overcomes these challenges by abstracting cells as ‘bags

of RNA' [31]. UCE (Fig. 1a) converts the RNA gene expression of a single cell into an expression weighted sample of its corresponding genes. Next, UCE represents the sample's genes by their protein products, using a large protein language model. This allows UCE to meaningfully represent any protein-coding gene, from any species based on sequence alone, regardless of whether the species had appeared in the training data. Finally, after incorporating additional metadata about genes' chromosomal locations, this representation is fed into a large transformer model [32]. UCE is able to map any cell, from any tissue, or any species, into one shared universal space, with no additional training.

In particular, UCE takes as an input (1) scRNA-seq count data and (2) the corresponding protein embeddings, generated by a large protein language model, ESM2 [33], for the genes in the dataset. The ESM2 protein language model takes amino acid sequences as an input and produces a numerical representation called a protein embedding. Given the expression count data for a cell, UCE takes a weighted and normalized sample, with replacement, of the cell's genes. This sample can only contain genes which had non-zero expression, and can contain multiple copies of each gene. These genes are then tokenized by converting them to the protein embedding representation of the protein that they code for [34]. Genes belonging to the same chromosome are grouped together by placing them in between special tokens and are then sorted by genomic location. A special token representing the entire cell, the 'CLS' token, is appended to the beginning of the cell representation [35]. This combined representation is passed into a transformer neural network. The embedding of a cell is taken as the embedding of the CLS token at the final layer of the transformer (Fig. 1a).

UCE is trained in a completely self-supervised manner, and thus does not make use of any cell type or dataset-based annotations. Although including cell type annotations in model architecture or training might seem like a compelling design choice, cell type labels may contain inaccuracies or biases, and including these labels can distort the structure of embedding spaces [36]. Ultimately, it would be preferable if the organization of cells within a model were to emerge naturally, rather than relying on human labeling.

To train the UCE model, a random subset (20%) of genes that were expressed are masked before sampling. These expressed genes are combined with a random subset of genes which had zero expression (non-expressed genes) to form a set of query genes. Each of these query genes' protein embedding tokens is combined with the UCE embedding of the cell they were generated from, and this joint embedding is passed into a fully connected neural network that predicts if that

gene was expressed.

UCE is a 33 layer model consisting of over 650 million parameters. UCE was trained across more than 300 datasets that are largely collected from the CellXGene corpus [37] consisting of over 36 million cells, for 40 days across 24 A100 80GB GPUs (Methods, Extended Data Table 2, Supplementary Table 2). The model's weights and implementation are freely available and the model will be hosted as an openly available resource for the research community to run inference on new datasets.

UCE creates an Integrated Mega-scale Atlas (IMA) of 36 million cells.

We apply UCE to generate an Integrated Mega-scale Atlas (IMA) of 36 million cells sampled from diverse biological conditions, demonstrating the emergent organization of UCE cell representations (Fig. 1b). We find that cells within the UCE space naturally cluster by biological conditions like cell type, while mixing among experimental conditions like batch (Fig. 1b). Since UCE is trained in a self-supervised manner, this organization represents an emergent behavior of the model. The IMA contains numerous cell type alignments, across tissues and species.

To investigate the emergent organization of the IMA, we inspect how tissue residency can influence the state of cell types. Although macrophages found in different tissues are characterized by diverse transcriptional identities [38], they align closely in the UCE space (Extended Data Table 1). For the purpose of our analysis below, we first determine the central location of each cell type and tissue combination in the IMA space, by averaging the UCE embeddings of the cells from that combination, creating a tissue and cell type 'centroid'.

Cells in the IMA have been pre-labeled by their cell type. As these labels were never used for training the UCE model, we use them to validate the quality of the learned representation.

For example, in the IMA, human macrophages are found in 73 different tissues and among these tissues, 72% (53) of tissue-specific macrophage centroids were embedded closest to a macrophage centroid from another tissue. Considering the 3 nearest centroids increases this percentage to 93% (Extended Data Table 1). Similar cross-tissue homogeneity can also be identified in other prolific cell types, like endothelial cells or neurons. This demonstrates that UCE, without any explicit training or labels, identifies that macrophages have a unique cellular identity that is shared across tissues. More broadly, it is an example of UCE's emergent organization that is consistent with known biology even though not explicitly trained for.

UCE embeds new datasets without additional model training.

We evaluated the universality of UCE representations by directly mapping new datasets which were not part of the training set into the embedding, without any additional training or refinement of the UCE model. This is referred to as a 'zero-shot' capability, since the model was never trained on any samples from the new dataset (Fig. 2a). Performance in a zero-shot setting demonstrates a novel and valuable capability, and is necessary to achieve universality. When a model's weights are updated, such as during fine-tuning, its underlying structure changes, and it may overfit to whatever data it is trained on. Because this underlying structure changes, the representations the model learns for data are no longer universal. While a variety of deep learning models have been proposed for this cell embedding task, we choose to compare the performance of UCE to other self-supervised transformer-based methods. This is because they do not rely on cell type annotation, are trained on large datasets, have high model capacity, and can be run in a zero-shot setting. In particular, we compare against Geneformer [28] and scGPT [29]. Geneformer represents cells as lists of genes sorted by their expression, while scGPT represents cells as lists of

genes with binned expression values.

We assess the performance of these methods on a completely new and yet unreleased dataset (as of the publication of this manuscript), Tabula Sapiens v2, which contains diverse human data from 581,430 cells, 27 tissues, 167 batches and 162 unique cell types. Tabula Sapiens is collected in a highly standardized way and represents one of the largest 'gold-standard' expert annotated datasets available. We use established metrics for embedding quality that measure the conservation of cell type information and the correction of batch effects (Methods). We compared several methods and found that UCE substantially outperforms the next best method Geneformer by 13.9% on overall score, 16.2% on biological conservation score, and 10.1% on batch correction score (Supplementary Table 1). To comprehensively assess the value of these zero-shot embeddings, we also compare UCE to fine-tuned methods that are conventionally used for this task. Notably, UCE even performs slightly better than non-zero-shot methods that require dataset-specific training: scVI [15] and scArches [18].

Tabula Sapiens v2 includes cells measured using both droplet or plate-based sequencing methods. Correcting technology-based batch effects, even for fine-tuned models, can be difficult. UCE zero-shot embeddings of the Tabula Sapiens v2 Ovary tissue, which contains $45,757$ cells profiled using 10x-primev3, and $3,610$ cells profiled using Smart-seq3, successfully corrects batch effects at the same level as fine tuned methods, while more accurately representing cell types. When scored using the single cell integration benchmark (SCIB), UCE's batch correction scores are close to that of scVI and scArches, while its bio conservation scores are higher (Supplementary Table 4).

For all cell types in Tabula Sapiens v2, we calculate the silhouette width score of each zero-

shot embedding method. For 67% of cell types, UCE has the highest silhouette score of any method (Extended Data Table 8). UCE outperforms Geneformer on 80% of cell types, tGPT on 73% of cell types, and scGPT on 83% of cell types. Notably, UCE accurately embeds B cells, while Geneformer and scGPT fail to do so (Supplementary Fig. 2a). In Tabula Sapiens v2, the silhouette width score of B cells is 93% higher in UCE versus scGPT and 25% higher versus Geneformer. Additionally, B cells within the UCE embedding space can be accurately mapped to an existing reference. We train a simple logistic classifier on the UCE embeddings of the Immune Cell Atlas [39], and then apply the classifier to B cell embeddings from Tabula Sapiens v2. This classifier accurately classifies the Tabula Sapiens v2 cells as memory and naive B cells (Supplementary Fig. 2b), which is confirmed with marker gene analysis (Supplementary Fig. 2c).

To further compare the performance of UCE and other foundation model approaches, we evaluate embedding quality on a complex dataset featuring 382 cell type clusters, the Human Brain Atlas [12]. UCE outperforms Geneformer and scGPT, and can be used to further investigate diverse cell types, like "Splatter neurons", by comparing them to the IMA (Supplementary Note 6). Overall, these results illustrate that UCE has the unique capability to meaningfully integrate new, previously unseen datasets into a universal cell representation space with no additional model training.

UCE embeds diverse cell types from organisms that were not part of the training data.

UCE is also able to align datasets from novel species without additional model training. This is due to the fact that UCE is not dependent on any particular genome—each gene of interest is translated to a corresponding protein sequence, which is then embedded in a universal protein space. The representation in this space is independent of species and importantly does not require

any judgment about whether particular pairs of genes are homologs or not. Since UCE can analyze cell atlas data from distinct species that were not part of the training set, the extent to which it succeeds in this task is a stringent test of whether UCE displays emergent behavior.

UCE's training data is composed of datasets from eight species: human, mouse, mouse lemur, zebrafish, pig, rhesus macaque, crab eating macaque and western clawed frog. We apply UCE to embed datasets from three novel species that were not included in the training set. For each species, we generate a zero-shot embedding and then determine the nearest cell type centroid from the IMA for each of the dataset's existing annotated cell types. For all three species we observed very high agreement between independent annotations of the novel species' data and the nearest cell type centroids in the IMA.

Within a dataset of green monkey lymph node and lung cells [40], for 13 of the 17 cell type centroids, the closest centroid from another species corresponds to the same cell type in the green monkey. This match extends to all 17 centroids when considering the three nearest centroids (Extended Data Table 1, Fig. 2c, 2d). Moreover, a population of lymph node cells that were originally labelled as B cells, form a distinct cluster in UCE space (Supplementary Fig. 3b). Differential expression analysis revealed that this cluster predominantly expresses a T cell marker, *Cd3d* (Supplementary Fig. 3a, 3c).

In the case of naked mole rat spleen and circulating immune cells [41], for 17 out of 24 cell types, the nearest cross species centroid matches the naked mole rat cell type (Extended Data Table 1, Supplementary Fig. 4b). In the case of chicken, we embed two distinct chicken datasets, chick retina [42] and developing chick heart [43] (Supplementary Fig. 5a, 5b). Different eye-specific neurons within the chick retina map to mouse lemur neurons, such as chick oligodendrocytes,

which are closest to mouse lemur oligodendrocytes (Extended Data Table 1). In chicken heart, 12 of 15 cell type centroids are matched within the nearest two cross species centroids (Extended Data Table 1). No bird species were included when training UCE.

We compare UCE's performance to a state-of-the-art method for cross-species label transfer, SATURN [34]. For each of the four datasets, we integrate them with a paired human dataset from the same tissue from Tabula Sapiens v1, except for chicken retina, which was paired with retina data from Orozvo et al [44]. In comparison to UCE, which is zero-shot and unsupervised, SATURN is a fine-tuned method and weakly-supervised: it is aware of each species' cell type annotations individually. Still, UCE outperforms SATURN for cell type label transfer between the human and the new species for three out of four datasets, chicken heart, chicken retina and naked mole rat spleen (Supplementary Note 5).

While UCE successfully integrates data from these new species, it still has limitations when assessing species that are more evolutionarily distant from the species in the training data. We create a zero-shot embedding of Tabula Drosophilae [45], a large multi-tissue fly cell atlas, and evaluate UCE's performance by transferring labels from Tabula Sapiens v1 and v2 to the fly cell atlas using cell type centroids. Predicted fly cell types are inaccurate using this method: many fly cells are incorrectly annotated as fat cells, including adult epithelial cells, adult muscle cells and adult trachea cells. Additionally, other adult muscle and epithelial cells are incorrectly annotated as leukocytes (Extended Data Table 4).

Altogether, these results highlight that UCE can be directly applied to investigate new and diverse datasets from previously unobserved species. While model performance does not necessarily extend to very distant species, such as fly, inclusion of more species from across the tree of

life could help solve this issue; both by improving models' performance as well as by providing a more appropriate reference to compare data to.

UCE learns a meaningful organization of cell types in previously unseen data.

Moving beyond metrics focused on individual cell type clusters, we also examined the structure of the universal embedding space as a whole, through the relative positioning of different cells within it. A meaningful arrangement of cell types emerges upon embedding all the cells from the Tabula Sapiens v2 dataset from the lung tissue (Fig. 3a). Not only do distinct cell types like T cells, monocytes and endothelial cells cluster together, but higher-level categories, such as immune cells and epithelial cells, are also clearly distinguished. When compared to the cell hierarchy derived using the Cell Ontology [46], UCE identified clusters showed higher similarity (as measured using Adjusted Rand Index) than other zero-shot embedding methods (Supplementary Fig. X).

To further assess the organization of all cells within the embedding, we compared distances between pairs of cell types across all tissues in the embedding space to their distances in the Cell Ontology tree (Fig. 3b). We hypothesized that cells that are known to be similar based on the cell ontology would likely also be closer together in the embedding space, and that the degree of closeness would be correlated with ontological similarity. The results validate this relationship: at each additional unit of separation between cell types in the cell ontology tree, there is a significant increase in the embedding distance in UCE between those cell types. We consistently observed this trend up to a distance of 5 hops in the ontology tree (Fig. 3b). However, beyond that, the effect levels off (Supplementary Fig. 6). This is expected due to the curse of dimensionality in high-dimensional spaces and the variability in the level of ontological refinement in different branches of the ontology (Supplementary Note 3).

We also noted significant colocalization among cells originating from the same developmental lineages, in particular from the mesoderm, endoderm, and ectoderm germ layers. For Tabula Sapiens v2, 90 out of 97 of the centroids for mesoderm-derived cell types had other mesoderm-derived cell type centroids as their closest neighbors. A similar pattern was observed for 46 of the 56 endoderm-derived cell types and 22 of the 30 ectoderm-derived cell types (Supplementary Fig. 7a). A neural network classifier trained to predict the germ layer of origin for individual held-out cell types using their universal embeddings showed an accuracy of over 80% (Supplementary Fig. 7b). When mapping unseen data from the bone marrow, we observe a clustering of cell types that aligns well with both developmental stage—where early progenitors are closer to each other in the embedding space—and lineage, with more developed cell types in the lymphoid and myeloid lineages clustering distinctly. (Supplementary Figure 16)

The accuracy of cell type organization in the Tabula Sapiens v2 lung dataset was evaluated by comparing it with other lung datasets in the IMA (Fig. 3c, Supplementary Fig. 8). Four different endothelial cell subtypes are observed to map correctly to their corresponding counterparts in the IMA. Similarly, lung ciliated cells correctly map to their counterpart in the larger corpus despite the presence of four different ciliated cell subtypes (Fig. 3c). Further analysis of the alignment of cell type centroids between Tabula Sapiens v2 and the IMA across all tissues showed an exact alignment for 41% of cell types (Methods). This alignment, based on the three nearest neighbor cell type centroids, is 65% more accurate compared to that measured in the original gene expression space (Fig. 3d). When focusing on the single nearest centroid, the alignment accuracy improves by 92%. Many inexact matches are caused by differences in the labeling resolution of different datasets in the IMA (Supplementary Tables 5, 6, Extended Data Table 9). These results

demonstrate that UCE can effectively learn a universal representation of cell biology that not only enables discrimination between individual cell types but also captures their relative similarities across scales with the potential to reveal deeper insights into development and function.

A workflow for decoding the function of newly discovered cell types.

UCE's zero-shot embedding capabilities unlock novel computational analyses of scRNA-seq data and aid in hypothesis generation. Beyond identifying novel cell type clusters, UCE differs from other methods in that the same cell type can also be easily compared against all previously assayed cells across tissues, disease states and species. Moreover, UCE is not biased in this process by existing annotations, opening the door for discovery of novel function (Fig. 4a). With existing fine tuning based methods, every searched dataset would need to be integrated, requiring repeated model retraining. Thus, UCE enables a new workflow for scRNA-seq data analysis that performs an unbiased search across the universe of cell biology.

We present an example of this analysis by using the recently identified kidney Norn cell as a case study. The kidney Norn cell is the long-sought erythropoietin (*Epo*) producing cell in the kidney, and is characterized as fibroblast-like. We perform a zero-shot embedding of mouse renal cells from [47], which produces a cluster of cells corresponding to Norn cells (Fig. 4b).

Using a simple logistic classifier trained on the embedding of mouse renal cells, we predict the existence of Norn cell clusters in many kidney datasets. Since this classifier takes universal cell embeddings as an input, we can directly apply it to all 36 million cells in the IMA, in a manner unbiased by cell type annotations ascribed by previous studies. To evaluate the model's predictions that these are Norn, or Norn-like cells, we investigate the expression of canonical Norn cell marker genes. Cells classified as Norn cells in the top 13 kidney datasets by Norn abundancy demonstrate

preferential expression of the Norn markers *Dcn*, *Lpar1*, *Col1a1*, *Cxcl12*, and *Cfh* (Extended Data Table 3). Notably, *Epo* transcripts, which are often missing from datasets and lowly expressed, are not typically differentially expressed in these cells. *Epo* cannot readily be used as a marker gene for Norn cells because of a variety of factors, including low overall expression, even in hypoxic Norn cells, the bursty nature of *Epo* transcription, and the fast degradation of *Epo* messenger RNA upon reoxygenation [47]. *Cxcl14*, another marker of Norn cells, displays mixed expression patterns in these predicted Norn cells (Fig. 4c). The same pattern of marker gene expression is also found in cells from other tissues, including lung and heart datasets (Fig. 4c). Additionally, these predicted cells also share a common set of genes that are lowly expressed in mouse renal Norn cells (Supplementary Fig. 9). The tissues with the highest number of predicted Norn cells were gonad, heart and lung. Finally, we also identify differences in gene expression specific to each tissue (Supplementary Figure 10). While *Epo* expression has been previously observed in the heart and lung tissue, the mechanisms and cell types associated with this expression, and their relation to kidney Norn cells have not been previously determined [48]. Overall, this demonstrates that UCE can serve as an unbiased tool for predicting novel similarities between cells.

UCE helps interrogate alternate lung disease outcomes.

Lastly, we apply UCE and our simple Norn cell classifier to investigate Norn-like cells in lung diseases. We generate an embedding of lung cells sampled from patients with idiopathic pulmonary fibrosis (IPF), chronic obstructive pulmonary disease (COPD), or patients from a control group [49]. We identify Norn-like lung cells that preferentially express Norn markers in all three groups (Fig. 4d).

For these Norn-like lung cells, we identify differences across disease groups (Fig. 4e). COPD

and IPF are both associated with elevated bloodstream *Epo*, but COPD has levels higher than IPF. Additionally, in patients with IPF, secondary erythrocytosis is absent or reduced compared to patients with COPD [50–52]. Given the identification of Norn-like cells in the lung, and Norn cell's production of *Epo*, it is possible that this difference in disease prognosis could be related to disease associated differences in Norn-like cells.

COPD predicted Norn-like cells express genes (*Lum*, *Crispld2*) involved in glycosaminoglycan pathways at higher levels than IPF or control predicted Norn-like cells [53,54]. Norn-like cells in both COPD and IPF express *Gpx3* and *Igfbp6* at significantly lower rates than the control group (Figure 4d, Extended Data Table 7).

Taken as a whole, these results indicate that cells with similar transcriptional states to Norn cells can be found in other tissues in the body, and such cells may play a previously undescribed role in disease. UCE greatly facilitates an analysis of this scale and diversity because it is a universal model that is agnostic to tissue, species or disease state. In order to ground this analysis, results are confirmed using expression of canonical marker genes. However, as universal analyses become more complex, it is possible that such assessments might be made based solely on the predictions of foundation models.

Discussion

UCE is a single-cell foundation model that is built from the ground up to represent cell biology across the wide array of single-cell datasets. We envision UCE as an embedding approach that enables researchers to map any new data, including entire atlases, into an accurate, meaningful and universal space. The embedding space that emerges from UCE is highly structured and diverse and aligns cell types across tissues and species. Additionally, these cell types organize themselves in a

pattern that reflects existing biological knowledge.

The UCE model has broad implications for the creation of large foundation models for single cell biology. For large foundation models to be truly useful for scientific discovery, they must have unique qualities that distinguish them from existing methods. Zero-shot embeddings are one such important capability because it enables an intrinsically meaningful representation that can extend insights beyond the data that has already been observed and annotated experimentally. Our results demonstrate that UCE can achieve such a generalizable representation across different datasets while maintaining accuracy on individual datasets, comparable to methods that require retraining for each specific dataset. To ensure that UCE's embeddings, or any downstream models trained on them, remain universally shareable, we have designed UCE to be used as-is, without any model fine-tuning. This allows UCE representations to serve as a universal way of connecting any single-cell gene expression dataset.

By building UCE, we enable novel analyses of scRNA-seq data. In order to establish the universality and accuracy of UCE, we perform a number of different benchmarks, mainly focused on correctly identifying and matching expert-annotated cell types. However, these analyses are still far from perfect, as they are generally limited by the resolution of cell type labels. To better understand single-cell foundation models, and especially how they scale, new analyses and benchmarks that surpass this resolution limit should be developed. Such benchmarks could focus on measuring a model's ability to encode response to perturbation or to integrate data from different modalities. Because of their "black box" natures, foundation models for biology represent a further step away from explainable predictions. In addition to benchmarks, which can point to what models performs the best, new interpretability tools must be developed to explain why these models make

their predictions [55–58].

UCE establishes a framework in which physical scales of biology, learned by large machine learning models, are directly connected. In this framework, molecular representations, embeddings of proteins based on sequence, are aggregated and used to construct representations of cells. Such an approach maximizes generalizability and the scope of genomic data that models can be trained on. Since these scales are connected, improvements in modeling in one scale can help improve results in the next. For example, by creating a model which could jointly represent molecules like proteins, RNA and DNA, one could enable a new cell based model to encode additional complex features, e.g. non-coding RNAs, regulatory DNA, alternative splicing, RNA velocity etc. Finally, current scRNA-seq foundation models, including UCE, do not account for any information contained in the raw RNA transcripts. By aligning these transcripts to the reference genome, vital data on genetic variation and crucial RNA-splicing processes are discarded [59]. As these models adopt more biologically-relevant features, they will increasingly be able to simulate the biological processes of cells, leading to the creation of "Virtual Cells".

In 2002, Nobel laureate Sydney Brenner identified many of the core motivations for the creation of cell atlases and virtual cells. Virtual cells should be the goal of biological foundation modeling, because cells are the "real units of function and structure in an organism" [60]. Brenner also identified the need for such models to be computationally efficient, predictive, and able to generate new cell types. We believe that UCE represents a significant advancement in the progress towards a virtual cell. Through learning a universal representation of every cell state and type, we expect that UCE will be a valuable tool for analysis, annotation and hypothesis generation as the scale and diversity of single-cell datasets continues to grow.

Methods

Overview of UCE.

UCE (Universal Cell Embedding) is a machine learning model for mapping single-cell gene expression profiles into a universal embedding space, denoted as $\mathcal{U}$. In this space, each cell c_i is represented as a d_{emb}-dimensional vector, where $d_{emb} = 1280$.

The model takes as input a dataset $\mathcal{D}$ with N cells $\{\mathbf{c}_i\}_{i=1}^{N}$. Cells in $\mathcal{D}$ can be drawn from one or more distinct scRNA-seq experiments. Each cell c_i in $\mathcal{D}$ is described by a gene expression vector $\mathbf{x}^i \in \mathrm{N}^{K_i}$, where K_i is the number of genes measured in c_i and can differ across $\mathcal{D}$. The gene expression vectors $\mathbf{x}^i \in \mathbb{N}^{K_i}$ are not subset to those with high variance. UCE defines a function $f_u : \{\mathbb{N}^{K_i} \rightarrow \mathbb{R}^{d_{emb}}\}_{i=1}^{N}$ that maps each gene expression vector $\mathbf{x}^i$ to its cell embedding vector $\mathbf{h}^i$.

Model input: Gene representation.

The expression of gene g in cell c_i is denoted by x_g^i, where g represents any protein-coding gene. The corresponding token embedding p_g is a pretrained embedding for the protein(s) encoded by the gene g. These embeddings are derived from a pretrained protein language model that takes an amino acid sequence as input and returns a d_p-dimensional embedding vector as output. To create p_g, we take the average of all proteins coded by gene g. In the context of UCE, we can formulate this as a dictionary that maps each gene g to a d_p-dimensional protein embedding vector. Specifically, we employ the ESM2 model, which yields embeddings of size $d_p = 5120$ [33, 34].

Protein language models are chosen for gene representation because they can generate universal representations of any protein sequence. Therefore, for new species, all that is required is to have the amino acid sequence of that species' protein coding genes. These genes do not need to have solved structures and orthology does not need to be calculated between them and the ex-

isting training data. Model training ablations demonstrate that among a class of different protein language models, UCE models trained with ESM2-15B performed the best (Supplementary Note 4, Supplementary Table 3).

Model input: Cell representation.

For each cell c_i in the input dataset $\mathcal{D}$, we identify two distinct sets of protein-coding genes: the expressed genes G_i^+ and the non-expressed genes G_i^-. These sets are defined as follows:

$$G_i^+ = \{g \mid x_g^i > 0\} \tag{1}$$

$$G_i^- = \{g \mid x_g^i = 0\} \tag{2}$$

For producing the cell embedding, a multi-set of 1024 non-unique genes G_i^s are sampled from the expressed genes G_i^+, with replacement. The probability of sampling a gene $g \in G_i^+$ is weighted by the log normalized expression of that gene, which can be formulated as:

$$P(g \mid c_i) = \frac{\log(x_g^i + 1)}{\sum_{g \in G_i^+} \log(x_g^i + 1)} \tag{3}$$

where x_g^i is the expression count of gene g in cell c_i, and the sum in the denominator is over all genes in G_i^+.

Once the multi-set G_i^s is compiled for each cell c_i, we arrange the genes within each chromosome according to their genomic positions. Different chromosomes are specified using special chromosome start and end tokens. Start tokens are unique to each chromosome and species. Every chromosome group is combined into a single sequence, with chromosome order randomly determined. A cell-level CLS token is appended to the start of the sequence. It is designed to capture

the cell-level embedding upon training the model. The final sequence of genes ordered by genomic location and separated by chromosome is referred to as the cell sentence S_i for cell c_i.

Ordering the genes using this information appears to have a moderate negative effect on model performance (Supplementary Note 4, Supplementary Table 3). Transformer models do not require any ordering of their inputs, and this metadata was only added to the model in order to enable potential applications, such as examining chromosomal aberrations or synteny between species. Future models of this type may choose to exclude this metadata, or include it using other methods besides positional encodings.

Transformer Architecture.

Each cell sentence S_i is fed into a transformer that consists of n_{lay} layers. Each layer contains a multi-head self-attention mechanism with n_{head} attention heads and a feedforward network operating over a hidden space of dimensionality d_{hid}. We also initialize sinusoidally-varying positional embeddings. Gene token embeddings are compressed using a single layer MLP to d_{emb}-dimensional vectors before passing through the transformer.

While the transformer architecture is highly performative, it requires significant computational cost to train and evaluate. One important aspect of this cost is the quadratic increase in the runtime of transformers' attention operations proportional to the number of tokens. Such a relationship requires that the number of genes sampled by UCE, 1024, be relatively limited, especially given that individual cells might have more than 1024 uniquely expressed genes. Future model architectures could explore increasing this context length, or using models with sub-quadratic scaling such as state space models [61]

Model output: Cell embedding.

The final output from the model is the cell embedding vector $\mathbf{h}^i_{cell} \in \mathcal{U}$ which corresponds to the d_{emb}-dimensional embedding of the CLS token in the final layer of the model following decoding with an additional MLP.

When cells with unrealistic random gene expression patterns, such as those generated by shuffling the expression of real cells, are inputted, the resulting cell embeddings default to a heterogeneous, out of distribution output (Supplementary Figure 11).

Model training: Cell representation.

At the time of training, we generate a set $G_i^{M+} \subset G_i^+$ by randomly selecting a certain percentage (r_{mask}) of genes from G_i^+, without replacement. This set is used for computing the loss during training, and is masked from the cell representation.

The probability of sampling a gene $g \in G_i^+ \setminus G_i^{M+}$ (Equation 3) is then updated to be:

$$P(g \mid c_i) = \frac{\log(x_k^i)}{\sum_{g \in G_i^+ \setminus G_i^{M+}} \log(x_j^i)}, \tag{4}$$

We also establish two additional gene sets to be used for loss computation: $G_i^{L+} \in G$ and $G_i^{L-} \in G$. G_i^{L+} and G_i^{L-} are randomly selected from the masked set of expressed genes G_i^{M+} and the set of unexpressed genes G_i^- respectively. Both G_i^{L+} and G_i^{L-} are of equal size, specifically $N_{loss}/2$. In the case of G_i^{L-}, the sampling is done without replacement unless $|G_i^-| < N_{loss}/2$. Similarly G_i^{L-}, is also sampled without replacement unless $|G_i^{M+}| < N_{loss}/2$. In this case, G_i^{M+} is used as-is along with additional samples drawn with replacement from the full set of expressed genes G_i^+.

Model training: Loss Function.

UCE is trained to reconstruct the binarized expression of genes in a cell, when those genes

are masked from the model by setting their expression to 0. To calculate the loss function for a given cell c_i, the cell embedding vector $\mathbf{h}^i_{emb}$ is individually concatenated with every gene g within both G_i^{L+} and G_i^{L-}. These concatenated vectors then serve as input to a feedforward multilayer perceptron (MLP), which computes the probability that gene g is expressed within cell c_i.

$\mathbf{h}^i_{cell}$ represents the embedding vector for cell c_i and p_g represents the token embedding for gene g. Then the concatenated vector $\mathbf{z}^i_g$ that serves as input to the MLP for cell c_i and gene g is:

$$p'_g = \text{MLP}(p_g) \tag{5}$$

$$\mathbf{z}^i_g = [\mathbf{h}^i_{cell} || p_g\prime] \tag{6}$$

where $||$ denotes the concatenation operation and p'_g is the compressed protein embedding.

The MLP then processes this concatenated input to produce the predicted probability that gene g is expressed:

$$p(y^i_g) = \text{MLP}(\mathbf{z}^i_g) \tag{7}$$

This probability is then used in the binary cross-entropy loss function. The true classification labels for each gene's expression status in cell c_i are represented by the vector $\mathbf{y}^i$. UCE is trained to accurately predict the expression of genes in G_i^{L+} and the lack of expression in G_i^{L-}. The model is trained using a binary cross-entropy loss, which is averaged across all N_{loss} genes and all N cells in the minibatch as follows:

$$L = -\frac{1}{N}\sum_{i=1}^{N}\frac{1}{N_{loss}}\sum_{j=1}^{N_{loss}}\left[y^i_j \log(p(y^i_j)) + (1-y^i_j)\log(1-p(y^i_j))\right] \tag{8}$$

For further details on hyperparmeter choices please see Supplementary Table 2. Different hyperparameter choices for this masked loss can impact final model quality. Reducing the masking percentage from 20% to 10% improved model performance, while increasing it to 40% degraded performance (Supplementary Note 4, Supplementary Table 3). In future models, an alternate loss, for example, a generative decoder style loss, could be used, and might improve performance since no genes would be masked during training.

Creating the IMA and dataset preprocessing.

The Integrated Mega-scale Atlas (IMA) used to train UCE was created by combining scRNA-seq datasets from multiple publicly available sources. The majority of IMA data (33.9 million cells and 285 datasets) is human and mouse data downloaded from CZI Cell X Gene (CxG) Census [37] version "2023-07-10" (July 10th, 2023). Duplicate cells were removed by selecting primary cells only. The remainder of the IMA is composed of 2.3 million cells from 28 datasets, from eight different species: human, mouse, zebrafish, rhesus macaque, crab-eating macaque, mouse lemur, frog, and pig.

For datasets from the CxG Census, preprocessing only involved filtering cells by minimum gene counts (200) and genes by a minimum cells count of 10. No highly variable gene selection was applied. For datasets collected from other sources, preprocessing was not uniform (Supplementary Note 1).

For visualization of the IMA (Fig. 1b), predicting green monkey cell types (Fig. 2d), matching new species centroids (Extended Data Table 1), and prediction of Norn-like cells (Fig. 4, Supplementary Fig. 9) a representative sample of the IMA was used in place of the full 36 million cells. This representative sample was used in order to speed up computationally intensive tasks like

UMAP calculation. The sample was created by randomly choosing 10,000 cells from each dataset, without replacement. For datasets with fewer than 10,000 cells, the entire dataset was included. In total, this representative sample has 2,969,114 cells. The average number of cells per dataset in the sample is 9486. For visualization and centroid calculation, cell types in the sample were coarsened by mapping them to a set of 51 coarse cell types (Extended Data Table 6).

Model Evaluation.

- **Zero-shot embedding quality and clustering** For evaluating the quality of embeddings, we used metrics from the single-cell integration benchmark [16].

- **Cell type organization** For each cell type dendrogram the Euclidean distance was used to perform hierarchical clustering across all cells.

- **Comparison to cell ontology** Here, we used the tree distance between any two cell types in Cell Ontology [46]. We used the most up to date version of Cell Ontology at the time of writing this paper (Release date: $2024/08/16$). To determine the Euclidean distance distribution, we sampled 100,000 random pairs of cells from Tabula Sapiens v2.

- **Zero-shot cell type alignment to IMA** For each cell type θ, a centroid was identified separately for data from Tabula Sapiens v2 (TSv2) c_θ^T and from IMA c_θ^I. For each cell type that is present in both TSv2 and the IMA, the 3 nearest neighbor cell type centroids $\mathbf{N}_\theta^T$ to the centroid in Tabula Sapiens c_θ^T were identified. These neighbors could be either from Tabula Sapiens or from the IMA.

  If this set of neighbors $\mathbf{N}_\theta^T$ to the anchor centroid from TSv2 data c_θ^T contains the centroid for the same cell type in IMA data c_θ^I, then this was counted as a correct match.

This analysis was performed per tissue, both in the UCE embedding space as well as in the original expression space (after log-normalization). In case of the original data representation, the set of 5704 shared genes across all 184 human datasets in the Integrated Mega-scale Atlas (IMA) were used to represent each cell. The expression counts in each cell are normalized by total counts over all genes, so that every cell has the same total count after normalization. This is followed by a log transformation. This follows standard Scanpy data preprocessing guidelines [62].

Description of other models.

- **Geneformer** Geneformer [28] is a transformer based foundation model. Geneformer represents a cell as a list of genes sorted by their expression. Geneformer was trained using masked language modeling, on 30 million cells. For all analyses, the *GF-12L-30M-i2048* model was used.

- **tGPT** tGPT [63] is a transformer based foundation model. tGPT represents cells, like Geneformer, as a list of genes sorted by expression. However, tGPT is trained as an autoregressive langauge model, rather than a masked langauge model. tGPT was trained on 22.3 cells.

- **scGPT** scGPT [29] is a transformer based foundation model. scGPT represents cells as a list of the expression values of its genes. scGPT is trained to iteratively decode those genes' expressions when some are masked, using generative pretraining. scGPT was trained on 33 million cells. For all analyses, the *whole-human* model was used.

- **scVI** scVI [15] is a variational autoencoder. scVI is trained to reconstruct gene expression values using a zero-inflated negative binomial loss. scVI was used with default parameters

for all experiments.

- **scArches** scArches builds on top of the scVI architecture to enable transfer learning of cell types. As used in this work, scArches first trains an scVI model on a reference dataset, fine-tunes a cell-type aware scANVI model using the cell types from the reference dataset, and then further finetunes this model on a transfer dataset using architecture surgery. scArches was used with default parameters for all experiments.

Differential expression analysis of predicted Norn cells.

A logistic classifier was trained to predict cell types from UCE embeddings on mouse kidney cells. This classifier was then applied to UCE embeddings from the representative sample of IMA datasets. Datasets were then split by tissue, and the datasets with the most predicted norn cells in each tissue were used for differential expression analysis. The top 13 kidney datasets, top 6 lung and top 6 heart datasets were chosen.

For each individual (full) dataset, RNA counts were log normalized, and then differential expression was run using default settings as implemented in Scanpy [62], comparing predicted Norn cells to all other cells in the dataset. The results of these differential expression tests were used to determine the log fold change of marker genes in predicted Norn cells (Fig. 4c, Supplementary Fig. 9).

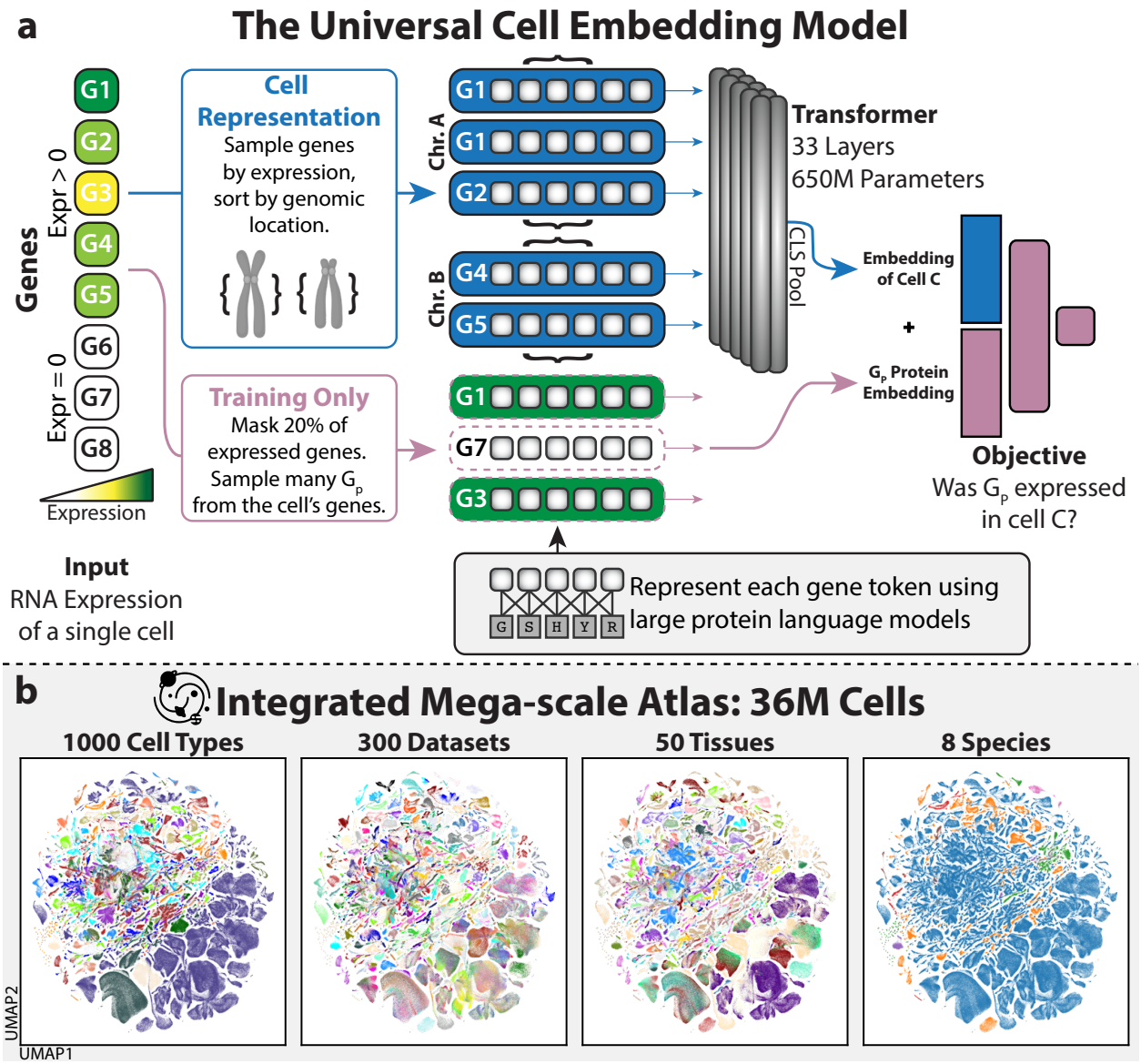

Figure 1: The Universal Cell Embedding Model is a large foundation model for single cell biology (a) Overview of the Universal Cell Embedding (UCE) model. UCE has a unique, biologically motivated representation of cells (blue) and training scheme (purple). Given the gene expression for a single cell, UCE samples with replacement genes that were expressed, weighted by their level of expression. Each gene is represented using a 'token' corresponding to its protein product. Gene tokens are represented numerically by using ESM2 protein embeddings, a 15 billion parameter protein language model that takes amino acid sequences as an input. The gene tokens are sorted by genomic location and grouped by chromosome. Chromosome groups are delineated by specific chromosome start tokens and end tokens, joined, and then passed into a transformer neural network. The embedding of the cell is determined by taking the final layer output of a special CLS token that is appended before all the other tokens. To train the UCE model, a portion of genes that were expressed are masked. The model next combines the protein embeddings corresponding to each of these genes with the embedding of the cell, and passes this joint representation through a neural network that predicts if a given gene was expressed in the cell or not. This objective function is then used to update the weights of the model. **(b)** UMAP visualizations of the universal cell embedding space. We apply UCE to embed 36 million cells, with more than 1,000 uniquely named cell types, from hundreds of datasets, dozens of tissues and eight species, creating an Integrated Mega-scale Atlas (IMA) spanning the universe of cell biology.

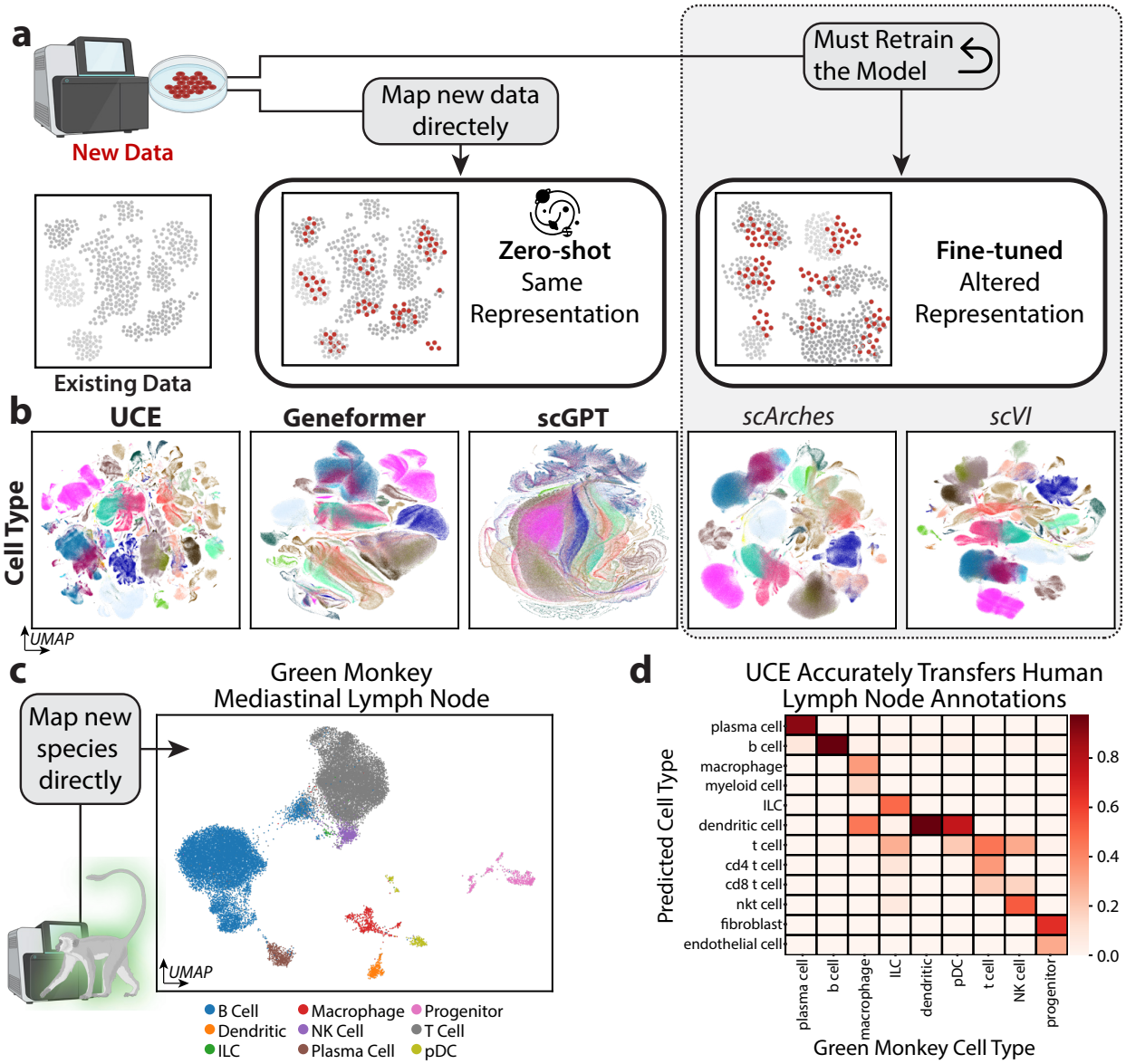

Figure 2: Zero-shot cell embedding capabilities of UCE (a) Comparison of zero-shot and fine-tuned single-cell embedding models. A zero-shot embedding model maps new data directly to the to the representation space, with no additional model training. In contrast, fine-tuned models must first be retrained on a given dataset, and only then can be applied on that dataset, fundamentally altering the model's representation space. **(b)** UMAP embeddings of UCE and other methods for Tabula Sapiens v1 and v2, colored by cell type. UCE zero-shot embeddings outperform other methods when scored using metrics from the single cell integration benchmark (Supplementary Table 1) [16]. **(c)** UMAP of cells from a new species, green monkey colored by cell type. UCE is able to generate high-quality zero-shot embeddings of novel species that were never seen during training. The UCE embedding for green monkey mediastinal lymph node [40] recaptures cell type clusters. Notably, a population of B cells (blue) clusters nearby to T cells, potentially due to expression of *Cd3* (Supplementary Fig. 1). **(d)** Green monkey lymph node cells can be accurately annotated using the IMA. A logistic classifier is first trained to predict cell types based on UCE embeddings of human lymph node cells. The classifier is then directly applied on green monkey cells to predict the cell types. Predicted cell types have high agreement with the original cell type annotations, demonstrating that UCE can be used to transfer cell type annotations to novel species.

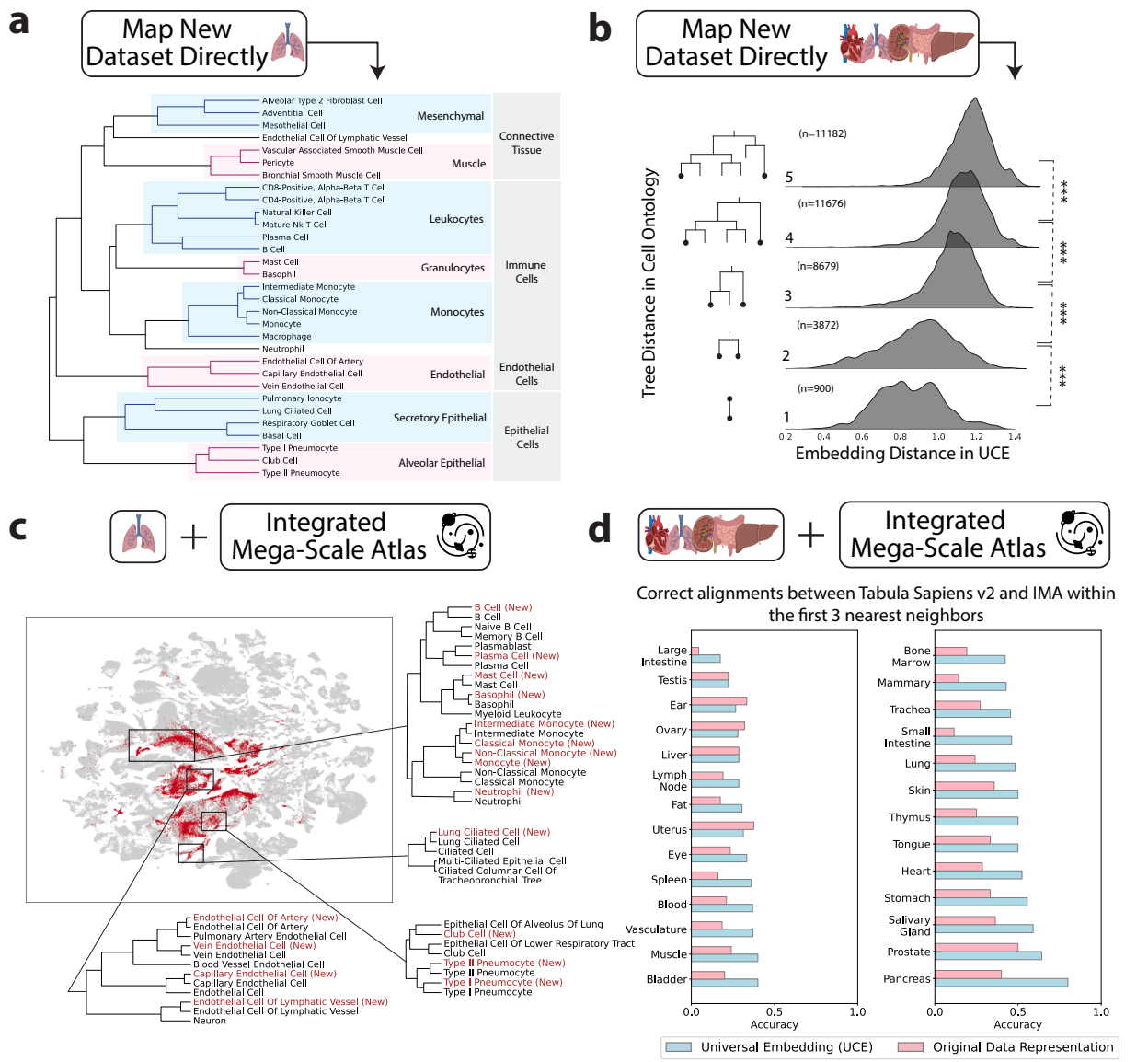

Figure 3: UCE learns meaningful organization of cell types (a) The UCE space generated for new, previously unseen data shows a meaningful arrangement of cell types. Lung data was used from new donors from the Tabula Sapiens Consortium. Dendrogram of hierarchical clustering of all annotated cell types in the UCE embedding space. Closely connected cell types in the dendrogram show meaningful relationships both at finer and coarser scale resolutions. **(b)** Evaluation of the organization of cell types in the embedding space when compared to Cell Ontology. The x-axis depicts the density of Euclidean distances between all pairs of cells across all tissues for these new donors from the Tabula Sapiens Consortium. The y-axis shows the corresponding tree distance between cell types as found in the Cell Ontology. Stars denote statistical significance, which was established using a one-sided t-test, p values in increasing order of distance: $10^{-106}, 10^{-201}, 10^{-25}, 10^{-131}$ **(c)** Mapping data from new donors to the Integrated Mega-scale Atlas (IMA) across multiple lung datasets. Red labels correspond to data from new donors, grey are from IMA datasets. All cell type labels from multiple datasets are displayed as-is, with no modifications or reformatting of text. Accurate alignment between the new dataset and IMA is observed at finer resolution. Four different subtypes of endothelial cells are shown to correctly map to their corresponding counterparts in the complete mega-scale atlas. In the case of lung ciliated cells, they map more closely to their matching counterpart as compared to all other ciliated cell subtypes also present in the IMA. **(d)** Quantification of cell type alignment between new dataset and IMA. Accuracy in 3-nearest centroid matches between new dataset and IMA cell types at the finest level of original annotation. Results are measured across all 27 tissues in Tabula Sapiens v2 for both the UCE space and the original gene expression space. Tissues are ordered by accuracy in the UCE space.

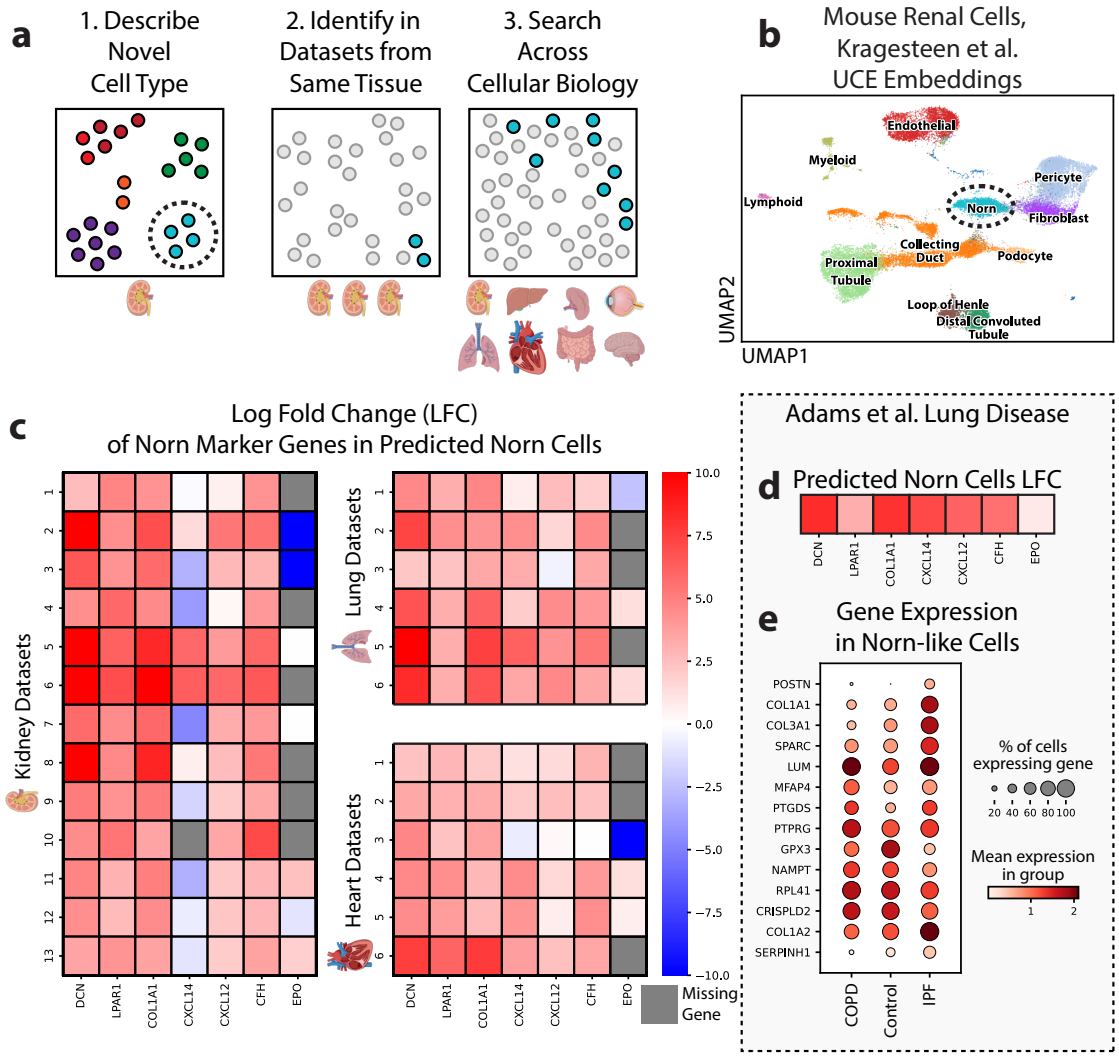

Figure 4: Norn Cell Case Study: UCE unlocks new analyses of single cell datasets (a) Overview of a novel single cell analysis workflow that UCE facilitates. Analysis begins with (1) the identification of a novel cell type (circled) within the embedding space, using methods such as clustering and confirmation using marker gene analysis. (2) Next, the novel cell type can be easily identified in other datasets profiled from the same tissue (for example, kidney). A simple classifier, such as a logistic classifier, is trained to predict cell types from universal cell embeddings, and is then applied to embeddings from other datasets of the same tissue (kidney), to confirm the cell type's existence and improve its characterization. (3) Finally, the same simple classifier can be applied to the embeddings of cells from any other tissue, to find cell types with similar biological functions or patterns of gene expression. **(b)** Identification of novel Norn cells in mouse kidney. UMAP visualization of zero-shot embedding of mouse renal cells from Kragesteen et al. [47]. Norn cells form a distinct cluster within the embedding space (circled). **(c)** Identification of Norn cells and Norn-like cells across tissues. A logistic classifier is trained to predict Norn cells from universal cell embeddings, and is then applied to other kidney datasets (left) and datasets from lung and heart (right). The log fold change of known Norn marker genes between cells predicted to be Norn cells and the remaining cells within each dataset is visualized. Cells which are predicted to be Norn-like preferentially express Norn markers in kidney, as well as in lung and heart. Notably, *Cxcl14* has a mixed pattern of expression among some datasets. **(d)** Cells predicted to be Norn-like cells within a lung disease dataset [49] express known Norn markers, as demonstrated by log fold change (LFC). **e** Differential gene expression in predicted Norn-like cells, grouped by disease status. There are significant differences in gene expression between cells from IPF, COPD and control patients. Patients with IPF and COPD are known to have elevated levels of blood stream *Epo*, with COPD patients having greater bloodstream *Epo* levels than patients with IPF.

Data availability

The full list of datasets used to train UCE are in Extended Data Table 2. Most of these datasets are available to download from CellXGene [37]. Tabula Sapiens v2, used for model evaluation, will be made available upon publication.

Datasets analyzed in the paper are publicly available to download. The green monkey lung and lymph node dataset is available with accession code GSE156755. The naked mole rat dataset is available with accession code GSE132642. The chicken retina dataset is available with accession code GSE159107. The chicken heart dataset is available with accession code GSE149457. The mouse kidney dataset is available with accession code GSE193321. The human lung disease dataset is available with acccesion code GSE136831.

Code availability

UCE was written in Python using the PyTorch library. The source code is available on Github at https://github.com/snap-stanford/uce.

Acknowledgements

We thank Rok Sosič, Kexin Huang, Charlotte Bunne, Hanchen Wang, Michihiro Yasunaga, Michael Moor, Minkai Xu, Mika Jain, George Crowley, Maria Brbić, Jonah Cool, Nicholas Sofroniew, Andrew Tolopko, Ivana Jelic, Ana-Maria Istrate and Pablo Garcia-Nieto for discussions and for providing feedback on our manuscript. We acknowledge support from Robert C. Jones for help with accessing and analyzing the Tabula Sapiens v2 dataset. We acknowledge support from the Chan Zuckerberg Initiative, including help with accessing and processing CxG datasets. We gratefully acknowledge the support of DARPA under Nos. N660011924033 (MCS); NSF under Nos. OAC-1835598 (CINES), CCF-1918940 (Expeditions), Stanford Data Science Initiative, Wu Tsai Neurosciences Institute, Amazon, Genentech, GSK, Hitachi, Juniper Networks, and KDDI. Y. RH. acknowledges funding support form GlaxoSmithKline. L.S. was supported by the American Slovenia Education Foundation (ASEF). Icons created with BioRender.com.

Author information

Y.RS., Y.RH., S.Q. and J.L. conceived the study. Y.RS, Y.RH., S.Q. and J.L. performed research, contributed new analytical tools, designed algorithmic frameworks, analyzed data and wrote the manuscript. Y.RS. and Y.RH. performed experiments and developed the software. A.A. and L.S. contributed to code and performed analyses. T.S. provided annotated data.

References

1. Vaishnav, E. D. *et al.* The evolution, evolvability and engineering of gene regulatory dna. *Nature* **603**, 455–463 (2022).

2. Onuchic, J. N., Luthey-Schulten, Z. & Wolynes, P. G. Theory of protein folding: the energy landscape perspective. *Annual review of physical chemistry* **48**, 545–600 (1997).

3. Moris, N., Pina, C. & Arias, A. M. Transition states and cell fate decisions in epigenetic landscapes. *Nature Reviews Genetics* **17**, 693–703 (2016).

4. Treutlein, B. *et al.* Reconstructing lineage hierarchies of the distal lung epithelium using single-cell rna-seq. *Nature* **509**, 371–375 (2014).

5. Waddington, C. H. *The strategy of the genes* (Routledge, 1957).

6. Schaum, N. *et al.* Single-cell transcriptomics of 20 mouse organs creates a tabula muris: The tabula muris consortium. *Nature* **562**, 367 (2018).

7. Regev, A. *et al.* The human cell atlas. *elife* **6**, e27041 (2017).

8. Rood, J. E., Maartens, A., Hupalowska, A., Teichmann, S. A. & Regev, A. Impact of the human cell atlas on medicine. *Nature medicine* **28**, 2486–2496 (2022).

9. Tanay, A. & Regev, A. Scaling single-cell genomics from phenomenology to mechanism. *Nature* **541**, 331–338 (2017).

10. Plass, M. *et al.* Cell type atlas and lineage tree of a whole complex animal by single-cell transcriptomics. *Science* **360**, eaaq1723 (2018).

11. Consortium*, T. S. *et al.* The tabula sapiens: A multiple-organ, single-cell transcriptomic atlas of humans. *Science* **376**, eabl4896 (2022).

12. Siletti, K. *et al.* Transcriptomic diversity of cell types across the adult human brain. *Science* **382**, eadd7046 (2023).

13. Li, H. *et al.* Fly cell atlas: A single-nucleus transcriptomic atlas of the adult fruit fly. *Science* **375**, eabk2432 (2022).

14. Eraslan, G., Avsec, Ž., Gagneur, J. & Theis, F. J. Deep learning: new computational modelling techniques for genomics. *Nature Reviews Genetics* **20**, 389–403 (2019).

15. Lopez, R., Regier, J., Cole, M. B., Jordan, M. I. & Yosef, N. Deep generative modeling for single-cell transcriptomics. *Nature methods* **15**, 1053–1058 (2018).

16. Luecken, M. D. *et al.* Benchmarking atlas-level data integration in single-cell genomics. *Nature methods* **19**, 41–50 (2022).

17. Argelaguet, R., Cuomo, A. S., Stegle, O. & Marioni, J. C. Computational principles and challenges in single-cell data integration. *Nature biotechnology* **39**, 1202–1215 (2021).

18. Lotfollahi, M. *et al.* Mapping single-cell data to reference atlases by transfer learning. *Nature biotechnology* **40**, 121–130 (2022).

19. Tarashansky, A. J. *et al.* Mapping single-cell atlases throughout metazoa unravels cell type evolution. *Elife* **10**, e66747 (2021).

20. Brown, T. *et al.* Language models are few-shot learners. *Advances in neural information processing systems* **33**, 1877–1901 (2020).

21. OpenAI. Gpt-4 technical report (2023). 2303.08774.

22. Anil, R. *et al.* Palm 2 technical report. *arXiv preprint arXiv:2305.10403* (2023).

23. Touvron, H. *et al.* Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971* (2023).

24. Kirillov, A. *et al.* Segment anything. *arXiv preprint arXiv:2304.02643* (2023).

25. Bommasani, R. *et al.* On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258* (2021).

26. Avsec, Ž. *et al.* Effective gene expression prediction from sequence by integrating long-range interactions. *Nature methods* **18**, 1196–1203 (2021).

27. Rives, A. *et al.* Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences* **118**, e2016239118 (2021).

28. Theodoris, C. V. *et al.* Transfer learning enables predictions in network biology. *Nature* 1–9 (2023).

29. Cui, H. *et al.* scgpt: Towards building a foundation model for single-cell multi-omics using generative ai. *bioRxiv* 2023–04 (2023).

30. Stegle, O., Teichmann, S. A. & Marioni, J. C. Computational and analytical challenges in single-cell transcriptomics. *Nature Reviews Genetics* **16**, 133–145 (2015).

31. Quake, S. R. The cell as a bag of rna. *Trends in Genetics* **37**, 1064–1068 (2021).

32. Vaswani, A. *et al.* Attention is all you need. *Advances in neural information processing systems* **30** (2017).

33. Lin, Z. *et al.* Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123–1130 (2023).

34. Rosen, Y. *et al.* Toward universal cell embeddings: integrating single-cell RNA-seq datasets across species with SATURN. *Nature Methods* **21**, 1492–1500 (2024).

35. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

36. Wang, H., Leskovec, J. & Regev, A. Metric mirages in cell embeddings. *BioRxiv* (2024).

37. Biology, C. S.-C. *et al.* Cz cellxgene discover: A single-cell data platform for scalable exploration, analysis and modeling of aggregated data. *bioRxiv* 2023–10 (2023).

38. Gordon, S., Plüddemann, A. & Martinez Estrada, F. Macrophage heterogeneity in tissues: phenotypic diversity and functions. *Immunological reviews* **262**, 36–55 (2014).

39. Conde, C. D. *et al.* Cross-tissue immune cell analysis reveals tissue-specific features in humans. *Science* **376**, eabl5197 (2022).

40. Speranza, E. *et al.* Single-cell rna sequencing reveals sars-cov-2 infection dynamics in lungs of african green monkeys. *Science translational medicine* **13**, eabe8146 (2021).

41. Hilton, H. G. *et al.* Single-cell transcriptomics of the naked mole-rat reveals unexpected features of mammalian immunity. *PLoS Biology* **17**, e3000528 (2019).

42. Yamagata, M., Yan, W. & Sanes, J. R. A cell atlas of the chick retina based on single-cell transcriptomics. *Elife* **10**, e63907 (2021).

43. Mantri, M. *et al.* Spatiotemporal single-cell rna sequencing of developing chicken hearts identifies interplay between cellular differentiation and morphogenesis. *Nature communications* **12**, 1771 (2021).

44. Orozco, L. D. *et al.* Integration of eQTL and a single-cell atlas in the human eye identifies causal genes for age-related macular degeneration. *Cell reports* **30**, 1246–1259.e6 (2020).

45. Li, H. *et al.* Fly cell atlas: A single-nucleus transcriptomic atlas of the adult fruit fly. *Science* **375**, eabk2432 (2022).

46. Bard, J., Rhee, S. Y. & Ashburner, M. An ontology for cell types. *Genome biology* **6**, 1–5 (2005).

47. Kragesteen, B. K. *et al.* The transcriptional and regulatory identity of erythropoietin producing cells. *Nature medicine* 1–10 (2023).

48. Haine, L. *et al.* Cytoprotective effects of erythropoietin: What about the lung? *Biomedicine & Pharmacotherapy* **139**, 111547 (2021).

49. Adams, T. S. *et al.* Single-cell rna-seq reveals ectopic and aberrant lung-resident cell populations in idiopathic pulmonary fibrosis. *Science advances* **6**, eaba1983 (2020).

50. Tassiopoulos, S. *et al.* Erythropoietic response to hypoxaemia in diffuse idiopathic pulmonary fibrosis, as opposed to chronic obstructive pulmonary disease. *Respiratory Medicine* **95**, 471–475 (2001).

51. Abdel-Aziz, C., Okaily, N. & Kasem, A. Erythropoietin: role in idiopathic pulmonary fibrosis revisited. *The Egyptian Journal of Chest Diseases and Tuberculosis* **69**, 716 (2020).

52. Tsantes, A. E. *et al.* Red cell macrocytosis in hypoxemic patients with chronic obstructive pulmonary disease. *Respiratory medicine* **98**, 1117–1123 (2004).

53. Safran, M. *et al.* The GeneCards suite. In Abugessaisa, I. & Kasukawa, T. (eds.) *Practical guide to life science databases*, 27–56 (Springer Singapore, Singapore, 2021).

54. Stelzer, G. *et al.* The genecards suite: from gene data mining to disease genome sequence analyses. *Current Protocols in Bioinformatics* **54**, 1.30.1–1.30.33 (2016).

55. Piran, Z., Cohen, N., Hoshen, Y. & Nitzan, M. Disentanglement of single-cell data with biolord. *Nature Biotechnology* (2024).

56. Piran, Z. & Nitzan, M. SiFT: uncovering hidden biological processes by probabilistic filtering of single-cell data. *Nature Communications* **15**, 760 (2024).

57. Lopez, R. *et al.* Learning causal representations of single cells via sparse mechanism shift modeling. In van der Schaar, M., Zhang, C. & Janzing, D. (eds.) *Proceedings of the Second Conference on Causal Learning and Reasoning*, vol. 213 of *Proceedings of Machine Learning Research*, 662–691 (PMLR, 2023).

58. Kunes, R. Z., Walle, T., Land, M., Nawy, T. & Pe'er, D. Supervised discovery of interpretable gene programs from single-cell data. *Nature Biotechnology* **42**, 1084–1095 (2024).

59. Amaral, P. *et al.* The status of the human gene catalogue. *Nature* **622**, 41–47 (2023).

60. Brenner, S. Nature's gift to science (nobel lecture). *Chembiochem* **4**, 683–687 (2003).

61. Gu, A., Goel, K. & Ré, C. Efficiently modeling long sequences with structured state spaces. *arXiv* (2021).

62. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biology* **19**, 15 (2018).

63. Shen, H. *et al.* Generative pretraining from large-scale transcriptomes for single-cell deciphering. *iScience* **26**, 106536 (2023).