Exploration and generation of cell transcriptomes over deep evolutionary time

Ying Xu[1], Catherine Gatt[1], Esra Kaymak[2], Kensei Kikuchi[2], Thomas Bourguignon[2], and Fabio Zanini[1,3,4]

[1] School of Clinical Medicine, UNSW Sydney, 2052, NSW, Australia
[2] Okinawa Institute of Science & Technology Graduate University, 1919-1 Tancha, Onna-son, 904-0495 Okinawa, Japan
[3] UNSW Cellular Genomics Futures Institute, 2052, NSW, Australia
[4] Evolution & Ecology Research Centre, 2052, NSW, Australia

Abstract

Whole-organism cell atlases have painted the cellular landscapes of individual species; however, comparing cells across the tree of life remains challenging. Most cross-species analyses are restricted to orthologous genes, while the complexity of atlas data constitutes an access barrier for many researchers. We developed a computational strategy to accelerate the exploration of cellular identities at scale. We integrated 30 atlases detailing the expression of 861,013 genes by 2,645,508 animal and plant cells and trained a universal model of cellular transcriptomes to track cell type diversification over evolutionary times. Transfer learning achieved cell type annotation of a *de novo* atlas of the insect *Cryptocercus punctulatus* within minutes. We devised a generative artificial intelligence approach to construct virtual cell atlases from genome sequences and used it to synthesise an atlas of the Tasmanian tiger, extinct since 1936. We then reduced the footprint of extant atlases by 100 times while retaining crucial information and developed interfaces to answer dozens of query types within seconds, boosting atlas exploration by ~50,000 times.

Main

Life on Earth revolves around two central elements: the genome, which stores heritable information, and the cell, which organises that information by selective expression of genes and proteins. The study of genome evolution, aided by innovative technologies such as ancient DNA sequencing [1] and protein language models [2], continues to reveal new connections between divergent species. Cellular evolution, i.e. how the orchestration of gene and protein expression changes over evolutionary time, remains comparatively understudied. Cell identities are clearly not private to individual species: sister cell types, with identical name and similar function across distinct organisms, are common (e.g. muscle cells, immune cells, neurons). These observations notwithstanding, a unified model connecting cellular identities across the tree of life is still missing.

Single-cell transcriptomic technologies, which profile the gene expression of thousands of individual cells in a single experiment, have revolutionised our understanding of cell types, states, and identities across organs and even entire organisms [3–10]. Though designed to serve as public repositories for the entire biological community, these “cell atlases” remain difficult to explore and link into an integrated, universal model. First, cell atlas data are large, complex, and scattered across the web in idiosyncratic formats.

Second, each species harbours a different genome sequence, which complicates cross-species comparisons. To overcome these challenges, new strategies are needed.

In physics and complex systems research, approximations are often used to identify accurate solutions to otherwise intractable problems. Here, we present atlas approximations, a computational strategy to enable scalable exploration of cellular identities across cell types, organs, and organisms. We constructed approximations on 30 whole-organism cell atlases from both plants and animals. We trained a universal deep learning model to track cell types across 25 of these species with well annotated genomes, integrating more than one billion years of cellular evolution. We constructed and experimentally validated a transfer learning model to annotate most cell types in a new species within minutes. Finally, we developed a widely applicable user experience to explore cell atlases within seconds, with or without programming skills.

Annotation approximations establishes trustworthy cellular identities across species

We located (online or through email requests), downloaded, and standardised whole-organism cell atlases comprising both animals and plants (**Supplementary Table 1**) [4–10], resulting in a streamlined data set for multiscale atlas exploration, from individual genes and cells to entire superkingdoms (**Figure 1A**). Overall, the entirety of our data set included more than 2.5 million cells and more than 800,000 genes across 30 species (**Figure 1B**). Through detailed manual evaluation of these data, we identified three key approximations with the potential to unlock broad comparative exploration of cellular identities: (i) annotation approximations, (ii) gene approximations, and (iii) cell approximations.

In raw atlas data, cell type annotations were very heterogeneous across organs and organisms. Deeper or more specific annotations were generally less credible. For instance, a T cell might be annotated as immune, T cell, CD4+ T cell, or naive CD4+ T cell, often despite a lack of expression of the CD4 gene itself. Seminumerical annotations such as “fibroblast 1”, difficult to interpret biologically, were not uncommon. To mitigate these issues, we manually coarse grained cell type annotations across all cells, organs, and organisms, aiming for limited depth but high interpretability and trustworthiness (**Supplementary Figure 1**). This annotation approximation workflow, implemented using declarative configuration files and iteratively refined upon community feedback, yielded a total of 386 broad, distinct cell types across all species.

A universal model of cellular identities via gene approximations

Equipped with clear cell annotations, we set out to address the issue that each species harbours a unique genome. We adapted the deep autoencoder model SATURN [11], which approximates genes from individual species into cross-species “macrogenes”, to work on dozens of species at once. Combining a cell-type rebalanced subsample with ESM protein embeddings [2] of 616,142 unique genes (**Figure 1A**), we trained an autoencoder with 1,000 macrogenes for 12 hours on a datacenter GPU and obtained a universal embedding of 212,189 cells across 25 organisms, 44 organs, and 351 approximate cell types, encompassing more than one billion years of cellular evolution (**Supplementary Table 2**). This enabled unified data visualisation across the tree of eukaryotic life (**Supplementary Figures 2-4**) and exploration of deeply conserved identities such as muscle, neurons, and immune cells (**Figure 1C**, **Supplementary Figures 5-6**). To illustrate quantitative applications of the model, we identified macrogenes heavily affected by markers of human muscle and traced which organisms’ muscle-like cells expressed each macrogene against their time since common ancestor (CA) with humans (**Figure 1D**, **Supplementary Figure 7**). We

performed nonlinear fits of logistic functions (dashed lines) and estimated that expression of macrogenes 52 and 203 emerged ~450 (Ordovician) and ~250 million (Triassic) years ago, while macrogene 673 has only begun to emerge and is projected to spread across all muscle cells ~140 million years in the future (**Figure 1E**).

Organismal evolution and cell differentiation both generate cellular heterogeneity. To disambiguate between these factors, we defined phylogeny-adjusted pseudotime, which computes a continuous metric renormalised between 0 and 1 within each species starting from a single root cell one from a basal organism. To demonstrate the concept, we constructed phylogeny-adjusted pseudotime in immune cells (**Figure 1F**) and discovered that in three out of three mammals, macrophages were predicted as the most evolutionary basal immune cells, with a gradient from innate to adaptive immunity, correctly recapitulating the evolutionary history (**Figure 1G**). A similar analysis identified smooth muscle as more primitive than cardiomyocytes (**Supplementary Figures 8-9**).

Annotating the nonmodel cockroach *Cryptocercus punctulatus* within minutes

One of the applications of a universal evolutionary model is to accelerate cell type annotation of cell atlases in nonmodel organisms. Insects, in particular, make up about half of described extant species [12], yet only one whole-organism atlas is available (*Drosophila melanogaster*) [13]. Using the frozen universal encoder learned by SATURN, we developed a transfer learning model to learn the gene to macrogene mappings of a new species without retraining the full model (**Figure 2A**). The transfer model could embed cell types with roundtrip mutual nearest neighbour accuracy indistinguishable from a full SATURN model during leave-one-organism-out tests starting from human (**Figure 2B, Supplementary Figure 10**), while only requiring minutes of fine-tuning on a desktop GPU (**Figure 2A**). To further validate SATURN-xfer on nonmodel organisms, we dissociated a single individual of the cockroach *Cryptocercus punctulatus* into individual nuclei, filtered them using cell strainer and a cell sorter with DAPI staining to distinguish them from debris, and constructed a single nuclei transcriptomic library (**Figure 2C**). After aligning the reads against the genome of the sister species *Cryptocercus meridianus*, we obtained 1,751 high-quality transcriptomes across 20 unsupervised clusters (**Figure 2D**). After embedding the *Cryptocercus* peptide sequences with ESM [2], we launched SATURN-xfer using *Drosophila melanogaster* as a guide species and obtained a joint embedding within 82 seconds on a desktop GPU (**Figure 2E**). Despite diverging ~400 million years ago [14], the two species possessed cellular transcriptomes similar enough for straightforward annotation of 70% of clusters (14/20) and 74% of cells, including cluster 9 as muscle (100% consistency) (**Figure 2F**) and cluster 17 as neurons (86% consistency) (**Figure 2G**). In the universal model, macrogenes upregulated in clusters 9 and 17 had high weights for human genes highly expressed in striated muscle (e.g. TTN, MYL1) and neurons (e.g. SYN3, TTL [15]), respectively. Some *Cryptocercus* cells embedded separately from all *Drosophila* cells, suggesting evolutionary specialisations (**Figure 2E**). Overall, these results suggest that the deep encoder from macrogene to latent space has learned transcriptomic properties that are generalisable across large evolutionary gaps.

A synthetic cell atlas of *Thylacinus cynocephalus*, the extinct Tasmanian tiger

As part of pretraining the universal model [11], both the encoder and its twin decoder are trained together. While the universal encoder can be used to annotate new experimental atlas data, the decoder can be leveraged for generative inference of an entire cell atlas in species that lack one (**Figure 3A**). To

demonstrate this concept, we constructed a synthetic human cell atlas from Tabula Sapiens [4] from the universal model in less than one minute (**Figure 3B**). The UMAP representation appeared less structured than the original publication [4], which is expected given that the back weights from macrogene to human gene space are not trained (zeroshot inference). To evaluate the faithfulness of the cellular heterogeneity in synthetic atlases, we computed how often nearest neighbors in the synthetic atlas belong to the same cell type as determined in the original publication (**Figure 3C**). We observed that for most cell types the synthetic atlas performs no worse than the original data, with significant degradation limited to specific cell types (e.g. pancreatic beta cells). Encouraged by these validations, we downloaded the recently annotated genome of *Thylacinus cynocephalus* or Tasmanian tiger, a carnivorous marsupial that went extinct about a century ago [16], and ran SATURN-gen zeroshot guided by Tabula muris (mouse) data to create a synthetic cell atlas of the thylacine (**Figure 3D**). Synthetic thylacine transcriptomes were meaningfully structured with regards to cell supertypes, with the vast majority of nearest neighbors belonging to the same cell supertype (**Figure 3E**, dots vs violin for random controls). Cell type-specific gene expression was also at least partially preserved: synthetic thylacine striated muscle cells specifically expressed homologs of human MYL1, MYL4, markers of striated muscle in living mammals (**Figure 3F**), while synthetic thylacine B cells expressed specifically an ortholog of the human B cell marker CD19 (**Figure 3G**). Taken together, these results indicate that generative artificial intelligence can provide meaningful cell biology insights about rare and extinct species.

Accelerating atlas exploration by orders of magnitude with cell approximations

We then developed cell approximations to address the central paradox of cell atlases: massive data is what makes an atlas informative but also reduces its usability by human researchers, due to higher skill, time, software, and hardware requirements (**Figure 4A**). We used lossless and lossy data compression techniques to reduce the size of all cell atlas by up to 99% (**Figure 4B**, *Homo sapiens* includes transcriptomic chromatin accessibility data, and **Supplementary Figure 11**), summarising single-cell data based on either biological annotation (cell type approximation) or unlabelled neighborhoods of a UMAP embedding (cell state approximation) (**Figure 4C**). For each bin and each gene, we recorded its average expression and the fraction of cells with nonzero expression, the two most commonly used metrics in summary visualisations (e.g. dot plots, heatmaps). For cell state compressions, we also computed and stored approximate UMAP embeddings via centroids and convex hulls.

We reasoned that a drastic reduction in data size could significantly accelerate user queries. To demonstrate this concept, we constructed a language-agnostic application programming interface (API) to answer a wide array of user queries from our 30 atlases (**Supplementary Table 3**). This includes identification of cell type-specific marker genes, of the highest expressor of a certain gene across organs, and of cross-species gene and cell type homologs; dot plots for a list of genes in one organ or across organs; gene-gene and cell type-cell type correlations; visually guided explorations; approximate UMAP embeddings; and so on. We then deployed the API onto a cloud server and demonstrated roundtrip query responses of 1-2 seconds from virtually anywhere in the world, representing a speedup of up to ~50,000 times compared to standard cell atlas analysis (**Figure 4D**, **Supplementary Table 4**). We also released language-specific packages in Python, R, JavaScript, and Bash to help query automation (**Supplementary Figure 12**), and developed a dedicated package, *scquill*, to perform cell approximations on arbitrary data sets, including additional metadata such as developmental stage, sex, or disease condition.

Democratising cell identity exploration via text interpretation on the web

Finally, we aimed to unlock atlas exploration for researchers with no programming skills. For this purpose, we trained a small language model to interpret English user text, ask follow-up questions, and provide textual and visual answers within a newly developed web application at https://atlasapprox.org (**Figure 4E, Supplementary Figures 11, 13-21**). This natural language model is open source, lightweight (~1 MB), and runs in the user's browser, requiring neither server maintenance nor subscription to cloud platforms; it is ideally suited not only to researchers with abundant or limited resources alike, but also for downstream adaptations to other research contexts. We supported this beginner-oriented workflow with text, code, and video tutorials. The unique combination of responsiveness and text interpretation creates a user experience that flows naturally from one request to the next along biological investigative lines, unlike anything we have seen elsewhere. For example, we could confirm within 15 seconds that tunicates (*Ciona intestinalis*) and humans, despite diverging half a billion years ago, share homologous muscle cell markers (**Figure 4F**).

Discussion

Cellular evolution has fascinated researchers for decades, from neuronal origins [9] to immune functions [17]. This study leverages deep learning and highly curated training data to integrate cellular transcriptomes across tens of organs, hundreds of cell types, hundreds of thousands of cells, and half a million genes into one coherent model encompassing one billion years of evolution. The scalability of this approach is partially credited to the base SATURN architecture [11], as well as to whole-organism cell atlases in plants [18–23] and animals [4–7,9,24,25], supported by consortia including the Human Cell Atlas [26], the Plant Cell Atlas [27], and the Tabulae [4,5,24]. Transcriptomic data for fungi has been collected at the bulk level across species [28] and single cell transcriptomics from marine plankton has been demonstrated recently [29], raising hopes for less travelled paths within the tree of life.

The universal model presented here builds upon recent efforts on cross-species data harmonisation [11,30]. The recently proposed Universal Cell Embeddings (UCEs) increased model size and complexity and focused on human and mouse while including data from individual tissues of eight vertebrates [31]. Our approach is related to but distinct from UCEs. First, we ingest data almost exclusively from whole-organism cell atlases, an approach pioneered by Chan Zuckerberg Biohub aiming for optimal data quality through coordinated collection of dozens of organs by the same consortium, from the same specimens, and at the same time [3]. Second, we manually curated and approximated cell annotations to maximise their biological trustworthiness. Third, we eschewed massive data bias towards mammals, choosing instead to include organisms deeply divergent from humans, such as corals, sponges, comb jellies, and plants. This strategy hinges on the insight that cell diversity, rather than raw cell numbers, might be the most important asset to train a universal model of cell identities. This hypothesis is confirmed by our ability to identify sister cells to human muscle, neurons, and macrophages in many animals phylogenetically distant from us.

The identification of sister cell types is key to annotate cells in nonmodel organisms [8,32]. As demonstrated by our mammal leave-one-out tests and *Cryptocercus* experiments, transfer learning can accelerate this process at no detectable accuracy cost. This supports the hypothesis that once the autoencoder is trained sufficiently along the tree of life topology, phylogenetic gaps as large as hundreds of millions of years can be partially bridged over at inference time. This conclusion is further buttressed by the demonstration of our generative model, which could synthesise cellular transcriptomes for the extinct thylacine starting only from

its annotated genome [16] and our universal decoder. The main limitation of this zero-shot generative model is distinguishing between closely related paralogs, however more accurate protein embeddings such as ESM-Cambrian and recent transformer-based architectures might help [31].

Cell atlases are primarily disseminated as monolithic binary files that require advanced bioinformatic skills and powerful hardware to access. Web platforms such as the Chan Zuckerberg Initiative CellXGene [33] and SPEED [34] mitigate this issue at the level of individual datasets but fall short of providing a unified interface across organs and organisms. Cell approximations provide a streamlined interface to cell atlases across 30 species for both programmers and noncoders alike, computing answers within two seconds [35]. The opportunity to recursively formulate queries about marker genes, highest expressors, cross-species homology, peptide sequences, cell embeddings, cell type abundances and so on encourages an inquisitive mindset to cell atlas exploration that is quite unlike any previous workflow we have experienced.

Overall, this study demonstrated that comparative analysis of cellular transcriptomes at a planetary scale can accelerate biological discovery on both extant and extinct species while also democratising data access, opening the doors to a deeper understanding of cellular identities and their evolutionary journey.

Acknowledgements

We would like to express our heartfelt thanks to Robert C Jones, Carsten Knutsen, Cristina Alvira, David Cornfield, Emily Wong, Junyue Cao, Bo Wang, Dania Nanes Sarfati, Simon Peters, Jules Duruz, and all authors of the original atlas publications for data access and scientific discussions, Dino P McMahon and Mark C Harrison for sharing the *Cryptocercus* genome, Simon Hellemans for help with the cockroach dissection, Danny Dien for early prototypes of the chatbot, Joanna Ahn for literature search, and Haolan Zhou for organisms' drawings. This work is supported by a Chan Zuckerberg Initiative Single Cell Data Insight Grant #DI2-0000000091.

References

1. Reich, D. *et al.* Genetic history of an archaic hominin group from Denisova Cave in Siberia. *Nature* **468**, 1053–1060 (2010).
2. Lin, Z. *et al.* Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123–1130 (2023).
3. Quake, S. R. A decade of molecular cell atlases. *Trends Genet.* **38**, 805–810 (2022).
4. Tabula Sapiens Consortium* *et al.* The Tabula Sapiens: A multiple-organ, single-cell transcriptomic atlas of humans. *Science* **376**, eabl4896 (2022).
5. Tabula Muris Consortium. Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. *Nature* (2018) doi:10.1038/s41586-018-0590-4.
6. Wagner, D. E. *et al.* Single-cell mapping of gene expression landscapes and lineage in the zebrafish embryo. *Science* **360**, 981–987 (2018).
7. Li, P. *et al.* Single-cell analysis of Schistosoma mansoni identifies a conserved genetic program controlling germline stem cell fate. *Nat. Commun.* **12**, 485 (2021).
8. Sebé-Pedrós, A. *et al.* Early metazoan cell type diversity and the evolution of multicellular gene regulation. *Nat Ecol Evol* **2**, 1176–1188 (2018).
9. Musser, J. M. *et al.* Profiling cellular diversity in sponges informs animal cell type and nervous system evolution. *Science* **374**, 717–723 (2021).
10. Plass, M. *et al.* Cell type atlas and lineage tree of a whole complex animal by single-cell transcriptomics. *Science* **360**, (2018).

11. Rosen, Y. *et al.* Toward universal cell embeddings: integrating single-cell RNA-seq datasets across species with SATURN. *Nat. Methods* **21**, 1492–1500 (2024).
12. Li, X. & Wiens, J. J. Estimating global biodiversity: The role of cryptic insect species. *Syst. Biol.* **72**, 391–403 (2023).
13. Li, H. *et al.* Fly Cell Atlas: A single-nucleus transcriptomic atlas of the adult fruit fly. *Science* **375**, eabk2432 (2022).
14. Misof, B. *et al.* Phylogenomics resolves the timing and pattern of insect evolution. *Science* **346**, 763–767 (2014).
15. Erck, C. *et al.* A vital role of tubulin-tyrosine-ligase for neuronal organization. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 7853–7858 (2005).
16. Feigin, C., Frankenberg, S. & Pask, A. A chromosome-scale hybrid genome assembly of the extinct Tasmanian tiger (Thylacinus cynocephalus). *Genome Biol. Evol.* **14**, evac048 (2022).
17. Yutin, N., Wolf, M. Y., Wolf, Y. I. & Koonin, E. V. The origins of phagocytosis and eukaryogenesis. *Biol. Direct* **4**, 9 (2009).
18. Shahan, R. *et al.* A single-cell Arabidopsis root atlas reveals developmental trajectories in wild-type and cell identity mutants. *Dev. Cell* **57**, 543–560.e9 (2022).
19. Xu, X. *et al.* Large-scale single-cell profiling of stem cells uncovers redundant regulators of shoot development and yield trait variation. *bioRxiv* 2024.03.04.583414 (2024) doi:10.1101/2024.03.04.583414.
20. Abramson, B. W. *et al.* The genome and preliminary single-nuclei transcriptome of Lemna minuta reveals mechanisms of invasiveness. *Plant Physiol.* **188**, 879–897 (2022).
21. Zhang, T.-Q., Chen, Y., Liu, Y., Lin, W.-H. & Wang, J.-W. Single-cell transcriptome atlas and chromatin accessibility landscape reveal differentiation trajectories in the rice root. *Nat. Commun.* **12**, 2053 (2021).
22. Zhang, L. *et al.* Asymmetric gene expression and cell-type-specific regulatory networks in the root of bread wheat revealed by single-cell multiomics analysis. *Genome Biol.* **24**, 65 (2023).

23. Marand, A. P., Chen, Z., Gallavotti, A. & Schmitz, R. J. A cis-regulatory atlas in maize at single-cell resolution. *Cell* **184**, 3041–3055.e21 (2021).
24. The Tabula Microcebus Consortium *et al.* Tabula Microcebus: A transcriptomic cell atlas of mouse lemur, an emerging primate model organism. *bioRxiv* 2021.12.12.469460 (2022) doi:10.1101/2021.12.12.469460.
25. Liao, Y. *et al.* Cell landscape of larval and adult Xenopus laevis at single-cell resolution. *Nat. Commun.* **13**, 4306 (2022).
26. Rozenblatt-Rosen, O., Stubbington, M. J. T., Regev, A. & Teichmann, S. A. The Human Cell Atlas: from vision to reality. *Nature* **550**, 451–453 (2017).
27. Plant Cell Atlas Consortium *et al.* Vision, challenges and opportunities for a Plant Cell Atlas. *Elife* **10**, (2021).
28. Krizsán, K. *et al.* Transcriptomic atlas of mushroom development reveals conserved genes behind complex multicellularity in fungi. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 7409–7418 (2019).
29. Gatt, C., Xie, Y., Wahi, K., Johansson, E. M. V. & Zanini, F. Integrating microscopy and transcriptomics from individual uncultured eukaryotic plankton. *eLife* (2024) doi:10.7554/elife.102991.1.
30. Tarashansky, A. J. *et al.* Mapping single-cell atlases throughout Metazoa unravels cell type evolution. *Elife* **10**, (2021).
31. Rosen, Y. *et al.* Universal cell embeddings: A foundation model for cell biology. *bioRxiv* 2023.11.28.568918 (2023) doi:10.1101/2023.11.28.568918.
32. Piovani, L. *et al.* Single-cell atlases of two lophotrochozoan larvae highlight their complex evolutionary histories. *Sci. Adv.* **9**, eadg6034 (2023).
33. Chan Zuckerberg Initiative. CZ CELLxGENE Discover. *CZ CELLxGENE Discover* https://cellxgene.cziscience.com/.
34. Chen, Y. *et al.* SPEED: Single-cell Pan-species atlas in the light of Ecology and Evolution for Development and Diseases. *Nucleic Acids Res.* **51**, D1150–D1159 (2023).
35. Zanini, F. *et al.* Northstar enables automatic classification of known and novel cell types from tumor

samples. *Sci. Rep.* **10**, 15251 (2020).

36. McInnes, L., Healy, J. & Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv [stat.ML]* (2018).
37. Redelings, B. D. & Holder, M. T. A supertree pipeline for summarizing phylogenetic and taxonomic information for millions of species. *PeerJ* **5**, e3058 (2017).
38. Liu, C. *et al.* Unravelling termite evolution with 47 high-resolution genome assemblies. *bioRxiv* 2025.01.20.633303 (2025) doi:10.1101/2025.01.20.633303.
39. Zstandard. http://facebook.github.io/zstd/.
40. The HDF Group. Hierarchical Data Format, version 5. *Hierarchical Data Format, version 5* https://www.hdfgroup.org/HDF5 (1997).
41. Grinberg, M. *Flask Web Development: Developing Web Applications with Python*. (O'Reilly Media, 2018).
42. Merkel, D. Docker: lightweight Linux containers for consistent development and deployment. *Linux J.* **2014**, 2 (2014).
43. *Nlp.js: An NLP Library for Building Bots, with Entity Extraction, Sentiment Analysis, Automatic Language Identify, and so More*. (Github).
44. React. https://react.dev/.
45. Plotly Technologies Inc. Collaborative data science. *Collaborative data science* https://plot.ly (2015).
46. Türei, D., Korcsmáros, T. & Saez-Rodriguez, J. OmniPath: guidelines and gateway for literature-curated signaling pathway resources. *Nat. Methods* **13**, 966–967 (2016).
47. Kilinc, M., Jia, K. & Jernigan, R. L. Improved global protein homolog detection with major gains in function identification. *Proc. Natl. Acad. Sci. U. S. A.* **120**, e2211823120 (2023).

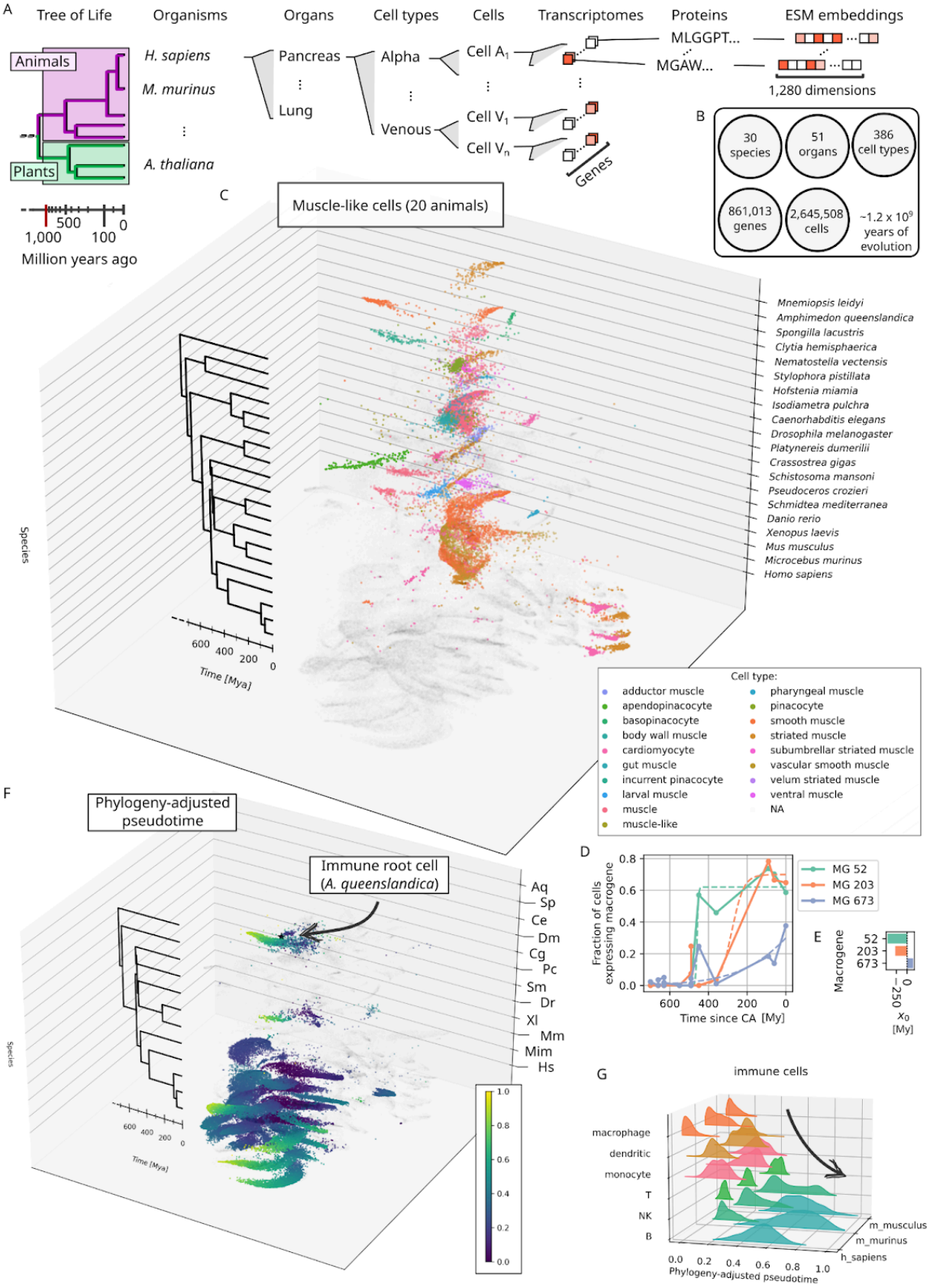
A
Tree of Life
Organisms
Organs
Cell types
Cells
Transcriptomes
Proteins
ESM embeddings
Animals
Plants
H. sapiens
M. murinus
A. thaliana
Pancreas
Lung
Alpha
Venous
Cell A1
Cell V1
Cell Vn
Genes
MLGGPT...
MGAW...
1,280 dimensions
500
1,000
100
0
Million years ago
B
30 species
51 organs
386 cell types
861,013 genes
2,645,508 cells
~1.2 x 10^9 years of evolution
C
Muscle-like cells (20 animals)
Mnemiopsis leidyi
Amphimedon queenslandica
Spongilla lacustris
Clytia hemisphaerica
Nematostella vectensis
Stylophora pistillata
Hofstenia miamia
Isodiametra pulchra
Caenorhabditis elegans
Drosophila melanogaster
Platynereis dumerilii
Crassostrea gigas
Schistosoma mansoni
Pseudoceros crozieri
Schmidtea mediterranea
Danio rerio
Xenopus laevis
Mus musculus
Microcebus murinus
Homo sapiens
Species
600
400
200
0
Time [Mya]
Cell type:
adductor muscle
apendopinacocyte
basopinacocyte
body wall muscle
cardiomyocyte
gut muscle
incurrent pinacocyte
larval muscle
muscle
muscle-like
pharyngeal muscle
pinacocyte
smooth muscle
striated muscle
subumbrellar striated muscle
vascular smooth muscle
velum striated muscle
ventral muscle
NA
D
Fraction of cells expressing macrogene
0.8
0.6
0.4
0.2
0.0
600
400
200
0
Time since CA [My]
MG 52
MG 203
MG 673
E
Macrogene
52
203
673
−250
0
x0 [My]
F
Phylogeny-adjusted pseudotime
Immune root cell (A. queenslandica)
Aq
Sp
Ce
Dm
Cg
Pc
Sm
Dr
Xl
Mm
Mim
Hs
Species
600
400
200
0
Time [Mya]
1.0
0.8
0.6
0.4
0.2
0.0
G
immune cells
macrophage
dendritic
monocyte
T
NK
B
m_musculus
m_murinus
h_sapiens
0.0
0.2
0.4
0.6
0.8
1.0
Phylogeny-adjusted pseudotime

Figure 1: Annotation and gene approximations enable deep evolutionary comparisons of cellular identities. A: Schematic of atlas approximations data, zooming in from the tree of life to each individual cell and protein. ESM: Evolutionary scale model [2]. **B:** Number of organisms, organs, approximate cell types, cells, and genes included in atlas approximations, and evolutionary time spanned by the species in the dataset. **C:** Unified Uniform Manifold Approximation and Projection (UMAP) of muscle-like cells across twenty animals [36], computed using annotation and gene approximations [11] and stratified in 3D sheets according to each organism's leaf in a synthetic tree of life [37]. NA: Not a muscle-like cell. **D:** Fraction of muscle-like cells expressing three macrogenes with high weights for muscle markers along evolutionary time. Each species is assigned an x-coordinate based on the time since common ancestor (CA) with humans. Logistic function fits are shown as dashed lines. **E:** Mid-height coordinate x_0 from the previous fits. **F:** Phylogeny-adjusted pseudotime for the immune cells. Black star: root cell. **G:** Distributions of phylogeny-adjusted pseudotime in immune cells across three mammals and six cell types. The arrow is a guide to the eye.

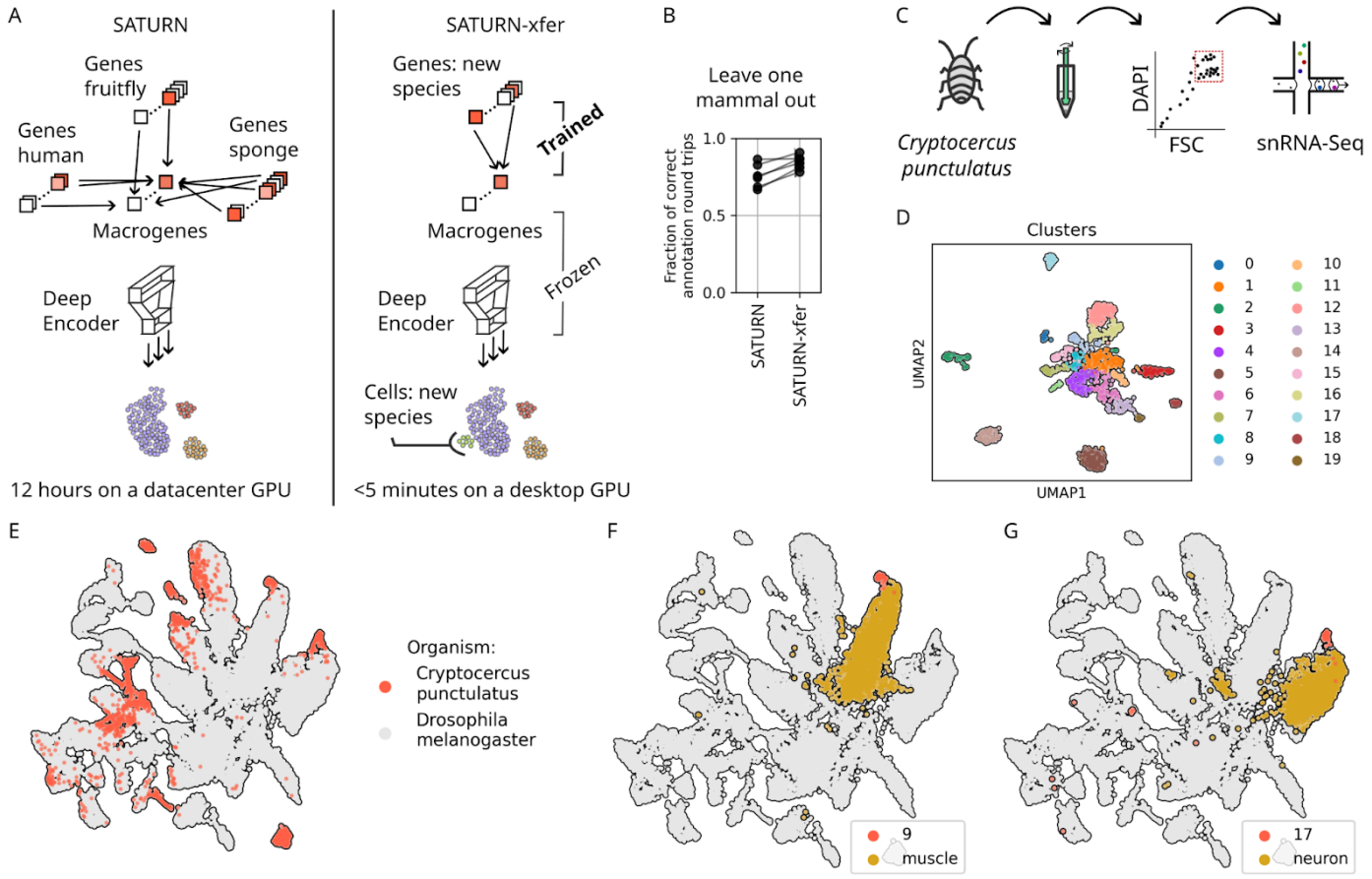

Figure 2: Transfer learning enables rapid annotation of nonmodel species. A: Comparison of the architecture of SATURN (left) and SATURN-xfer (right). **B:** Accuracy of SATURN-xfer across leave-one-out tests in mammals, demonstrating performance on par with a full model. **C:** Schematic of the experimental protocol for single-nuclei transcriptomics of the cockroach *Cryptocercus punctulatus*. **D:** UMAP embedding of *Cryptocercus* cells, coloured by Leiden clustering. **E:** UMAP after SATURN-xfer with 30 epochs of fine tuning. Only *Cryptocercus* and *Drosophila melanogaster* cells are visualised for simplicity. **F:** Same cells as

in E, this time colouring *Cryptocercus* cluster 9 (red) and *Drosophila* muscle cells (yellow) (left), and *Cryptocercus* cluster 17 (red) and *Drosophila* neurons (yellow) (right).

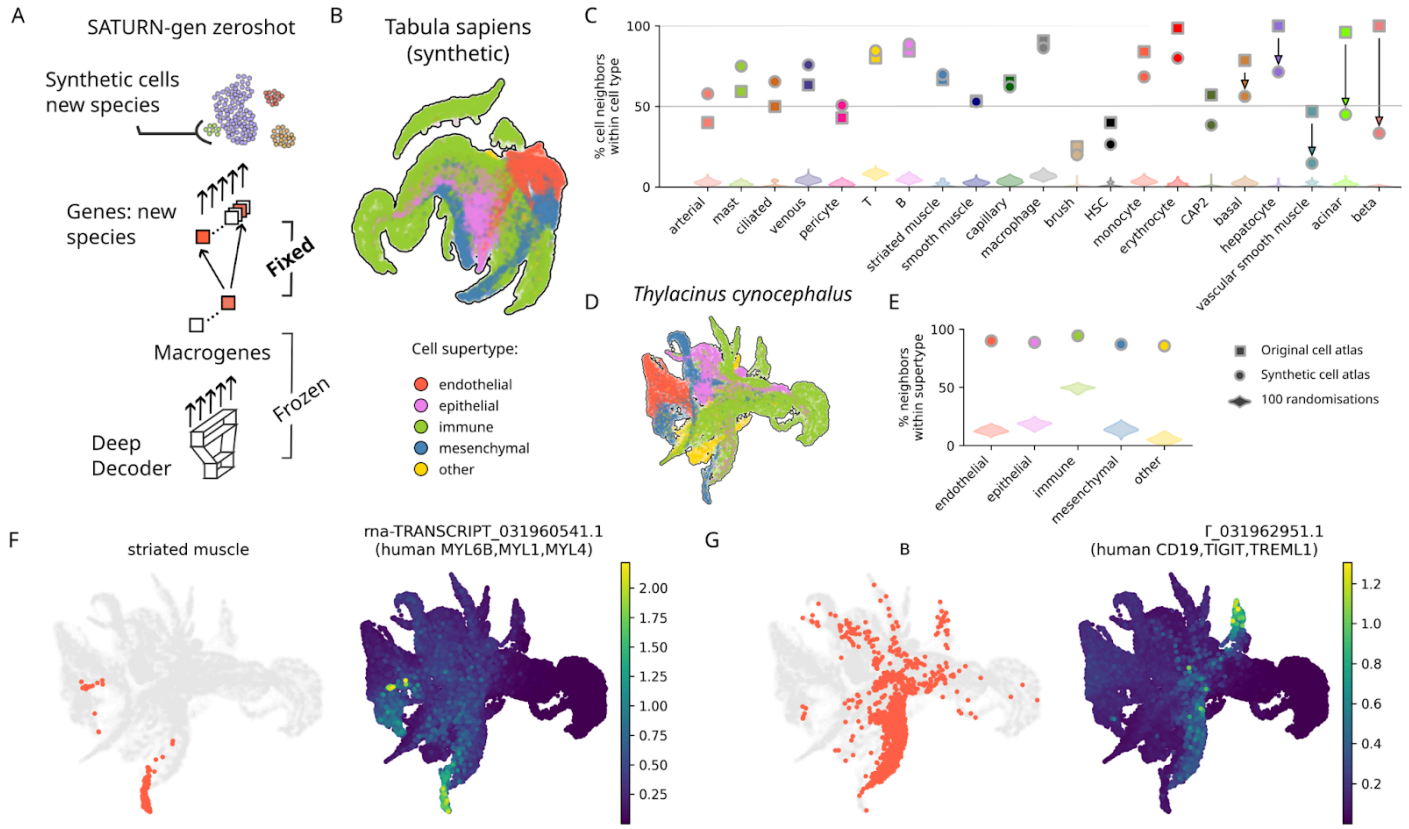

Figure 3: Generative learning creates a synthetic cell atlas of the extinct thylacine. A: Schematic of SATURN-gen zeroshot. **B:** UMAP representation of Tabula sapiens (human) synthetic cell transcriptomes by SATURN-gen zeroshot, coloured by cell supertype. **C:** Percentage of cell nearest neighbors belonging to the same cell type in the original cell atlas (squares), the synthetic atlas (circles), and 100 randomisations of the cell type labels (violins). **D:** UMAP representation of synthetic *Thylacinus cynocephalus* cell transcriptomes by SATURN-gen zeroshot guided by Tabula muris. Colours as in **B**. **E:** Percentage of nearest neighbor cells within each cell supertype in the synthetic atlas (circles) and 100 cell randomisations of the cell type labels (violins). **F-G:** UMAP representations as in **D**, colored to highlight specific cell types (top) and based on synthetic gene expression for cell type markers with human homologs (bottom) for striated muscle cells (**F**) and B cells (**G**).

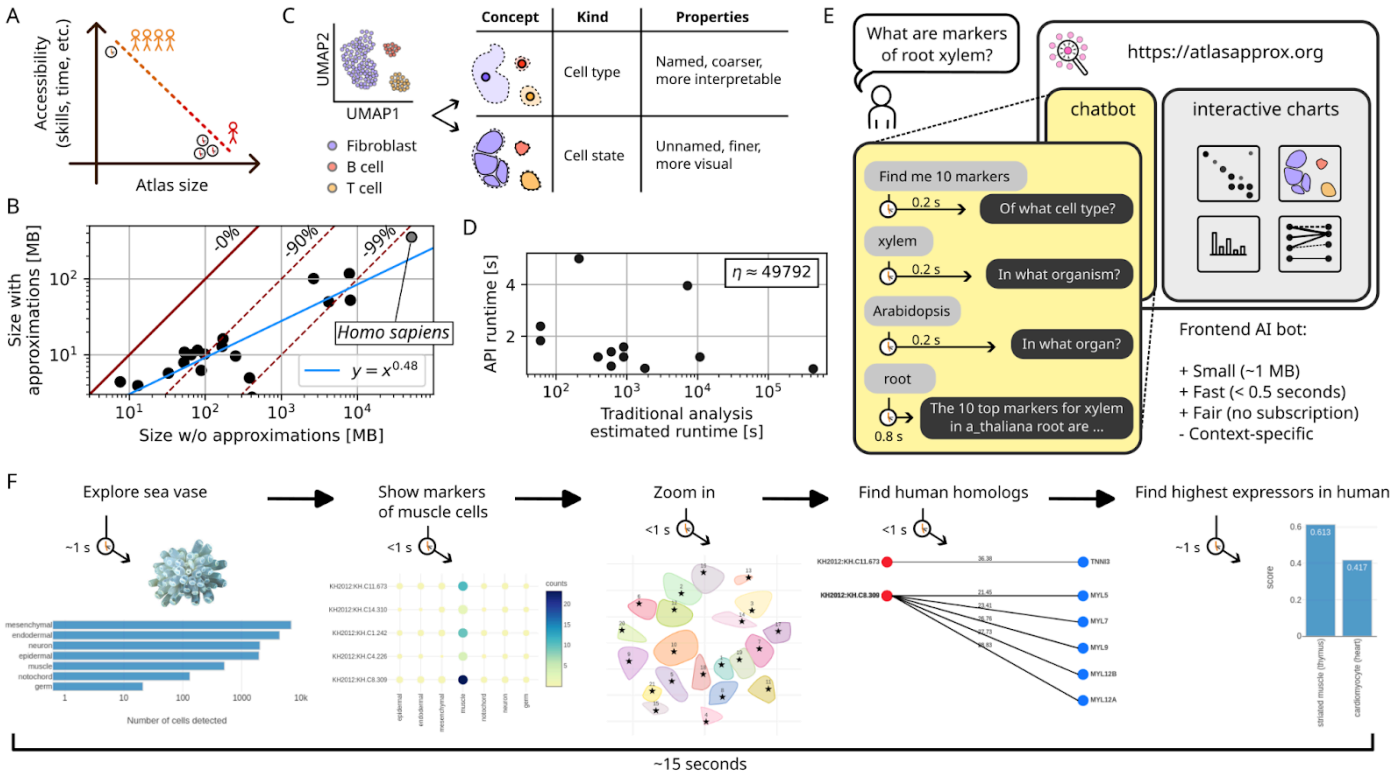

Figure 4: Cell approximations unlock atlas rapid exploration for everyone. A: Schematic of the cell atlas paradox: larger atlases contain more biological information yet incur technical skill and time barriers, reducing usability for most researchers. **B:** Cell approximations save up to 99% storage and memory compared to full atlases (solid burgundy line), with especially high compression ratios for larger atlases (blue line: empirical fit). **C:** Two types of cell approximations provided with properties specific for either one. **D:** Typical query-to-response times for various requests (marker genes, cell type abundance, etc.) of the atlasapprox programmatic interface (API), compared to typical times for traditional atlas analysis. η is the average speedup ratio across request types. **E:** Excerpt of the conversation-driven web interface to cell approximations across all 25 organisms, including a typical human-chatbot interaction and key differences from popular large language models. **F:** Typical zero-programming workflow on the web interface with estimated response times for each individual query.