A Cross-Species Generative Cell Atlas Across 1.5 Billion Years of Evolution: The TranscriptFormer Single-cell Model

James D Pearce[1], Sara E Simmonds[1,*], Gita Mahmoudabadi[1,*], Lakshmi Krishnan[1,*], Giovanni Palla[1], Ana-Maria Istrate[1], Alexander Tarashansky[1], Benjamin Nelson[1], Omar Valenzuela[1], Donghui Li[1], Stephen R Quake[1,2,3,#], Theofanis Karaletsos[1,#]

[1]Chan Zuckerberg Initiative, Redwood City, California, USA
[2]Department of Bioengineering, Stanford University; Stanford, CA, USA.
[3]Department of Applied Physics, Stanford University, Stanford, CA, USA.

*Contributed equally
[#]Corresponding authors: Theofanis Karaletsos (tkaraletsos@chanzuckerberg.com) and Stephen R. Quake (steve@quake-lab.org)

Abstract

Single-cell transcriptomics has revolutionized our understanding of cellular diversity, but integrating this knowledge across evolutionary distances remains challenging. Here we present TranscriptFormer, a family of generative foundation models representing a cross-species generative cell atlas trained on up to 112 million cells spanning 1.53 billion years of evolution across 12 species. TranscriptFormer jointly models genes and transcripts using a novel generative architecture, enabling it to function as a virtual instrument for probing cellular biology. In zero-shot settings, our models demonstrate superior performance on both in-distribution and out-of-distribution cell type classification, with robust performance even for species separated by over 685 million years of evolutionary distance. TranscriptFormer can also perform zero-shot disease state identification in human cells and accurately transfers cell type annotations across species boundaries. Being a generative model, TranscriptFormer can be prompted to predict cell type-specific transcription factors and gene-gene interactions that align with independent experimental observations. This work establishes a powerful framework for integrating and interrogating cellular diversity across species as well as offering a foundation for *in silico* experimentation with a generative single-cell atlas model.

Introduction

Single-cell genomics has revolutionized our understanding of cellular diversity of tissues and organisms. The dramatic improvement of single-cell technologies enabled the profiling of cell type transcriptomes of diverse organisms (*1–7*), leading to a holistic understanding of how cellular heterogeneity shapes developmental processes (*8*), tissue homeostasis and its dysregulation in disease states.

Scalable and cheap assays have enabled scientists to routinely generate scRNA-seq data from millions of cells, and have prompted large data collection centers to harmonize and standardize data around schemas and ontologies. These large datasets have spurred a variety of applications, and have established open access cell atlases (e.g., CZ CELLxGENE (*9*), Single Cell Portal (*10*), Human Cell Atlas Data Portal (*11*, *12*), Bgee (*13*)).

The recent emergence of single-cell foundation models (e.g., scGPT, Geneformer, UCE, scFoundation, GeneCompass) (*14–18*) was driven by the availability of extensive standardized datasets and progress in training large-scale machine learning systems. Foundation models have become a powerful framework for learning from large datasets and applying that knowledge to various downstream tasks.These models exhibit rich representations and novel emergent capabilities. Several single-cell foundation models, trained on human (scMulan (*19*), scFoundation (*17*), scGPT (*14*)), human/mouse (GeneCompass (*20*), Nicheformer (*21*), scPRINT (*22*)), or multispecies datasets (UCE (*16*), Xu et al.(*23*)), have demonstrated transfer learning capabilities, which often manifest in zero-shot settings (without retraining), including cell type annotation, metadata prediction, and gene-gene interaction network construction. While other methods are also proving powerful (e.g., scVI (*24*), Harmony(*25*)), the foundation model paradigm remains a promising approach for synthesizing vast amounts of data to learn a helpful representation for various downstream tasks as part of analysis pipelines.

Here, we present TranscriptFormer, a family of generative large-scale single-cell foundation models representing a digital cell atlas, trained on up to 112 million cells spanning 1.53 billion years of evolutionary history across species. TranscriptFormer is a generative autoregressive joint model over genes and their expression levels per cell across species, with a transformer-based architecture including a novel coupling between gene and transcript heads, expression-aware multi-head self-attention, causal masking, and a count likelihood to capture transcript-level variability. It represents a significant step toward constructing a generative cell atlas as a foundational component for virtual cell models (*26*). To train the models, we compiled a data corpus of scRNA-seq cell atlases from 8 animal species, a fungus and a protist and integrated it with the CZ CELLxGENE (CxG) database, which contains harmonized scRNA-seq data from human and mouse for a total of 12 species. We assessed how pretraining across a broader corpus of evolutionary data relates to generalizability in key tasks across species by constructing three versions of TranscriptFormer trained on subsets of the corpus: TF-Metazoa (all 12 species), TF-Exemplar (human and 4 model organisms), and TF-Sapiens (human-only data). TranscriptFormer was designed to incorporate gene orthology similarity, enabling the integration of transcript-level information across vast evolutionary distances. To this end, we

used Evolutionary Scale Model 2 (ESM-2) amino acid embeddings for coding genes, which capture evolutionary and functional similarity, allowing orthologous genes across species to be represented in a shared, species-agnostic embedding space (*27*), similar to UCE (*16*) and scPRINT (*22*).

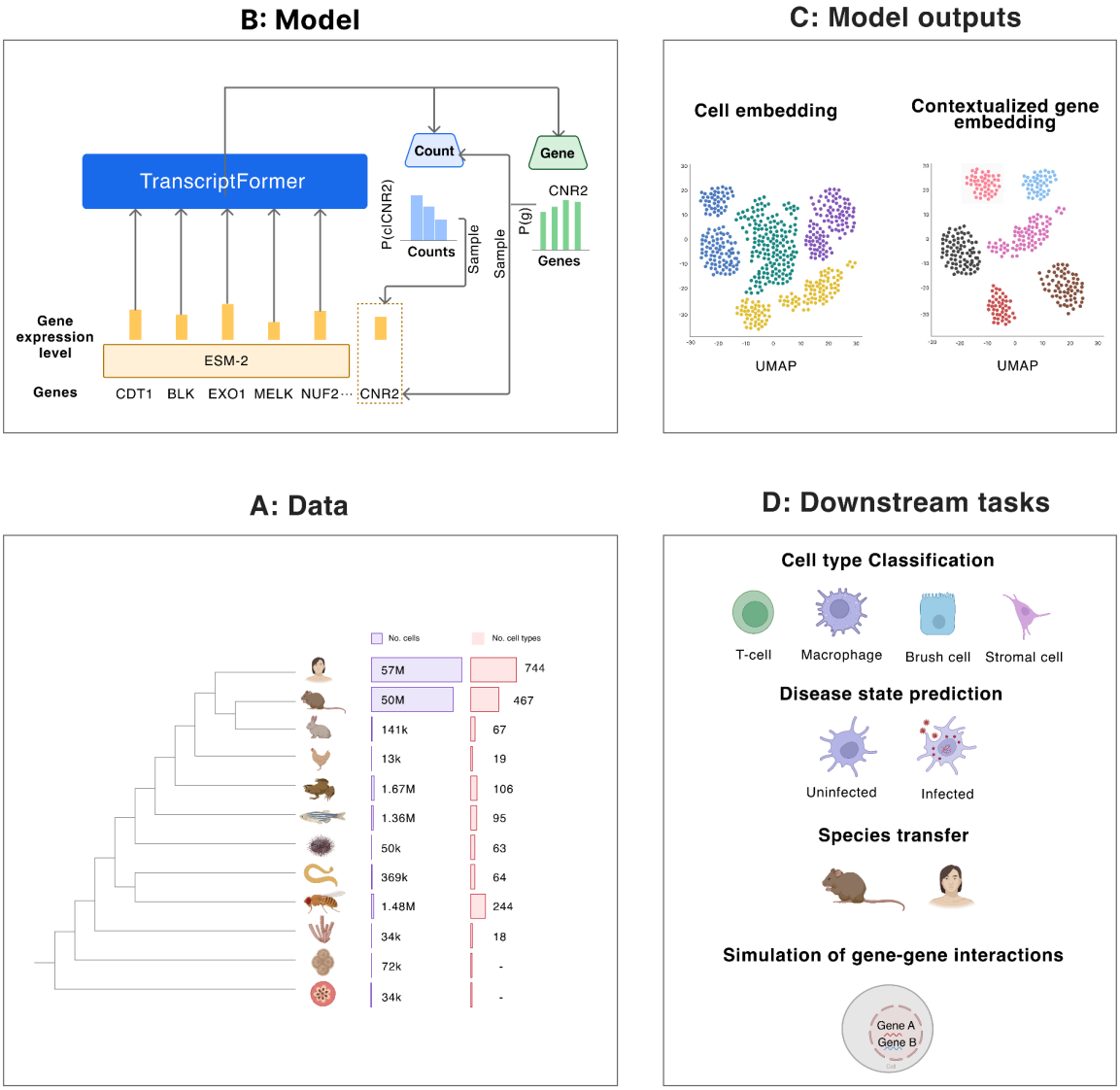

Fig. 1. Overview of TranscriptFormer pretraining data, model, output and downstream tasks. (A) The largest model TF-Metazoa was trained on a large compendium of 112 M cells from single-cell RNA-seq datasets across 12 species spanning 1.53 bn years of evolution. **(B)** Inputs to the model are cell sentences of protein-coding embeddings of transcripts and their associated raw counts. TranscriptFormer processes inputs to generate representations of expressed genes and cells. **(C)** Outputs of the model are learned cell embeddings and contextualized gene embeddings. **(D)** Embeddings can be used for a variety of downstream tasks, such as cell type classification, disease state prediction, and simulation of gene-gene interactions.

We evaluated TranscriptFormer's performance on various downstream tasks in zero-shot settings. The model demonstrates state-of-the-art performance in cell type classification and cell-level disease state prediction for both in-distribution species (seen during pretraining) and out-of-distribution species (unseen during pretraining). Investigation of the learned TranscriptFormer features and embeddings reveals the emergence of biological structure such as cell type-level groupings in the model's gene-level features. This indicates the model's ability to learn biologically relevant relationships by ingesting large amounts of single-cell data without labels. Finally, we demonstrate the utility of the model for biological inquiries at the example of transcription factor interactions by performing virtual experiments with the model using a novel form of model prompting aligned with the generative AI paradigm.

Results

The TranscriptFormer Model Family

TranscriptFormer is an autoregressive generative model of both protein coding gene tokens and corresponding gene expression counts arranged as a cell sentence with random ordering. It models probabilities over the sequence of expressed genes and their corresponding counts per cell, and can adjust these probabilities given the context of all preceding genes and expression levels. The model introduces innovations in architecture design, likelihood, and also captures joint beliefs about gene tokens and corresponding expression as an explicit feature of its design. It is trained via log-likelihood maximization according to its probabilistic setup. Notably, it is trained to model only cell sentences and is not provided with metadata such as cell type, species, or other labels apart from assay technology during pretraining. This focuses the model to learn biologically meaningful features based on cellular measurements only instead of human annotations. As with UCE, we represent gene tokens as corresponding ESM-2 protein embeddings in a shared embedding space. We explain the details of the model in the Methods section and provide an overview of the technical means used to build it. The outputs of the model are both these probabilities per gene token-expression tuple, as well as embeddings of cell sentences up to a position and entire cells (Fig. 1C). We use these outputs for downstream tasks (Fig. 1D), like cell type classification, disease state prediction, species transfer, and simulation of gene-gene interactions.

TranscriptFormer belongs to the emerging family of single-cell foundation models, but exhibits distinct characteristics compared to existing approaches. Like UCE it is trained across multiple species; however, unlike UCE, TranscriptFormer is trained as a generative model and can be used as both an embedding model and as a probabilistic generator of genes . We provide a visual sketch of the architecture in (Fig. 1B), and a deeper dive in (Methods).

We trained three versions of the TranscriptFormer architecture with different datasets:

TF-Metazoa: TranscriptFormer-Metazoa (444m trainable parameters) trained on 112 million cells from all twelve species described in (Methods, Fig. 1B). This dataset includes six vertebrate species (*Homo sapiens, Mus musculus, Oryctolagus cuniculus, Gallus gallus, Xenopus laevis, Danio rerio*), four invertebrate species (*Lytechinus variegatus, Caenorhabditis elegans, Drosophila melanogaster, Spongilla lacustris*), and two outgroups: a fungus (*Saccharomyces cerevisiae*) and a protist (*Plasmodium falciparum*). This phylogenetically diverse collection spans approximately 1.53 billion years of evolutionary history (*28*); Fig. 1A).

TF-Exemplar: TranscriptFormer-Exemplar (542m trainable parameters) trained on 110 million cells from human (*H. sapiens*) and four major model organisms: mouse (*M. musculus*), zebrafish (*D. rerio*), fruit fly (*D. melanogaste*r), and *C. elegans*.

TF-Sapiens: TranscriptFormer-Sapiens (368m trainable parameters) trained on 57 million cells exclusively from human (*H. sapiens*).

In addition, we utilize the concurrently proposed **ESM2-CE (ESM-2 Cell Embedding)** baseline method (unpublished work) based on ESM-2 embeddings. ESM2-CE uses the same input representations as TranscriptFormer but omits the generative transformer architecture. Instead, it computes a fixed-length cell embedding by averaging the ESM-2-derived amino acid embeddings across the input sequence. This method isolates the contribution of protein-based features and binarizes expression to average accordingly and provides a strikingly simple but effective baseline for assessing the added value of generative modeling in TranscriptFormer across evolution.

TranscriptFormer generalizes across 685 M years of cell type evolution

We are broadly interested in two key scientific hypotheses. First, will a generative cross-species transcript foundation model learn generalizable biological features given diverse input species? Second, will such a model learn biologically meaningful, interpretable biological structures without explicit supervision of species or cell type labels?

Given the emphasis of our work on cross-species atlases, we started with an inquiry to cell type classification across evolution, in particular for held-out species, including multiple species with matching tissues. We demonstrated that the TranscriptFormer family of models consistently exhibits leading performance across these tasks, showing that the combination of broad training across species and generative pretraining on transcripts and genes yields a strong universal single-cell RNA foundation model that captures meta-data like cell types robustly.
Moving beyond performance across species, we studied results for human data which also comprises the majority of our cell atlas data in CxG. We began by testing the generalization of TranscriptFormer relevant to human cell types against other baseline models on held-out human subjects that have not been trained on. We then assessed whether the generative nature of the

model offers an understanding of subtle shifts in transcriptional profiles related to disease in a held-out data set of COVID lung infection. Across the board we saw that TranscriptFormer matches or outperforms human-specific and other cross-species models in these tasks as well.

Out-of-distribution species generalization across evolutionary distances

To assess the generalization capacity of TranscriptFormer across phylogenetically distant species, we evaluated its performance on cell type classification using the linear probing protocol (see Methods) on out-of-distribution species, spanning an estimated 685 million years of evolutionary distance. We applied the model to five single-cell datasets derived from coral, zebrafish, sea lamprey, tropical clawed frog, and mouse lemur (see Methods). Each cell atlas represents a whole-organism or multi-tissue sampling, with the exception of sea lamprey, which is restricted to brain tissue. This evaluation tested the model's ability to generalize to previously unseen species across vast evolutionary distances.

Evidence supporting the role of evolutionary diversity in enhancing model generalization is provided by the superior performance of TF-Metazoa, which was trained on twelve phylogenetically diverse species. In evaluations across all out-of-distribution species, TF-Metazoa consistently outperformed both TF-Exemplar and UCE, achieving the highest overall average F1 score (0.778 ± 0.002) (Fig. 2A, B). Of note is that the additional species pretraining data for TF-Metazoa was relatively small (2 million cells), but benefits this task disproportionately. This suggests that the diversity of the input data distribution is a core aspect of generalization, apart from sheer size of the pretraining dataset. TF-Exemplar also performed well (0.770 ± 0.002), but both models substantially outperformed UCE (0.701 ± 0.002) (Fig. 2A, B), underscoring the advantage of evolutionarily broad pretraining for generalization to unseen biological contexts.

While all models show decreasing performance with increasing evolutionary distance from humans, TranscriptFormer models maintain high F1 scores (>0.65) even for species separated by over 685 million years of evolution (stony coral) (Fig. 2A, B). In contrast, benchmark models show steeper performance drops, with UCE's performance deteriorating dramatically for sea lamprey and stony coral (F1 ≤0.5, Fig. 2A, B). This resilience to evolutionary divergence suggests that TranscriptFormer effectively captures conserved gene expression signatures that define evolutionarily related cell types across diverse species.

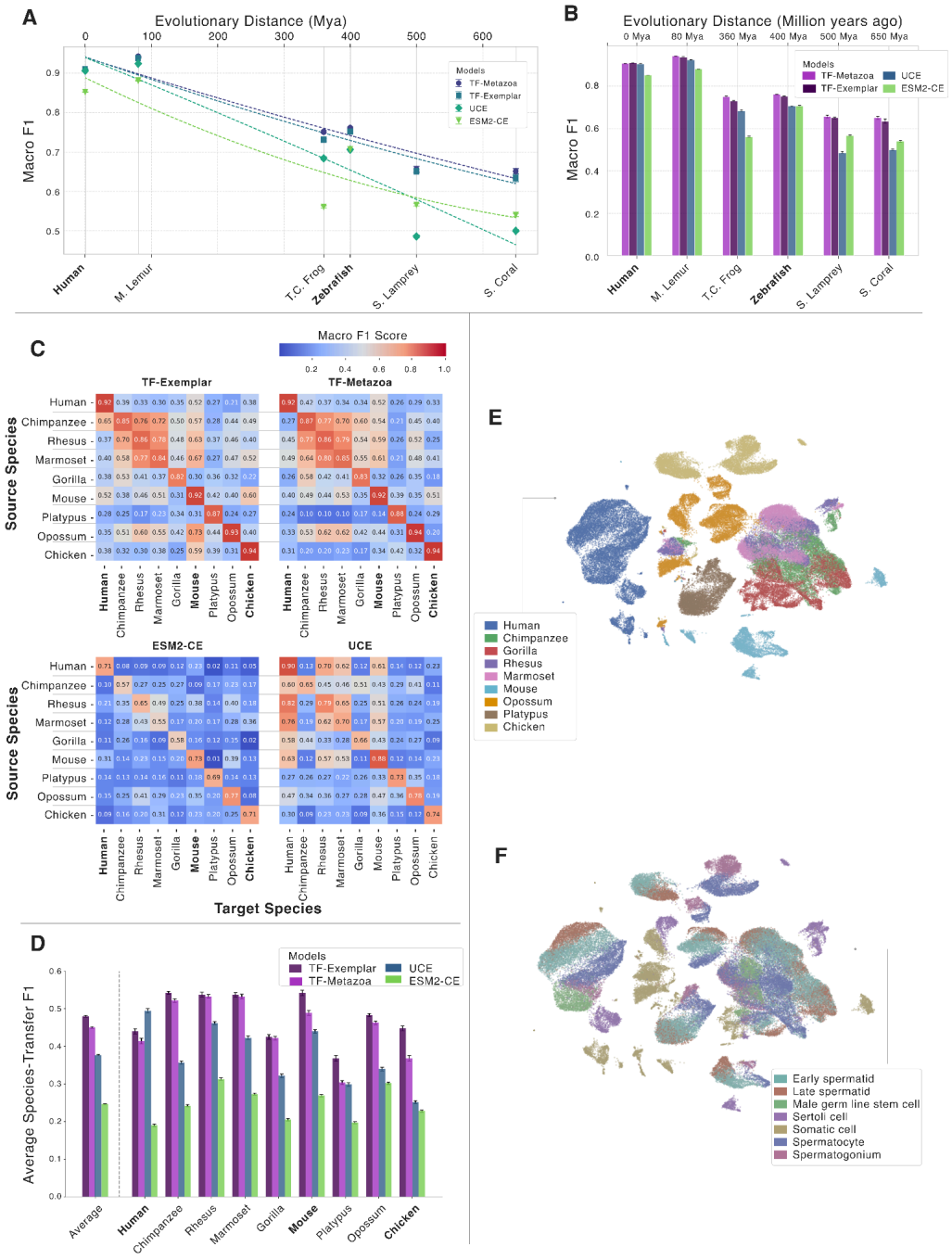

Fig. 2. TranscriptFormer enables accurate cross-species cell type classification across hundreds of millions of years of evolution, even for held-out species. (A, B) Cell type classification performance on cell atlases (macro F1 score) as a function of evolutionary distance from humans. TF-Metazoa and TF-Exemplar outperform UCE and EMS2-CE across all species and maintain stronger performance at larger evolutionary distances. Bold species are included in TranscriptFormer pretraining. **(C)** Transfer learning of cell type classification on testis tissue dataset across mammals and chicken. Confusion matrix showing macro F1 scores for all target and reference species pairs for each model. **(D)** Bar plot showing classification performance by species confirms that TranscriptFormer maintains higher

accuracy across held-out species (e.g., platypus, opossum) compared to UCE and ESM2-CE. Bold species are included in TranscriptFormer pretraining. **(E)** UMAP visualizations and PCAs colored by species and cell types.

A spermatogenesis case study reveals effective cell type transfer across species boundaries

In addition to assessing TranscriptFormer generalization capabilities across out-of-distribution cell atlases, we also studied its ability to integrate multiple species in embedding space and map cell types across multiple species using a study of spermatogenesis anchored on a rich dataset covering nine vertebrate species (*29*). Specifically, this dataset comprises snRNA-seq data from testes across major mammalian and avian lineages, including human, four non-human primates (chimpanzee, gorilla, rhesus monkey, marmoset), mouse, opossum, platypus, and chicken. For the analysis, we employed a reference mapping approach (see Methods) to evaluate the ability to transfer cell types annotated on one species to others, a form of transfer learning across species.

On this dataset, TranscriptFormer models consistently outperform benchmarks, with TF-Exemplar achieving the highest average transfer learning F1 score across all species (0.480 ± 0.002), followed by TF-Metazoa (0.450 ± 0.002), UCE (0.377 ± 0.002), and ESM 2-CE (0.246 ± 0.001) (Fig. 2C). Notably, TranscriptFormer models almost double the transfer learning performance of the ESM2-CE baseline (F1 = 0.480 vs. 0.246; Fig. 2C). This dramatic improvement demonstrates that TranscriptFormer has learned substantial cross-species generalization capabilities by internalizing a broad pretraining distribution and modeling transcripts beyond what can be achieved with a stronger reliance on protein sequence features as used in the ESM2-CE baseline or UCE.

TranscriptFormer outperforms UCE in transfer learning across all species when averaging over target and source, with the exception of human, where UCE (0.495 ± 0.006) outperforms both TF-Exemplar (0.440 ± 0.007) and TF-Metazoa (0.414 ± 0.006). Notably, TranscriptFormer maintains strong species transfer even for more distant species such as chicken (310 Mya), with species-average scores of 0.448 ± 0.006 for TF-Metazoa and 0.368 ± 0.007 for TF-Exemplar, compared to 0.251 ± 0.004 for UCE and 0.228 ± 0.003 for ESM2-CE (Fig. 2D). These results suggest that TranscriptFormer has a deeper understanding of generalizable biology, for instance evolutionarily conserved gene expression programs, enabling effective knowledge transfer even between distantly related organisms.

Dimensionality reduction of TF-Metazoa embeddings shows substantial overlap among non-human primate species (chimpanzee, rhesus monkey, marmoset, and gorilla), suggesting conserved transcriptional programs during spermatogenesis (Fig. 2E). We observed that human cells cluster separately from other primates, likely due to technical batch effects (SI. Fig. 2). Despite this technical variation, cell types cluster consistently across species (Fig. 2E), highlighting the biological relevance of the learned embeddings. TranscriptFormer learns to group cells in a biologically relevant fashion by species and cell types (Fig. 2E), without the model being trained or run with species or cell type labels.

TranscriptFormer achieves competitive performance on human cell type classification

In order to assess cell type classification performance on unseen human data, we used Tabula Sapiens 2.0, a comprehensive human reference cell atlas containing single-cell data across 26 tissues (Methods) (*30*). To ensure independence from our pretraining data, we filtered the dataset to include nine donors (TSP17 to TSP30) that were absent from Tabula Sapiens 1.0. Given its release date of December 4, 2024, Tabula Sapiens 2.0 was explicitly held out from TranscriptFormer pretraining and is unlikely to be included in the pretraining data of any of the comparison models, making it a suitable evaluation set.

When evaluated on the Tabula Sapiens 2.0 dataset, TranscriptFormer matches or slightly outperforms existing state-of-the-art models. Using a linear probing approach (Methods), our TF-Metazoa variant achieved the highest average macro F1 score (0.910 ± 0.001) across 26 human tissues, followed by TF-Exemplar (0.907 ± 0.001) and TF-Sapiens (0.906 ± 0.001), compared to the next best model, UCE (0.906 ± 0.001) (Fig. 3A).

The results reveal that this benchmark is approaching saturation, with TranscriptFormer and baseline models like UCE, scGPT, and Geneformer all achieving close to perfect accuracy on the majority of cell types (Fig. 3B). Despite this ceiling effect, the multi-species variants (TF-Metazoa and TF-Exemplar) performed marginally better than the human-only model (TF-Sapiens) despite the same number of active parameters during inference and identical pretraining protocols, suggesting cross-species pretraining does not dilute performance for human data and may even improve it. Performance differences across the comparison models and TranscriptFormer are most pronounced for challenging cell types like myeloid leukocytes, T-cells, and innate lymphoid cells, where the TranscriptFormer variants, scGPT, and UCE maintain higher classification F1 scores where other models exhibit more substantial performance drops (Fig. 3B, SI. Fig. 1).

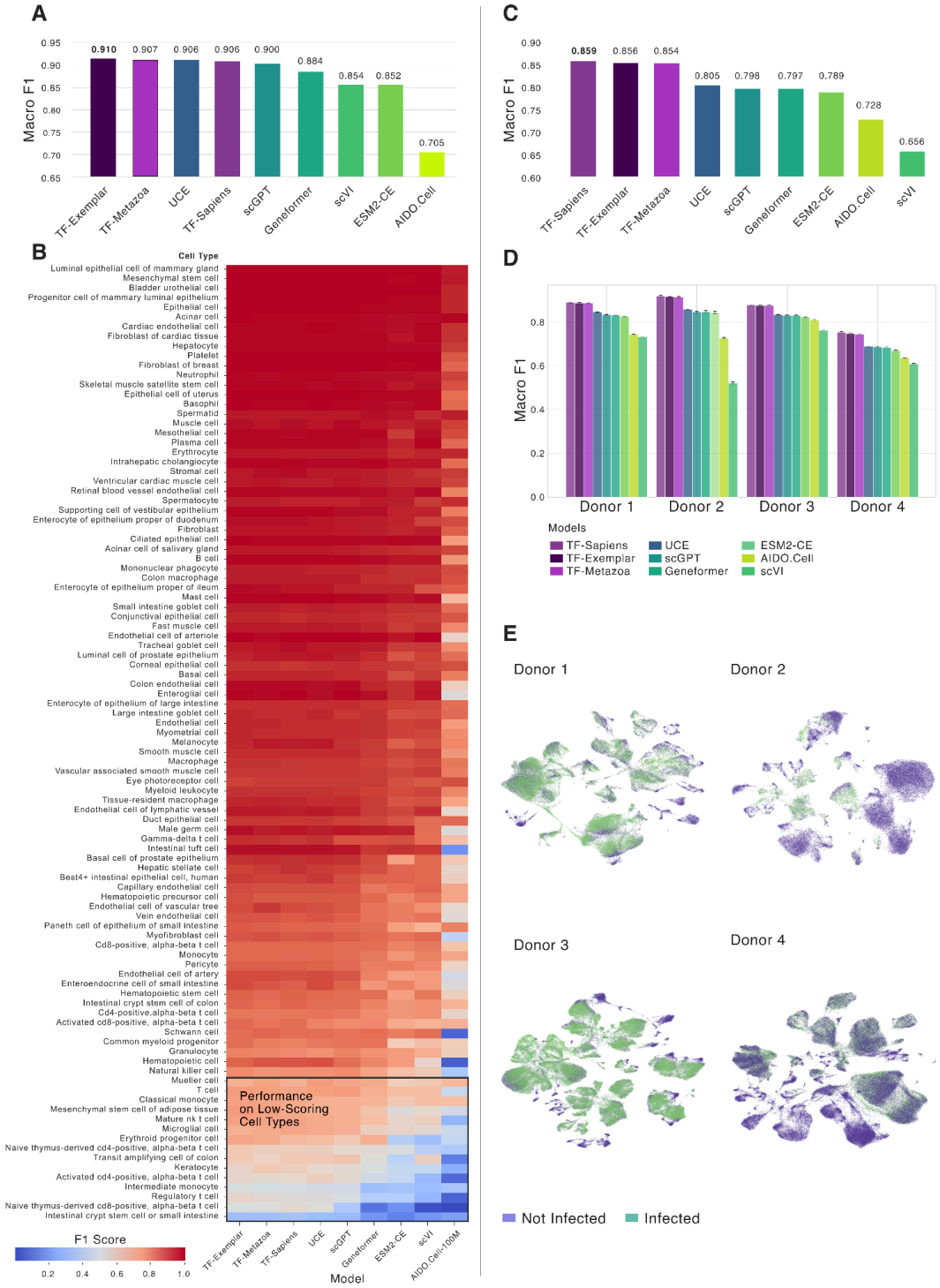

Fig. 3. TranscriptFormer outperforms existing models in cell type classification and disease state prediction. (A) Macro F1 scores for cell type classification on the Tabula Sapiens 2.0 benchmark dataset, averaged across 26 tissues. **(B)** Per-cell-type F1 scores across models reveal that most models perform well on abundant or easily classifiable cell types. TranscriptFormer variants maintain higher performance on low-scoring cell types such as T-cells and their subtypes (inset box). Multi-species pretraining (TF-Metazoa, TF-Exemplar) confers a performance advantage over the human-only TF-Sapiens, even

under matched training and model size conditions. **(C)** SARS-COV-2 dataset average disease state prediction F1 scores of uninfected and infected cells across models. **(D)** SARS-CoV-2 dataset disease state prediction F1 scores for each of the four donors across models. **(E)** TF-Sapiens embedding UMAP of uninfected (purple) and infected (green) cells for each donor.

Robust disease state identification from single-cell human transcriptomes

To assess the capability of TranscriptFormer to distinguish transcriptional profiles within a species in cell states it has not trained on, we assessed its utility for disease classification on a human lung SARS-COV-2 dataset (*31*) which contains non-infected and infected cells across four donors.

We applied a linear probing methodology per donor (Methods). The results show that TF-Sapiens achieves the highest macro F1 score (0.859), with TF-Exemplar and TF-Metazoa following closely (0.856 and 0.854, respectively) (Fig. 3C). These results surpass baseline models, including UCE (0.805), scGPT (0.798), and Geneformer (0.797) (Fig. 3C), and suggest that generative pretraining for TranscriptFormer is useful for capturing novel shifts in transcriptional profiles.

Differences in cell type susceptibility to infection could, in principle, confound this analysis—raising the possibility that the model might rely primarily on cell type identity to infer disease state. However, upon closer examination, we find that UCE performs comparably to TranscriptFormer variants in cell type classification (Fig. 3A), yet exhibits a more pronounced decline in performance on the disease labeling task. This divergence suggests that the observed results are not solely driven by a cell type-based shortcut, but instead reflect a deeper relationship between cellular features and disease state.

UMAP visualizations of TranscriptFormer embeddings reveal separation in some regions between infected and non-infected cell populations (Fig. 3D), highlighting the model's ability to capture infection-induced transcriptional changes. This separation demonstrates TranscriptFormer's potential for studying host-pathogen interactions at single-cell resolution, enabling researchers to identify cell-specific responses to viral infection and potentially uncover novel mechanisms of pathogenesis and cellular defense.

Context-aware gene representations encode multi-level biological structure

Unlike cell embeddings that summarize overall transcriptional states by averaging across a model’s gene embeddings, the probing of individual gene embeddings provides a more granular window into a model’s learned representations of different biological contexts and enable a more detailed understanding of the transcriptomic trends observed through cell embeddings. More specifically, TranscriptFormer’s Contextualized Gene Embeddings (CGEs) are the gene-level representations of the model after processing the context from all genes that precede it in a cell sentence (Fig. 4A, see Methods: CGE) (which if averaged would yield

cell-embeddings). As we will demonstrate in the following paragraphs, studying CGEs enables us to explore the zero-shot learnings of the model such as cell type, tissue and donor contexts.

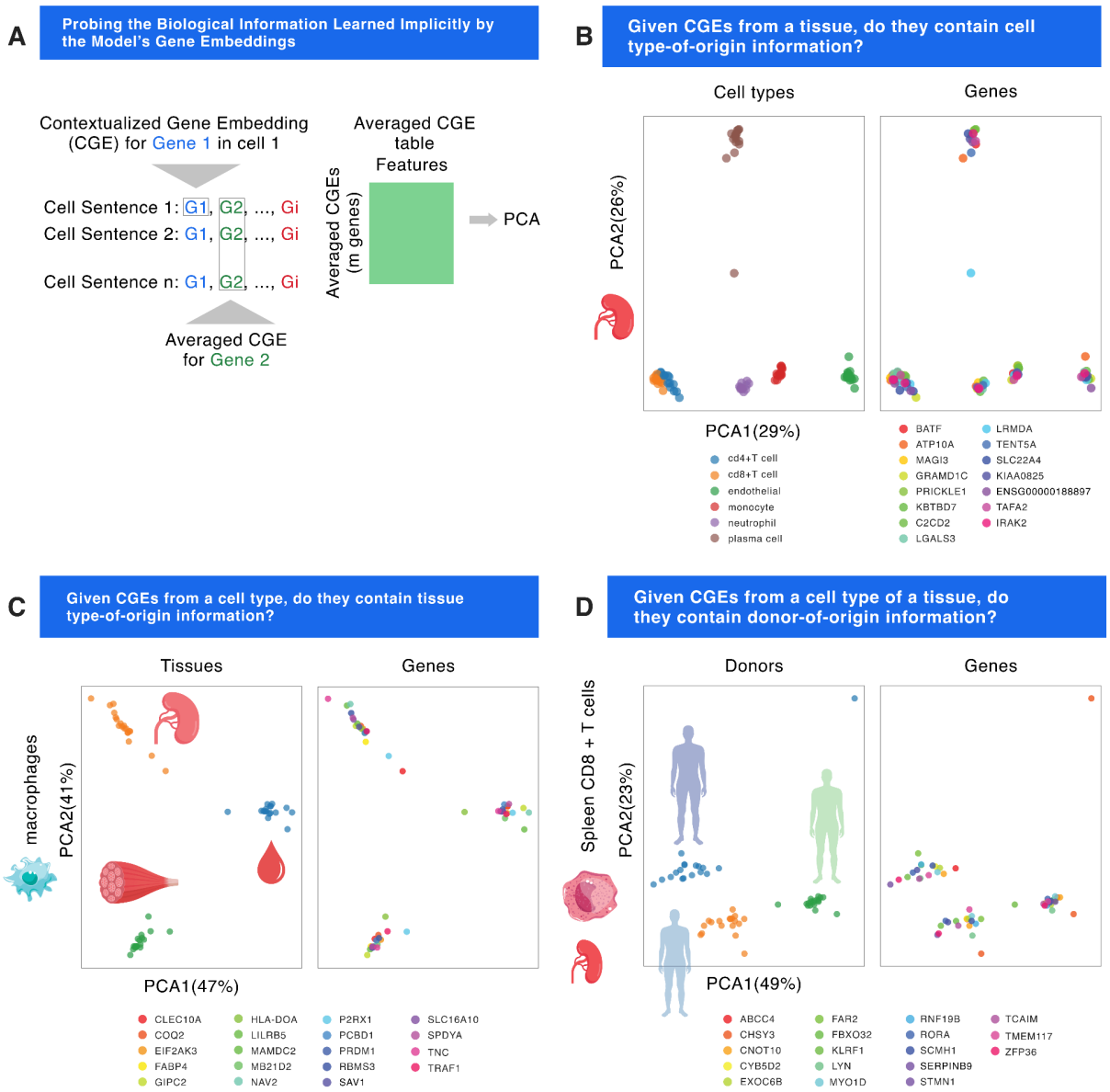

Fig. 4. Contextualized Gene Embeddings (CGEs) can be used as probes into the model's implicitly learned biological contexts. A) Schematic representation of the averaged CGEs underlying the PCA plots shown in B-D. B) PCA of CGEs from spleen averaged based on six cell types, color-coded by cell type and gene. C) PCA of CGEs from macrophages across three different tissues (spleen, muscle and blood) averaged by tissue and color-coded by tissue and gene. D) PCA of CGEs from spleen CD8+ T-cells averaged across three different donors (TSP21, TSP25, TSP27) and color-coded by donor and gene.

Model inference on three tissues from an unseen dataset (Tabula Sapiens 2.0) demonstrates that despite identical input ESM-2 embeddings for given genes, the resulting CGEs cluster primarily by cell type. This clustering pattern persists across all three tissues, validating the model's ability to capture cell type-specific context across a diverse range of genes (Fig. 4B, SI. Fig. 3), without having access to cell type labels. Further analysis of shared cell types between tissues reveals that CGEs incorporate the subtle variations in transcriptomic profiles of identical cell types across different tissues in a zero-shot manner, highlighting the context-sensitivity of the model's gene representations (Fig. 4C, Supplemental Information SI. Fig. 4).

TranscriptFormer also exhibits sensitivity to donor-specific transcriptomic variations, with CGEs from the same cell type and tissue separating based on donor (Fig. 4D, SI. Fig. 5). We further show that donor effects are stronger than batch effects, however we cannot completely rule out batch effects (SI. Fig. 8). This finding underscores TranscriptFormer's capacity to detect inter-individual variability in part stemming from biological factors such as genetic polymorphisms or environmental exposures. Moreover, we used variance partitioning analysis to quantify the contribution of cell type, tissue and donor contexts to CGEs. We show that cell type information dominates the CGE space (explaining >95% of variance in PC1 and PC2, after accounting for tissue and donor), while tissue and donor information contribute as secondary factors (explaining <2% in PC1 and <7% in PC2, after accounting for the other two variables) with this hierarchy remaining consistent across various gene selection criteria (see Methods, SI. Fig. 6, SI. Fig. 7).

These results demonstrate TranscriptFormer's ability to encode cell type, tissue type, and donor information in a zero-shot manner without explicit category tokens available during pretraining or testing. This work demonstrates the utility of gene representations and their ability to capture nuanced biological contexts in gene embeddings provides a useful foundation for future analyses of cell type specific and potentially donor-specific gene regulation, advancing our interpretation of complex transcriptomic data.

Using TranscriptFormer as a virtual instrument via prompting

TranscriptFormer models the joint distribution of genes and expression levels across cells, similar to a generative language model. We exploited this generative capability of the model to utilize it as a virtual instrument, and demonstrated how to prompt the model to represent complex inquiries into cellular biology that would otherwise be tackled by compiling datasets laboriously.

Inferring gene-gene regulatory relationships

Our first prompt asked the model to predict interactions between transcription factors and other protein coding genes using its generative interface. We obtained point-wise conditional mutual information (PMI) estimates between these transcription factors-gene pairs (see Methods), which we used to identify high-confidence interactions (adjusted p-value <= 0.05) and cross-reference with known biological relationships documented in the STRING database v12.0

(*32*). PMI was chosen because it highlights specific co-occurrences between genes and transcription factors while correcting for genes with high baseline expression.

Our model accurately recovers key regulatory targets for six transcription factors with diverse biological functions (Fig. 5A). **E2F8**, a repressor of E2F-activated genes, is linked to replication factors like *CDT1* and mitotic regulators such as *KIF18B* (*33*, *34*). The model predicts 227 significant *E2F8* targets, of which 87 are known in (*32*). **FOXM1**, a G_2/M-phase driver, is associated with targets like *TPX2*, *HJURP*, and *KIF18B*, in line with its role in mitotic gene activation (*35*, *36*)(*37*). Of 224 predicted FOXM1 targets, 105 are known. **PTTG1** predictions highlight cell cycle genes such as *BIRC5* and *PBK*, consistent with its transcriptional activation of *c-MYC* and *Cyclin D3* (*38*). The model identified 106 *PTTG1* targets, 70 of which are known. **SPIB** targets include B-cell markers like *CD19* and *CXCR5*, supporting its function in immune lineage commitment (*39*, *40*). Of 23 predicted *SPIB* targets, 9 are known. **MYBL2** is correctly linked to late cell-cycle regulators (*CDC45*, *BIRC5*), reflecting its cooperation with FOXM1 in mitotic transcription (*41*). Among 87 predicted *MYBL2* targets, 54 are known. Finally, **UHRF1** is associated with replication stress and checkpoint factors like *CHEK1* and *TIMELESS*, consistent with its epigenetic role in coordinating DNA methylation and genome stability (*42*, *43*). Of the 90 predicted *UHRF1* targets 54 are known.

Inferring cell type-specific transcription factor interactions

We then created more complex prompts by conditioning the model on a set of marker genes corresponding to specific human cell types from the (unseen in pretraining) Tabula Sapiens 2.0 dataset and inquired for probabilities of generating transcription factors. Thus prompted, TranscriptFormer predicted a rich collection of cell type-specific transcription factors. Our model-generated heatmap (Fig. 5B) recapitulates structural features observed in empirical data from Tabula Sapiens 2.0. We employed a methodology that parallels the approach used in the Tabula Sapiens 2.0 study, with the key distinction that the gene expression levels were generated by the model rather than measured experimentally. We arranged transcription factors according to the cell type for which they showed the highest probability of expression, creating a visualization that mirrors Figure 3A in the Tabula Sapiens 2.0 publication (*30*). Our generated heatmap exhibits remarkably similar global structures: vertical bands corresponding to ubiquitously expressed transcription factors that function across diverse cell types, and a distinctive diagonal trace representing cell type-specific transcription factors with highly specialized expression patterns. While different cell type annotations between our evaluation data and the original Tabula Sapiens 2.0 dataset prevent a one-to-one comparison, we did recover several key transcription factors and broad cell types associations mentioned in (*30*), including *IKZF1* and *IKZF3* for T-cells, *CDX1* for stem cells, *SOX30* and *TCFL5* for male germ cells, and *HEY1*, *ERG*, and *LHX6* for endothelial cells.

This ability to generate probabilistic associations based on learned patterns demonstrates TranscriptFormer's internalization of complex gene regulatory networks during pretraining, enabling biologically meaningful predictions even for unseen data distributions. More importantly, this capability sketches a future path for using generative models as virtual

instruments in place of cell atlases, whereby an integrated view of cells is not just a look-up table of data, but also a query-able interactive model of the knowledge contained that can simulate complex views into the data.

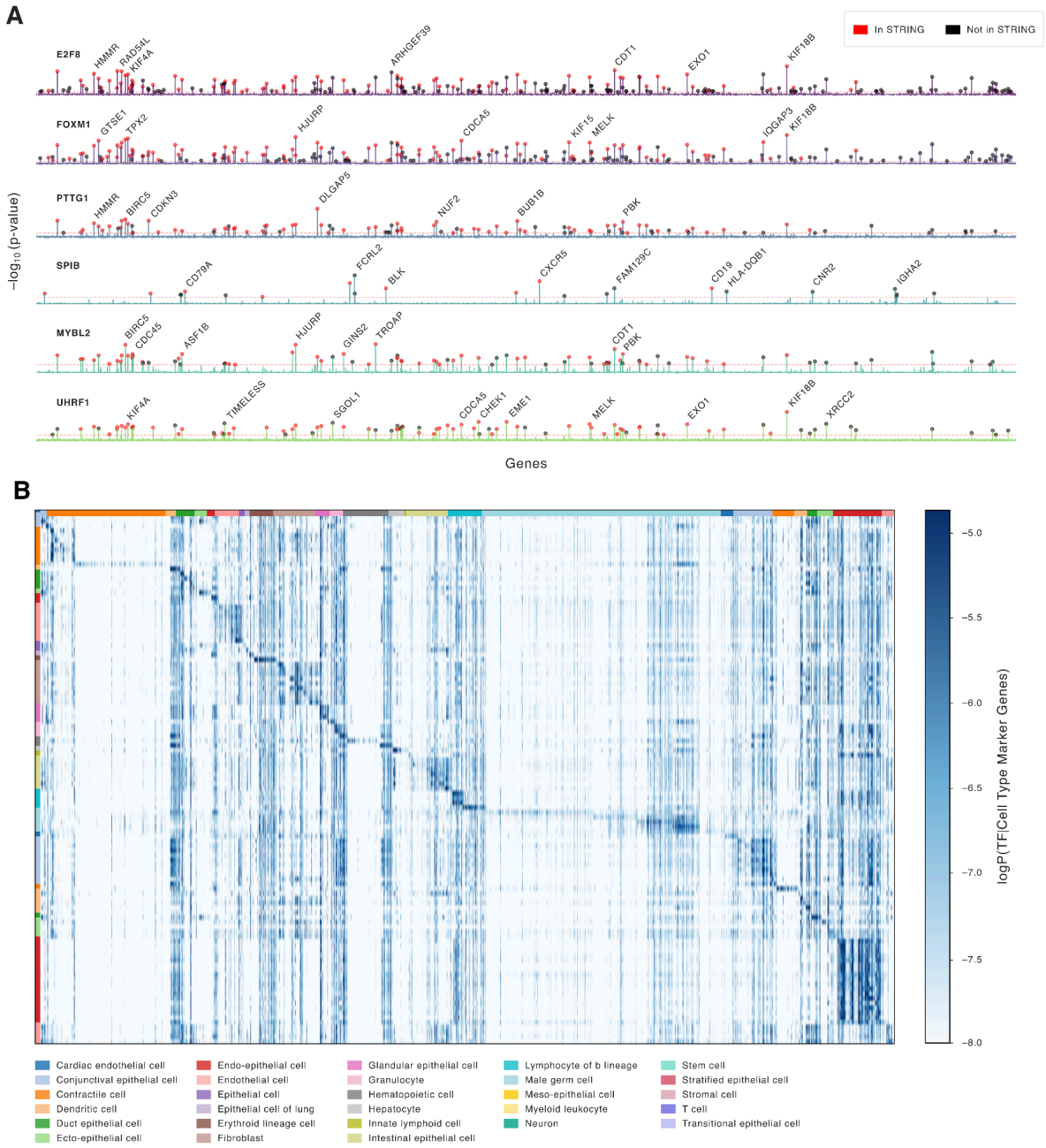

Fig. 5. TranscriptFormer learns biologically meaningful gene-gene relationships and predicts cell type-specific human transcription factors. (A) Pointwise conditional mutual information (PMI) spectra for selected transcription factors, computed using TranscriptFormer's autoregressive architecture showing significantly enriched gene-level associations, with genes annotated in STRING v12.0 shown in red. **(B)** TranscriptFormer-predicted gene expression probabilities for transcription factors across 112 human cell

types from Tabula Sapiens 2.0. TF-Sapiens generates known structural features: a diagonal pattern of cell type-specific transcription factors and vertical bands of broadly expressed regulators.

Conclusions

TranscriptFormer represents a significant advancement in biological foundation models by successfully integrating transcriptomic data across an unprecedented evolutionary span of 1.53 billion years. Our findings demonstrate that generative pretraining on both genes and expression counts from diverse species creates representations with emergent biological properties that surpass previous approaches. The performance of TF-Metazoa over single-species models on both in-distribution and out-of-distribution tasks provides compelling evidence that broader evolutionary pretraining enhances biological generalization. This suggests that evolutionarily conserved gene expression signatures can be effectively captured by foundation models, enabling robust cross-species analysis of cell types and states.

Beyond benchmarking, TranscriptFormer introduces a paradigm shift in how we interact with cellular data by functioning as a virtual instrument for biological inquiry. The model's ability to generate meaningful cell type-specific transcription factor predictions and gene-gene interactions through prompting demonstrates its potential as a computational tool for hypothesis generation. Context-aware gene embeddings further reveal how TranscriptFormer learns to encode multi-level biological structure—from cell types to tissues to donor-specific variations—without explicit supervision. This emergent capability points toward future models that could serve not merely as repositories of data but as interactive knowledge bases capable of simulating complex cellular phenomena.

The development of TranscriptFormer signals an advance toward a unified generative cell atlas that transcends species boundaries and enables comparative biology at unprecedented scale. As single-cell technologies continue to generate vast datasets across the tree of life, foundation models like TranscriptFormer will become increasingly valuable for integrating this knowledge, identifying universal principles of cellular organization, and potentially predicting cell states under novel conditions.

Future work will focus on expanding the diversity of species represented, and mechanisms to better integrate diverse datasets by handling batch effects in a unified fashion, a limitation commonly shared across many foundation models. Similar to other single-cell foundation models trained on cell atlas datasets, TranscriptFormer is also not specialized for zero-shot perturbation prediction, which is a fruitful direction for future technical enhancements. Further, it will be exciting to work on incorporating additional modalities beyond transcriptomics, and developing more sophisticated prompting strategies to fully realize the potential of generative foundation models as virtual instruments for biological discovery. TranscriptFormer establishes a framework for this vision, demonstrating how machine learning can help unify our understanding of cellular diversity across evolutionary time.

Data availability

Pretraining human and mouse datasets are available from CZ CELLxGENE (see Extended Data Table 1 for dataset IDs) and GEO (GSE247719). Pretraining datasets are publicly available for *Caenorhabditis elegans* (GSE126954, GSE229022, GSE98561), *Danio rerio* (GSE202639, Zebrahub), *Drosophila melanogaster* (Aging Fly Cell Atlas, Fly Cell Atlas, Alzheimer's Disease Fly Cell Atlas), *Gallus gallus* (GSE181577), *Lytechinus variegatus* (GSE184538), *Oryctolagus cuniculus* (RabbitGastrulation2022), *Plasmodium falciparum* (Malaria Cell Atlas), *Saccharomyces cerevisiae* (GSE125162), *Spongilla lacustris* (GSE134912), *Xenopus laevis* (GSE195790) (see Extended Data Table 2 for details). Evaluation datasets: *Danio rerio* (GSE234216, GSE130487), *Microcebus murinus* (Tabula Microcebus), *Petromyzon marinus* (E-MTAB-11087), *Stylophora pistillata* (GSE166901), *Xenopus tropicalis* (GSE113074), Spermatogenesis dataset (*Homo sapiens* E-MTAB-11063, *Pan troglodytes* E-MTAB-11064, *Gorilla gorilla* E-MTAB-11065, *Macaca mulatta* E-MTAB-11068, *Callithrix jacchus* E-MTAB-11069, *Mus musculus* E-MTAB-11071, *Monodelphis domestica* E-MTAB-11072, *Ornithorhynchus anatinus* E-MTAB-11070, *Gallus gallus* E-MTAB-11073, *Homo sapiens* Tabula Sapiens 2.0 (subset of Tabula Sapiens), COVID-19 human lung (CZ CELLxGENE) (see Extended Data Table 2 for details).

Code availability

The pretrained models are publicly available on CZI's virtual cells platform at https://virtualcellmodels.cziscience.com/paper/transcriptformer-2025-04 and GitHub at https://github.com/czi-ai/transcriptformer

Acknowledgements

We gratefully acknowledge the CZI AI infrastructure team for providing support with the GPU cluster required for training our models. We also want to thank the CZI Brand and Creative team for support in making and improving visual assets for this work. The CZI SciTech team was instrumental in sharing assets on our platform. We want to also acknowledge the scientific and technology leadership teams across the Chan Zuckerberg Science ecosystem including Andrea Califano, Sandra Schmid, Shana Kelley, Scott Fraser, Amy Herr, and Patricia Brennan for continuous feedback on this work.

References and Notes

1. J. M. Musser, K. J. Schippers, M. Nickel, G. Mizzon, A. B. Kohn, C. Pape, P. Ronchi, N. Papadopoulos, A. J. Tarashansky, J. U. Hammel, F. Wolf, C. Liang, A. Hernández-Plaza, C. P. Cantalapiedra, K. Achim, N. L. Schieber, L. Pan, F. Ruperti, W. R. Francis, S. Vargas, S. Kling, M. Renkert, M. Polikarpov, G. Bourenkov, R. Feuda, I. Gaspar, P. Burkhardt, B. Wang, P. Bork, M. Beck, T. R. Schneider, A. Kreshuk, G. Wörheide, J. Huerta-Cepas, Y. Schwab, L. L. Moroz, D. Arendt,

Profiling cellular diversity in sponges informs animal cell type and nervous system evolution. *Science* **374**, 717–723 (2021).

2. A. J. Massri, L. Greenstreet, A. Afanassiev, A. Berrio, G. A. Wray, G. Schiebinger, D. R. McClay, Developmental single-cell transcriptomics in the *Lytechinus variegatus* sea urchin embryo. *Development* **148**, dev198614 (2021).

3. A. G. Cole, J. Steger, J. Hagauer, A. Denner, P. Ferrer Murguia, P. Knabl, S. Narayanaswamy, B. Wick, J. D. Montenegro, U. Technau, Updated single cell reference atlas for the starlet anemone *Nematostella vectensis*. *Front. Zool.* **21**, 8 (2024).

4. J. A. Briggs, C. Weinreb, D. E. Wagner, S. Megason, L. Peshkin, M. W. Kirschner, A. M. Klein, The dynamics of gene expression in vertebrate embryogenesis at single-cell resolution. *Science* **360** (2018).

5. D. E. Wagner, C. Weinreb, Z. M. Collins, J. A. Briggs, S. G. Megason, A. M. Klein, Single-cell mapping of gene expression landscapes and lineage in the zebrafish embryo. *Science* **360**, 981–987 (2018).

6. Tabula Muris Consortium, A single-cell transcriptomic atlas characterizes ageing tissues in the mouse. *Nature* **583**, 590–595 (2020).

7. R. Shahan, C.-W. Hsu, T. M. Nolan, B. J. Cole, I. W. Taylor, L. Greenstreet, S. Zhang, A. Afanassiev, A. H. C. Vlot, G. Schiebinger, P. N. Benfey, U. Ohler, A single-cell *Arabidopsis* root atlas reveals developmental trajectories in wild-type and cell identity mutants. *Dev. Cell* **57**, 543–560.e9 (2022).

8. C. Qiu, B. K. Martin, I. C. Welsh, R. M. Daza, T.-M. Le, X. Huang, E. K. Nichols, M. L. Taylor, O. Fulton, D. R. O'Day, A. R. Gomes, S. Ilcisin, S. Srivatsan, X. Deng, C. M. Disteche, W. S. Noble, N. Hamazaki, C. B. Moens, D. Kimelman, J. Cao, A. F. Schier, M. Spielmann, S. A. Murray, C. Trapnell, J. Shendure, A single-cell time-lapse of mouse prenatal development from gastrula to birth. *Nature* **626**, 1084–1093 (2024).

9. CZI Cell Science Program, S. Abdulla, B. Aevermann, P. Assis, S. Badajoz, S. M. Bell, E. Bezzi, B. Cakir, J. Chaffer, S. Chambers, J. M. Cherry, T. Chi, J. Chien, L. Dorman, P. Garcia-Nieto, N. Gloria, M. Hastie, D. Hegeman, J. Hilton, T. Huang, A. Infeld, A.-M. Istrate, I. Jelic, K. Katsuya, Y. J. Kim, K. Liang, M. Lin, M. Lombardo, B. Marshall, B. Martin, F. McDade, C. Megill, N. Patel, A. Predeus, B. Raymor, B. Robatmili, D. Rogers, E. Rutherford, D. Sadgat, A. Shin, C. Small, T. Smith, P. Sridharan, A. Tarashansky, N. Tavares, H. Thomas, A. Tolopko, M. Urisko, J. Yan, G. Yeretssian, J. Zamanian, A. Mani, J. Cool, A. Carr, CZ CELLxGENE Discover: a single-cell data platform for scalable exploration, analysis and modeling of aggregated data. *Nucleic Acids Res.* **53**, D886–D900 (2025).

10. L. Tarhan, J. Bistline, J. Chang, B. Galloway, E. Hanna, E. Weitz, Single Cell Portal: an interactive home for single-cell genomics data, *bioRxivorg* (2023)p. 2023.07.13.548886.

11. A. Regev, S. A. Teichmann, E. S. Lander, I. Amit, C. Benoist, E. Birney, B. Bodenmiller, P. Campbell, P. Carninci, M. Clatworthy, H. Clevers, B. Deplancke, I. Dunham, J. Eberwine, R. Eils, W. Enard, A. Farmer, L. Fugger, B. Göttgens, N. Hacohen, M. Haniffa, M. Hemberg, S. Kim, P. Klenerman, A. Kriegstein, E. Lein, S. Linnarsson, E. Lundberg, J. Lundeberg, P. Majumder, J. C. Marioni, M. Merad, M. Mhlanga, M. Nawijn, M. Netea, G. Nolan, D. Pe'er, A. Phillipakis, C. P. Ponting, S. Quake, W. Reik, O. Rozenblatt-Rosen, J. Sanes, R. Satija, T. N. Schumacher, A. Shalek, E. Shapiro, P. Sharma, J. W. Shin, O. Stegle, M. Stratton, M. J. T. Stubbington, F. J. Theis, M. Uhlen, A. van Oudenaarden, A. Wagner, F. Watt, J. Weissman, B. Wold, R. Xavier, N. Yosef, Human Cell Atlas Meeting Participants, The Human Cell Atlas. *Elife* **6** (2017).

12. J. E. Rood, S. Wynne, L. Robson, A. Hupalowska, J. Randell, S. A. Teichmann, A. Regev, The Human Cell Atlas from a cell census to a unified foundation model. *Nature* **637**, 1065–1071 (2025).

13. F. B. Bastian, J. Roux, A. Niknejad, A. Comte, S. S. Fonseca Costa, T. M. de Farias, S. Moretti, G. Parmentier, V. R. de Laval, M. Rosikiewicz, J. Wollbrett, A. Echchiki, A. Escoriza, W. H. Gharib, M. Gonzales-Porta, Y. Jarosz, B. Laurenczy, P. Moret, E. Person, P. Roelli, K. Sanjeev, M. Seppey, M. Robinson-Rechavi, The Bgee suite: integrated curated expression atlas and comparative transcriptomics in animals. *Nucleic Acids Res.* **49**, D831–D847 (2021).

14. H. Cui, C. Wang, H. Maan, K. Pang, F. Luo, N. Duan, B. Wang, scGPT: toward building a foundation model for single-cell multi-omics using generative AI. *Nat. Methods* **21**, 1470–1480 (2024).

15. C. V. Theodoris, L. Xiao, A. Chopra, M. D. Chaffin, Z. R. Al Sayed, M. C. Hill, H. Mantineo, E. M. Brydon, Z. Zeng, X. S. Liu, P. T. Ellinor, Transfer learning enables predictions in network biology. *Nature* **618**, 616–624 (2023).

16. Y. Rosen, Y. Roohani, A. Agarwal, L. Samotorčan, Tabula Sapiens Consortium, S. R. Quake, J. Leskovec, Universal Cell Embeddings: A Foundation Model for Cell Biology, *bioRxiv* (2023)p. 2023.11.28.568918.

17. M. Hao, J. Gong, X. Zeng, C. Liu, Y. Guo, X. Cheng, T. Wang, J. Ma, X. Zhang, L. Song, Large-scale foundation model on single-cell transcriptomics. *Nat. Methods* **21**, 1481–1491 (2024).

18. X. Yang, G. Liu, G. Feng, D. Bu, P. Wang, J. Jiang, S. Chen, Q. Yang, Y. Zhang, Z. Man, Z. Liang, Z. Wang, Y. Li, Z. Li, Y. Liu, Y. Tian, A. Li, J. Dong, Z. Hu, C. Fang, H. Miao, L. Cui, Z. Deng, H. Jiang, W. Cui, J. Zhang, Z. Yang, H. Li, X. He, L. Zhong, J. Zhou, Z. Wang, Q. Long, P. Xu, The X-Compass Consortium, H. Wang, Z. Meng, X. Wang, Y. Wang, Y. Wang, S. Zhang, J. Guo, Y. Zhao, Y. Zhou, F. Li, J. Liu, Y. Chen, G. Yang, X. Li, GeneCompass: Deciphering Universal Gene Regulatory Mechanisms with Knowledge-Informed Cross-Species Foundation Model, *bioRxiv* (2023)p. 2023.09.26.559542.

19. H. Bian, Y. Chen, X. Dong, C. Li, M. Hao, S. Chen, J. Hu, M. Sun, L. Wei, X. Zhang, “ScMulan: A multitask generative pre-trained language model for single-cell analysis” in *Lecture Notes in Computer Science* (Springer Nature Switzerland, Cham, 2024) pp. 479–482.

20. X. Yang, G. Liu, G. Feng, D. Bu, P. Wang, J. Jiang, S. Chen, Q. Yang, H. Miao, Y. Zhang, Z. Man, Z. Liang, Z. Wang, Y. Li, Z. Li, Y. Liu, Y. Tian, W. Liu, C. Li, A. Li, J. Dong, Z. Hu, C. Fang, L. Cui, Z. Deng, H. Jiang, W. Cui, J. Zhang, Z. Yang, H. Li, X. He, L. Zhong, J. Zhou, Z. Wang, Q. Long, P. Xu, X-Compass Consortium, H. Wang, Z. Meng, X. Wang, Y. Wang, Y. Wang, S. Zhang, J. Guo, Y. Zhao, Y. Zhou, F. Li, J. Liu, Y. Chen, G. Yang, X. Li, GeneCompass: deciphering universal gene regulatory mechanisms with a knowledge-informed cross-species foundation model. *Cell Res.* **34**, 830–845 (2024).

21. A. C. Schaar, A. Tejada-Lapuerta, G. Palla, R. Gutgesell, L. Halle, M. Minaeva, L. Vornholz, L. Dony, F. Drummer, M. Bahrami, F. J. Theis, Nicheformer: a foundation model for single-cell and spatial omics, *bioRxiv* (2024)p. 2024.04.15.589472.

22. J. Kalfon, J. Samaran, G. Peyré, L. Cantini, scPRINT: pre-training on 50 million cells allows robust gene network predictions. *Nat. Commun.* **16**, 3607 (2025).

23. Y. Xu, C. Gatt, E. Kaymak, K. Kikuchi, T. Bourguignon, F. Zanini, Exploration and generation of cell transcriptomes over deep evolutionary time, *Evolutionary Biology* (2025). https://www.biorxiv.org/content/10.1101/2025.02.19.639005v2.

24. M. D. Luecken, M. Büttner, K. Chaichoompu, A. Danese, M. Interlandi, M. F. Mueller, D. C. Strobl, L. Zappia, M. Dugas, M. Colomé-Tatché, F. J. Theis, Benchmarking atlas-level data integration in single-cell genomics. *Nat. Methods* **19**, 41–50 (2022).

25. I. Korsunsky, N. Millard, J. Fan, K. Slowikowski, F. Zhang, K. Wei, Y. Baglaenko, M. Brenner, P.-R. Loh, S. Raychaudhuri, Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat.*

Methods **16**, 1289–1296 (2019).

26. C. Bunne, Y. Roohani, Y. Rosen, A. Gupta, X. Zhang, M. Roed, T. Alexandrov, M. AlQuraishi, P. Brennan, D. B. Burkhardt, A. Califano, J. Cool, A. F. Dernburg, K. Ewing, E. B. Fox, M. Haury, A. E. Herr, E. Horvitz, P. D. Hsu, V. Jain, G. R. Johnson, T. Kalil, D. R. Kelley, S. O. Kelley, A. Kreshuk, T. Mitchison, S. Otte, J. Shendure, N. J. Sofroniew, F. Theis, C. V. Theodoris, S. Upadhyayula, M. Valer, B. Wang, E. Xing, S. Yeung-Levy, M. Zitnik, T. Karaletsos, A. Regev, E. Lundberg, J. Leskovec, S. R. Quake, How to build the virtual cell with artificial intelligence: Priorities and opportunities, *ArXiv* (2024). https://www.ncbi.nlm.nih.gov/pubmed/39398201.

27. Z. Lin, H. Akin, R. Rao, B. Hie, Z. Zhu, W. Lu, N. Smetanin, R. Verkuil, O. Kabeli, Y. Shmueli, A. Dos Santos Costa, M. Fazel-Zarandi, T. Sercu, S. Candido, A. Rives, Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123–1130 (2023).

28. S. Kumar, G. Stecher, M. Suleski, S. B. Hedges, TimeTree: A resource for timelines, timetrees, and divergence times. *Mol. Biol. Evol.* **34**, 1812–1819 (2017).

29. F. Murat, N. Mbengue, S. B. Winge, T. Trefzer, E. Leushkin, M. Sepp, M. Cardoso-Moreira, J. Schmidt, C. Schneider, K. Mößinger, T. Brüning, F. Lamanna, M. R. Belles, C. Conrad, I. Kondova, R. Bontrop, R. Behr, P. Khaitovich, S. Pääbo, T. Marques-Bonet, F. Grützner, K. Almstrup, M. H. Schierup, H. Kaessmann, The molecular evolution of spermatogenesis across mammals. *Nature* **613**, 308–316 (2023).

30. S. R. Quake, The Tabula Sapiens Consortium, Tabula Sapiens reveals transcription factor expression, senescence effects, and sex-specific features in cell types from 28 human organs and tissues, *bioRxiv* (2024)p. 2024.12.03.626516.

31. T. T.-H. Wu, K. J. Travaglini, A. Rustagi, D. Xu, Y. Zhang, L. Andronov, S. Jang, A. Gillich, R. Dehghannasiri, G. J. Martínez-Colón, A. Beck, D. D. Liu, A. J. Wilk, M. Morri, W. L. Trope, R. Bierman, I. L. Weissman, J. B. Shrager, S. R. Quake, C. S. Kuo, J. Salzman, W. E. Moerner, P. S. Kim, C. A. Blish, M. A. Krasnow, Interstitial macrophages are a focus of viral takeover and inflammation in COVID-19 initiation in human lung. *J. Exp. Med.* **221** (2024).

32. D. Szklarczyk, R. Kirsch, M. Koutrouli, K. Nastou, F. Mehryary, R. Hachilif, A. L. Gable, T. Fang, N. T. Doncheva, S. Pyysalo, P. Bork, L. J. Jensen, C. von Mering, The STRING database in 2023: protein-protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic Acids Res.* **51**, D638–D646 (2023).

33. J. Christensen, P. Cloos, U. Toftegaard, D. Klinkenberg, A. P. Bracken, E. Trinh, M. Heeran, L. Di Stefano, K. Helin, Characterization of E2F8, a novel E2F-like cell-cycle regulated repressor of E2F-activated transcription. *Nucleic Acids Res.* **33**, 5458–5470 (2005).

34. K. Yoshida, I. Inoue, Regulation of Geminin and Cdt1 expression by E2F transcription factors. *Oncogene* **23**, 3802–3812 (2004).

35. J. Laoukili, M. R. H. Kooistra, A. Brás, J. Kauw, R. M. Kerkhoven, A. Morrison, H. Clevers, R. H. Medema, FoxM1 is required for execution of the mitotic programme and chromosome stability. *Nat. Cell Biol.* **7**, 126–136 (2005).

36. Y. Wang, H. Wang, Z. Yan, G. Li, G. Hu, H. Zhang, D. Huang, Y. Wang, X. Zhang, Y. Yan, Q. Lu, M. Cheng, S. Luo, The critical role of dysregulated Hh-FOXM1-TPX2 signaling in human hepatocellular carcinoma cell proliferation. *Cell Commun. Signal.* **18**, 116 (2020).

37. Q. Li, M. Sun, Y. Meng, M. Feng, M. Wang, C. Chang, H. Dong, F. Bu, C. Xu, J. Liu, Q. Ling, Y. Qiao, J. Chen, Kinesin family member 18B activates mTORC1 signaling via actin gamma 1 to promote the recurrence of human hepatocellular carcinoma. *Oncogenesis* **12**, 54 (2023).

38. Y. Tong, T. Eigler, Transcriptional targets for pituitary tumor-transforming gene-1. *J. Mol. Endocrinol.*

43, 179–185 (2009).

39. L. A. Garrett-Sinha, G. H. Su, S. Rao, S. Kabak, Z. Hao, M. R. Clark, M. C. Simon, PU.1 and Spi-B are required for normal B cell receptor-mediated signal transduction. *Immunity* **10**, 399–408 (1999).

40. S. Horiuchi, T. Koike, H. Takebuchi, K. Hoshino, I. Sasaki, Y. Fukuda-Ohta, T. Kaisho, D. Kitamura, SpiB regulates the expression of B-cell-related genes and increases the longevity of memory B cells. *Front. Immunol.* **14**, 1250719 (2023).

41. S. Sadasivam, S. Duan, J. A. DeCaprio, The MuvB complex sequentially recruits B-Myb and FoxM1 to promote mitotic gene expression. *Genes Dev.* **26**, 474–489 (2012).

42. A. L. Tien, S. Senbanerjee, A. Kulkarni, R. Mudbhary, B. Goudreau, S. Ganesan, K. C. Sadler, C. Ukomadu, UHRF1 depletion causes a G2/M arrest, activation of DNA damage response and apoptosis. *Biochem. J.* **435**, 175–185 (2011).

43. B. P. Madakashira, E. Magnani, S. Ranjan, K. C. Sadler, DNA hypomethylation activates Cdk4/6 and Atr to induce DNA replication and cell cycle arrest to constrain liver outgrowth in zebrafish. *Nucleic Acids Res.* **52**, 3069–3087 (2024).

1 Methods

1.1 Data

We curated an extensive collection of single-cell RNA sequencing datasets for model pretraining and evaluation from public sources comprising 115 million cells from 465 tissues, 1,865 cell types in 129 disease states across 23 species.

1.1.1 Pretraining Data Selection and Curation

For pretraining our largest model TranscriptFormer-Metazoa, we selected datasets that span the major animal lineages. This diverse dataset includes six vertebrate species (*Homo sapiens*, *Mus musculus*, *Oryctolagus cuniculus*, *Gallus gallus*, *Xenopus laevis*, *Danio rerio*), four invertebrate species (*Lytechinus variegatus*, *Caenorhabditis elegans*, *Drosophila melanogaster*, *Spongilla lacustris*), and two outgroups: a fungus (*Saccharomyces cerevisiae*) and a protist (*Plasmodium falciparum*). This phylogenetically diverse collection spans approximately 1.53 billion years of evolutionary history ([1]; Figure 1A).

Most of these data were sourced from CZ CELLxGENE using the Discover API [2], comprising 57 million human cells (644 datasets) and 27 million mouse cells (88 datasets) (Extended Data Table 1) We filtered these datasets to exclude spatial assays and duplicate cells. Both normal (67 millions cells) and diseased (17 million cells across 123 diseases) cells were retained.

To capture as many cell types and states as possible from other species, we selected whole-organism, developmental, and aging cell atlases for the remaining 16 datasets (Extended Data Table 2). These were obtained from various public repositories and processed into a standardized format.

Processing We converted each dataset to AnnData objects, verified the presence of raw gene count layers, harmonized metadata fields (`organism`, `assay`, `suspension_type`), and mapped and updated gene features to Ensembl gene stable IDs (v113) for each species. Two datasets (*Spongilla lacustris*[3] and Zebrahub[4]) required reprocessing from raw sequencing reads, which we accomplished using Cell Ranger v9.1 with default settings and custom genome references (odSpoLacu1.1 and GRCz11, respectively). To ensure data quality, we applied minimal filtering to remove cells with fewer than 200 non-zero in-vocabulary genes, preserving the biological heterogeneity of the data.

1.1.2 Evaluation Data Selection and Curation

We assembled a comprehensive set of evaluation datasets to benchmark TranscriptFormer against existing methods.

Tabula Sapiens 2.0 A human scRNA-seq reference cell atlas [5]. We downloaded single-cell RNA sequencing data from the Tabula Sapiens Consortium via CZ CELLxGENE. To ensure independence from our training data, we filtered the "All Cells" dataset to include nine donor IDs absent from Tabula Sapiens 1.0 (TSP17 to TSP30). We excluded pancreas data due to insufficient cell counts (>100 cells).

Spermatogenesis dataset A multi-species snRNA-seq dataset of testes from the major lineages of mammals and birds [6]. Data from nine species (*H. sapiens*, *Gorilla gorilla*, *G. gallus*, *Callithrix jacchus*, *Macaca mulatta*, *Monodelphis domestica*, *M. musculus*, *Ornithorhynchus anatinus*, *Pan troglodytes*) were downloaded from two sources (GEO and Bgee). Data were aligned and merged to create a dataset with raw counts (GEO) and harmonized metadata (Bgee).

Tabula Microcebus A scRNA-seq reference cell atlas for mouse lemur (*Microcebus murinus*), a model primate species [7]. We obtained data from the Tabula Microcebus Consortium. We filtered out cells marked as low quality, contaminated, or lacking cell type assignments according to the original authors' criteria (179,519 cells).

Tropical clawed frog cell atlas A scRNA-seq developmental cell atlas for *X. tropicalis* covering embryonic stages [8]. The full embryonic dataset was subset to the early tailbud stage when dozens of cell types have differentiated (34,335 cells; 54 cell types).

Zebrafish cell landscape A scRNA-seq developmental cell atlas for *D. rerio* [9]. The dataset was subset to adult stage and split by tissue (N=10) (184,808 cells; 91 cell types). All embryonic stages were pooled.

Stony coral cell atlas A scRNA-seq reference cell atlas for stony coral (*Stylophora pistillata*) [10]. The dataset was subset to include adult stage only (15,252 cells; 13 cell types).

Sea lamprey brain A snRNA-seq dataset of brain tissue from *Petromyzon marinus* [11]. This dataset required reprocessing of the raw sequencing reads. We used Cell Ranger v9.1 with default settings and a custom genome reference (Pmarinus7.0).

COVID-19 human lung The scRNA-seq experimental dataset of four healthy donors' lung sections infected with SARS-CoV-2 [12]. Data were downloaded from CZ CELLxGENE.

Processing We converted all datasets to AnnData objects, harmonized metadata fields (`organism`, `cell type`), and mapped gene features to Ensembl gene stable IDs (v113). For quality control, we removed duplicate cell barcodes, and cells with zero gene counts.

For all cell type prediction tasks, we applied consistent curation steps across evaluation datasets. We removed cells with ambiguous cell type annotations (labeled as "cell," "unassigned," "unknown," or "Unclassified") and filtered out genes with zero counts across all cells. We also standardized all count matrices to be raw counts, ensuring consistency across all datasets for our analyses.

1.2 Modeling Cells as Bags of Transcripts

Cells can be represented as *bags of transcripts*, where each transcript t_k corresponds to an observed mRNA molecule. These transcripts are unordered, and identical transcripts are indistinguishable from one another, meaning that the representation of a cell is based on the *multiset* of observed transcripts rather than their order.

Mathematically, let a cell be represented as a collection of transcripts:

$$\mathcal{T} = \{\underbrace{t_1, t_1, \ldots, t_1}_{c_1 \text{ times}}, \underbrace{t_2, t_2, \ldots, t_2}_{c_2 \text{ times}}, \ldots, \underbrace{t_K, t_K, \ldots, t_K}_{c_K \text{ times}}\}, \tag{1}$$

where K is the number of unique transcripts observed. Because transcripts are sampled from expressed genes, the same transcript can appear multiple times within a cell. By aggregating transcripts at the gene level, we define a *gene expression profile* instead as a collection of genes, where each gene g_j is associated with a *count* c_j, denoting the number of times a transcript from that gene was observed.

Formally, the gene expression profile representation of a cell is:

$$\mathcal{G} = \{(g_1, c_1), (g_2, c_2), \ldots, (g_M, c_M)\}, \tag{2}$$

where:

- g_j is a unique gene observed in the cell,
- c_j is its associated *expression level*, defined as the number of times a transcript from gene g_j appears,
- M is the number of unique genes detected in the cell (non-zero counts).

This transformation from transcripts to gene-level counts allows for a more efficient representation of single-cell data, capturing the overall expression profile rather than individual transcript occurrences. The next section describes a generative model that formalizes this process, modeling the sequential selection of genes and their associated expression levels.

1.3 A Generative Model for Gene Expression

The generative model for gene expression represents the process by which a cell's transcriptomic profile is formed. Given a set of genes and their associated expression levels, we model the generation of a cell's expression profile as a sequential process involving two key steps: (1) selecting a gene from a categorical distribution and (2) sampling its expression count from a zero-truncated Poisson (ZTP) distribution.

Step 1: Gene Selection At each step j, a gene g_j is sampled from a categorical distribution over the gene vocabulary. The probability of selecting a gene depends on previously chosen genes and their counts. Formally, the probability of selecting g_j at step j is given by:

$$P(g_j \mid g_{<j}, c_{<j}) = \text{Categorical}(\pi_j), \tag{3}$$

where $\pi_j = f_\theta(g_{<j}, c_{<j})$ is a probability distribution over the gene vocabulary, parameterized by the function f_θ with learned parameters θ.

Step 2: Count Sampling Once a gene g_j is selected at step j, its expression level c_j is sampled from a zero-truncated Poisson (ZTP) distribution. The ZTP distribution ensures that every selected gene has a strictly positive count:

$$P(c_j \mid g_j, g_{<j}, c_{<j}) = \text{ZTP}(\lambda_j), \tag{4}$$

where $\lambda_j = f_\phi(g_j, g_{<j}, c_{<j})$ is the Poisson rate parameter, estimated by the function f_ϕ with learned parameters ϕ.

Joint Distribution over Genes and Counts The generative process described above defines a structured distribution over both the genes present in a cell and their corresponding expression levels. Given the sequential nature of the model, the joint probability distribution over all genes and counts in a cell is factorized as:

$$P(\mathcal{G}) = P(g_0)P(c_0|g_0)\prod_{j=1}^{M} P(g_j \mid g_{<j}, c_{<j})P(c_j \mid g_j, g_{<j}, c_{<j}), \tag{5}$$

Since the categorical distribution for gene selection and the ZTP distribution for count sampling are both conditioned on the previously sampled genes and counts, this formulation captures dependencies between genes in a cell's expression profile. The parameters θ and ϕ are jointly learned, allowing the model to infer gene co-expression patterns and context-dependent count distributions.

This generative framework provides a structured probabilistic model for single-cell transcriptomics, enabling explicit modeling of gene-gene interactions and expression dependencies within

a cell capturing the structured nature of transcriptomic data.

1.4 Model Architecture

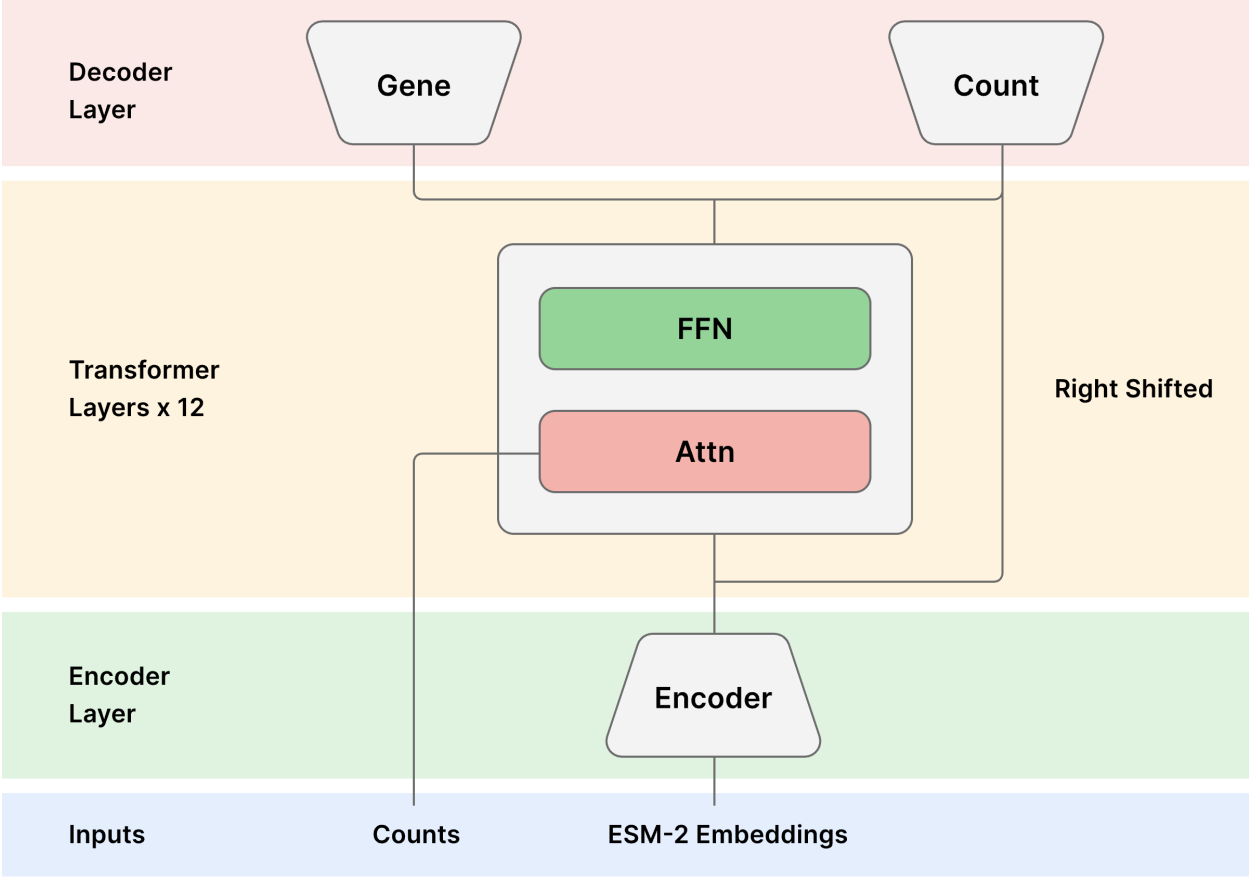

Figure 1: Model architecture diagram for TranscriptFormer models.

Input Embeddings TranscriptFormer embeds gene sequences across diverse species and sequencing technology into continuous representations. For each gene g_j, we map it to its associated amino acid and proteins sequences. These sequences are embedded using ESM-2 [13], a protein language model. If there are multiple proteins associated with a gene then the mean over all ESM-2 protein embeddings is used.

Additionally, we introduce an *assay token* $\mathbf{a} \in \mathbb{R}^d$ that encodes information about the sequencing technology used. The transformer input, $Z^{(0)} \in \mathbb{R}^{(M+1)\times d}$, combines the gene embeddings (processed through an MLP) with the assay token:

$$Z^{(0)} = \text{concat}(\mathbf{a}, \text{MLP}(E)) \tag{6}$$

where $E \in \mathbb{R}^{M \times d_{esm2}}$ is the protein embeddings tensor. The MLP serves to project the ESM-2 embedding dimension d_{esm2} to the model dimension d.

Transformer Architecture The model follows a standard transformer encoder architecture with L stacked layers. The output of the l-th layer is defined recursively as:

$$Z^{(l)} = \mathrm{TransformerLayer}(Z^{(l-1)}) \tag{7}$$

where $Z^{(l-1)} \in \mathbb{R}^{(M+1) \times d}$ is the output from the previous layer. We employ pre-normalization, GeLU activation as well as removal of MLP and attention bias terms. We omit positional encodings from the input since the gene expression profiles have no inherit sequence.

Expression-Aware Multi-Head Self-Attention We encode count information as a bias term in the attention matrix at each transformer layer. We modified the attention mechanism to incorporate the counts as follows:

$$\text{Attention}(Q, K, V, C) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}} + \log(1 + C)\right) V \tag{8}$$

where the count matrix $C \in \mathbb{R}^{(M+1) \times (M+1)}$ is constructed by repeating the count vector $\mathbf{c} = (1, c_1, c_2, \ldots, c_M)$ across all rows. Q, K and V are the standard query, key and value matrices respectively [14]. In Supplementary Methods we show that this bias term ensures that higher-count genes exert greater influence in the attention computation in a way that is consistent with the bag-of-transcripts representation of cells. For numerical stability, we apply soft capping to the softmax logits. The standard causal attention mask is applied across gene tokens, while the assay token is never masked.

Decoders Heads The gene decoder head predicts the gene Categorical distribution parameters π_j at each position j using a two-layer MLP followed by softmax:

$$\pi_j = \text{softmax}(\text{MLP}_\theta(\mathbf{z}_j^{(L)})) \tag{9}$$

where $\mathbf{z}_j^{(L)} \in \mathbb{R}^d$ is the jth component of $Z^{(L)}$.

The input to the count decoder is obtained by concatenating the final hidden states with the transformer input for the next position $j + 1$, which is then passed through a MLP to get the pre-normalized count logits $\mathbf{u} \in \mathbb{R}^M$ as so

$$u_j = \text{MLP}_\phi(\text{concat}(\mathbf{z}_j^{(L)}, \mathbf{z}_{j+1}^{(0)})). \tag{10}$$

To improve training stability, a softmax normalization is applied to $\mathbf{u}$ (across the sequence dimension) and then rescaled to ensure that the predicted counts sum to the observed total $N = \sum c_j$.

Additionally a learnable gene-level bias $\mathbf{b} \in \mathbb{R}^{|V|}$ is applied to all genes in the input sequence to obtain the Poisson rate parameter vector:

$$\lambda = N \cdot \text{softmax}(\mathbf{u} + \mathbf{b}_{g_0:g_M}). \tag{11}$$

Cell and Contextual Gene Embeddings TranscriptFormer produces two types of embeddings: cell embeddings and contextual gene embeddings (CGEs). Cell embeddings represent the entire expression profile of a cell as a single vector, computed as the average over all transformer outputs (excluding the assay token) from the final layer:

$$\mathbf{z}_{\text{cell}} = \frac{1}{M} \sum_{j=1}^{M} \mathbf{z}_j^{(L)} \tag{12}$$

where $\mathbf{z}_j^{(L)}$ denotes the output of the L-th transformer layer for the j-th gene position. Contextual gene embeddings (CGEs), on the other hand, capture gene-specific representations within their cellular context. Each CGE corresponds directly to an individual component $\mathbf{z}_j^{(L)}$ of the transformer's final layer output.

1.4.1 Loss Functions

The model is trained using two distinct loss functions corresponding to the two decoder heads: a cross-entropy loss for gene prediction and a zero-truncated Poisson negative log-likelihood (NLL) loss for count prediction.

Gene Decoder Loss The gene decoder predicts a probability distribution over the gene vocabulary at each sequence position. The standard cross-entropy loss is used to supervise the gene predictions. Given the predicted gene probabilities π_j at position j the gene loss is given by:

$$\mathcal{L}_{\text{gene}} = -\sum_{j=1}^{M} \log \pi_j. \tag{13}$$

This loss encourages the model to assign high probability to the correct gene identity at each position.

Count Decoder Loss The count decoder predicts a set of expected expression counts per sequence position. Since gene expression counts follow a discrete, non-negative distribution, we model them using a Poisson likelihood. However, because counts of zero are excluded from the input sequence, we employ a zero-truncated Poisson negative log-likelihood (NLL) loss.

Given the predicted rate parameter λ_j for position j and the observed ground truth count c_j,

the loss is computed as:

$$\mathcal{L}_{\text{count}} = -\sum_{j=1}^{M} c_j \log \lambda_j - \lambda_j - \log(1 - e^{-\lambda_j}). \tag{14}$$

The additional term $\log(1 - e^{-\lambda_j})$ accounts for the truncation of zero values, ensuring proper normalization.

Total Loss The final loss function is the sum over gene and count losses:

$$\mathcal{L} = \mathcal{L}_{\text{gene}} + \mathcal{L}_{\text{count}}. \tag{15}$$

This formulation ensures that the model jointly learns to predict gene identities and their corresponding expression levels in a biologically meaningful manner.

1.5 Model Configurations and Versions

We trained three variants of the TranscriptFormer architecture, each using a different subset of the available dataset, but all sharing an identical core model configuration. Each model is a 12-layer Transformer with 16 attention heads per layer, using a model dimension of 2048 and GELU activation. The architecture employs full-sequence self-attention with a maximum sequence length of 2,047 tokens (representing genes), no forward or attention biases, and a dropout rate of 0.1 applied throughout. The gene embeddings are initialized from pretrained representations and held fixed during training. This consistent architecture results in a Transformer module with approximately 302 million parameters, with overall trainable parameters varying across models primarily due to differences in vocabulary size and decoder head dimensionality.

To handle the exceptionally large output vocabulary in the TF-Metazoa model (over 247,000 genes across 12 species), we reduced the output embedding dimensionality from 2048 to 512, which yielded no loss in performance during preliminary evaluations.

- **TF-Metazoa**: Trained on 112 million cells spanning all twelve species described in Section 1.1. The model includes 444 million trainable parameters and 633 million non-trainable parameters (from frozen pretrained embeddings). Vocabulary size: 247,388.

- **TF-Exemplar**: Trained on 110 million cells from five model organisms: human (*H. sapiens*), mouse (*M. musculus*), zebrafish (*D. rerio*), fruit fly (*D. melanogaster*), and *C. elegans*. Total trainable parameters: 542 million; non-trainable: 282 million. Vocabulary size: 110,290.

- **TF-Sapiens**: Trained on 57 million human-only cells. This model has 368 million trainable parameters and 61 million non-trainable parameters. Vocabulary size: 23,829.

1.6 Model Training

Training Setup Training was conducted on CZI's compute cluster equipped with 1000 H100 GPUs, using 16-bit mixed precision for improved efficiency and memory utilization. The training loop was implemented using PyTorch Lightning with Distributed Data Parallel (DDP) backend. We used PyTorch's Flex Attention [15] to build the transformer blocks. A global batch size of approximately 4–5 million tokens was used, achieved with a per-GPU batch size of 4–16 and gradient accumulation steps of 6. Gradients were clipped to a maximum norm of 1.0 to improve stability.

Training Data Sampling During training, expression profiles were sampled from each species s with sampling probability p_s, where the sampling probability is give by:

$$p_s = \frac{n_s^\alpha}{\sum_{s'} n_{s'}^\alpha} \tag{16}$$

where n_s is the number of training cells from species s, and $\alpha = 0.6$ controls the sampling skew. This strategy increases the relative sampling frequency of low-resource species while preventing overfitting to high-resource species like human or mouse.

Sequence Processing We trained with a sequence length of 2048, which includes the gene expression profiles and assay token. Genes were randomly shuffled in each batch to enforce permutation invariance. Sequences longer than 2048 were truncated, and shorter sequences were padded using a special `[PAD]` token. The model was trained using causal masking over this gene vocabulary, with raw count values clipped to a maximum of 30. Cell with less than 200 expressed gene were filtered out of training. Additionally, outlier cells with expression levels exceeding 3 standard deviations above or below the dataset average were filtered.

Optimization We used the AdamW [16] optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.95$, and weight decay set to 0.1. The learning rate was scheduled using a linear warm-up over the first 10% of training steps, followed by cosine decay to the minimum learning rate. The peak learning rate was set to 1.5×10^{-4}, decaying to a minimum of 1.0×10^{-5}. Learning rates were scaled relative to a baseline batch size of 1,536 tokens to maintain consistency across model variants. Each model was trained for a total of approximately 3.5 trillion tokens, equivalent to 15 epochs for the TF-Metazoa model.

1.7 Analysis

1.7.1 Cross-Species Reference Mapping via kNN Classification

To assess the ability of TranscriptFormer to generalize transcriptional representations across species, we performed a cross-species reference mapping experiment using a curated single-nucleus RNA-seq dataset of spermatogenesis from nine vertebrate species [6].

We adopted a transfer learning framework in which a k-nearest neighbors (kNN) classifier was trained on labeled cell embeddings from one species (source) and evaluated on embeddings from a different species (target). For each source-target species pair, we embedded all cells using a frozen TranscriptFormer model and extracted their representation from the final encoder layer. A kNN classifier ($k = 10$) was trained on the source species' cell embeddings with known cell type labels, and then used to predict labels for the target species' cells based on nearest-neighbor majority voting in embedding space.

Classification performance was quantified using the F1 score, calculated per cell type and averaged to obtain a macro F1 score (see Supplementary Methods) for each source–target species pair. This reference mapping approach evaluates the model's ability to transfer cell type identity across evolutionary distances without fine-tuning.

1.7.2 COVID-19 disease classification task

For the disease classification task, we downloaded the dataset from CELLxGENE from the corresponding publication Wu et al. [12], see. We used the original labels of "infected" cells (True or False) provided by the original authors. We adopted a linear probing evaluation as discussed in Supplementary Methods. We generated embeddings for each TranscriptFormer model variant as well as the comparison models. We ran 5-fold cross-validation for each, and reported macro F1 score in Figure 3C,D. We followed the performance averaging and uncertainty calculation as described in Supplementary Methods.

1.7.3 Contextualized Gene Embedding

To systematically evaluate contextualized gene embeddings (CGEs), we developed several types of CGE analyses. These included cross-donor, cross-cell type, and cross-tissue comparisons, as well as a flexible variance partitioning framework for quantifying the influence of different biological covariates.

Gene Selection For each analysis type, raw count matrices from Tabula Sapiens 2.0 datasets were subset to the relevant cell types (typically those containing large numbers of cells) and preprocessed using Scanpy [17]. Standard preprocessing steps were applied, which included gene filtering (minimum of 3 cells), library-size normalization to 10,000 counts per cell, and log-transformation. Highly variable genes (HVGs) were identified using the Seurat method [18]. In

cross-tissue and cross-donor analyses, genes were retained if they were expressed in at least 10% of cells within the target cell type. For cross-cell type comparisons, genes were required to exceed the same threshold in at least three distinct cell types. This was done to increase the probability that the selected genes would have a corresponding CGE in a large number of cells, as the model selects only the top 2047 expressed genes per cell to form each cell sentence. A fixed number of valid genes ($n = 20$) were randomly selected per analysis, and those without a corresponding CGE in any cells were excluded from the analysis. We created an auxiliary function for retrieving additional metadata (gene functional annotation) for the selected genes based on their Ensemble IDs.

Embedding Aggregation and Dimensionality Reduction CGEs were obtained from inference runs through the TranscriptFormer model. Cells were grouped by donor (cross-donor), tissue (cross-tissue), or cell type (cross-cell type), and gene embeddings were averaged within each group. This process also recorded the number of cells contributing to each averaged CGE.

The averaged CGEs were projected into two dimensions using principal component analysis (PCA). PCA results were visualized in multi-panel scatter plots: (1) colored by the biological grouping variable (e.g., donor, tissue, or cell type), (2) colored by gene identity, and (3) optionally colored by the $\log_2$-transformed number of contributing cells. These plots allowed direct visual comparison of embedding structure across biological axes.

Variance Partitioning Analysis and Gene Selection Tests To assess the impact of gene selection criteria on the observed CGE trends, we implemented a flexible analysis that supports parameterization of gene filtering based on HVG status, expression thresholds, and the minimum number of cell types in which a gene must be expressed. For each configuration, CGEs were aggregated across donor, tissue, and cell type, followed by PCA.

We evaluated four gene selection criteria (see Supplementary Methods for additional details on the selected criteria):

- **Baseline** HVG filtering enabled, expression threshold $= 10\%$, min cell types $= 3$
- **Test 1** HVG filtering disabled, expression threshold $= 10\%$, min cell types $= 3$
- **Test 2** HVG filtering disabled, expression threshold $= 1\%$, min cell types $= 3$
- **Test 3** HVG filtering disabled, expression threshold $= 1\%$, min cell types $= 1$

Each configuration was run using the same number of randomly selected genes. For each result, variance partitioning [19] was performed by fitting linear models to PCA coordinates and computing partial R^2 values for donor, tissue, and cell type. The results are shown in SI Fig. 7.

1.7.4 Pairwise Gene-Gene Interaction Analysis

To assess whether TranscriptFormer captures biologically meaningful regulatory relationships between transcription factors (TFs) and their target genes, we computed pointwise mutual information (PMI) [20] values for each TF–gene pair, conditioned on a specified assay token a:

$$\mathrm{PMI}(g_i, TF_j|a) = \log\left(\frac{P(g_i, TF_j|a)}{P(g_i|a)P(TF_j|a)}\right) = \log\left(\frac{P(g_i|TF_j, a)}{P(g_i|a)}\right) \tag{17}$$

Marginal probabilities $P(g_i|a)$ were obtained via single-step decoding from TF-Sapiens by "prompting" on the assay token, while conditional probabilities $P(g_i|TF_j, a)$ were derived from two-step decoding, i.e. prompting with TF_j as the initial token ($c_j = 1$). To reduce spurious effects from low-probability genes, we filtered out genes with marginal probabilities below 10^{-8}, retaining 17,625 protein-coding genes for analysis.

To assess statistical significance of interactions, we computed Z-scores for each PMI value:

$$Z_{ij}^{\mathrm{PMI}} = \frac{\mathrm{PMI}(g_i, TF_j|a) - \mu_j^{\mathrm{PMI}}}{\sigma_j^{\mathrm{PMI}}}$$

where μ_j^{PMI} and σ_j^{PMI} denote the mean and standard deviation, respectively, of the PMI distribution for transcription factor TF_j, estimated across all considered genes. This per-TF background modeling accounts for variability in interaction distributions between different transcription factors. Two-sided p-values were then calculated using the standard normal distribution as:

$$p_{ij} = 2 \cdot \Phi(-|Z_{ij}^{\mathrm{PMI}}|)$$

where Φ is the standard normal cumulative distribution function. Resulting p-values were adjusted for multiple hypothesis testing using the Benjamini–Hochberg procedure [21] to control the false discovery rate (FDR). We applied an FDR threshold of $\alpha = 0.05$ to identify statistically significant TF–target gene interactions inferred by the model.

1.7.5 Cell Type-Specific Transcription Factor Generation

To assess the ability of TranscriptFormer to capture cell type-specific transcription factor (TF) activity, we implemented a conditional generation strategy inspired by the Tabula Sapiens 2.0 (TS2) study [5]. We began by identifying the top 100 marker genes for each of 112 cell types spanning 28 tissues in the TS2 dataset. Marker genes $\mathcal{G}_{\mathrm{marker}}$ were selected based on average expression profiles within each cell type, resulting in a high-confidence gene set for conditioning the model.

Using these cell type-specific gene signatures, we probed the TF-Sapiens model by prompting it with the set $\mathcal{G}_{\mathrm{marker}}$ alongside an assay token a, and performed a single forward decoding step to obtain the conditional log-probability distribution over all 1,635 human transcription factors in

the TS2 study:

$$\log P(TF_j \mid \mathcal{G}_{\text{marker}}, a)$$

which we interpret as pseudo-expression values for transcription factor TF_j in the context of the marker gene set. This probabilistic output reflects the model's internal representation of TF–cell type associations, even though it was not explicitly trained on the TS2 dataset.

To identify dominant TFs for each cell type, we grouped transcription factors by their maximal pseudo-expression value across the 112 cell types. That is, for each transcription factor TF_j, we identified the most associated cell type c^* as:

$$c_j^* = \arg\max_{c \in \mathcal{C}} \log P(TF_j \mid \mathcal{G}_{\text{marker}}^{(c)}, a)$$

where $\mathcal{C}$ is the set of all 112 cell types, and $\mathcal{G}_{\text{marker}}^{(c)}$ denotes the marker gene set for cell type c. This allowed us to map transcription factors to the most likely associated cell type according to the model's learned internal representations.

The resulting matrix of pseudo-expression values across all TF–cell type combinations was visualized as a heatmap (Fig. 5B). Transcription factors were arranged according to their assigned cell type, revealing distinctive structural features: vertical bands corresponding to ubiquitously expressed TFs and diagonal patterns highlighting cell-specific regulatory factors.

References

[1] Sudhir Kumar, Glen Stecher, Michael Suleski, and S Blair Hedges. TimeTree: A resource for timelines, timetrees, and divergence times. *Mol. Biol. Evol.*, 34:1812–1819, 2017.

[2] CZI Cell Science Program, Shibla Abdulla, Brian Aevermann, Pedro Assis, Seve Badajoz, Sidney M Bell, Emanuele Bezzi, Batuhan Cakir, Jim Chaffer, Signe Chambers, J Michael Cherry, Tiffany Chi, Jennifer Chien, Leah Dorman, Pablo Garcia-Nieto, Nayib Gloria, Mim Hastie, Daniel Hegeman, Jason Hilton, Timmy Huang, Amanda Infeld, Ana-Maria Istrate, Ivana Jelic, Kuni Katsuya, Yang Joon Kim, Karen Liang, Mike Lin, Maximilian Lombardo, Bailey Marshall, Bruce Martin, Fran McDade, Colin Megill, Nikhil Patel, Alexander Predeus, Brian Raymor, Behnam Robatmili, Dave Rogers, Erica Rutherford, Dana Sadgat, Andrew Shin, Corinn Small, Trent Smith, Prathap Sridharan, Alexander Tarashansky, Norbert Tavares, Harley Thomas, Andrew Tolopko, Meghan Urisko, Joyce Yan, Garabet Yeretssian, Jennifer Zamanian, Arathi Mani, Jonah Cool, and Ambrose Carr. CZ CELLxGENE discover: a single-cell data platform for scalable exploration, analysis and modeling of aggregated data. *Nucleic Acids Res.*, 53:D886–D900, 2025.

[3] Jacob M Musser, Klaske J Schippers, Michael Nickel, Giulia Mizzon, Andrea B Kohn, Constantin Pape, Paolo Ronchi, Nikolaos Papadopoulos, Alexander J Tarashansky, Jörg U Ham-

mel, Florian Wolf, Cong Liang, Ana Hernández-Plaza, Carlos P Cantalapiedra, Kaia Achim, Nicole L Schieber, Leslie Pan, Fabian Ruperti, Warren R Francis, Sergio Vargas, Svenja Kling, Maike Renkert, Maxim Polikarpov, Gleb Bourenkov, Roberto Feuda, Imre Gaspar, Pawel Burkhardt, Bo Wang, Peer Bork, Martin Beck, Thomas R Schneider, Anna Kreshuk, Gert Wörheide, Jaime Huerta-Cepas, Yannick Schwab, Leonid L Moroz, and Detlev Arendt. Profiling cellular diversity in sponges informs animal cell type and nervous system evolution. *Science*, 374:717–723, 2021.

[4] Merlin Lange, Alejandro Granados, Shruthi VijayKumar, Jordão Bragantini, Sarah Ancheta, Sreejith Santhosh, Michael Borja, Hirofumi Kobayashi, Erin McGeever, Ahmet Can Solak, Bin Yang, Xiang Zhao, Yang Liu, Angela M Detweiler, Sheryl Paul, Honey Mekonen, Tiger Lao, Rachel Banks, Yang-Joon Kim, Adrian Jacobo, Keir Balla, Kyle Awayan, Samuel D'Souza, Robert Haase, Alexandre Dizeux, Olivier Pourquie, Rafael Gómez-Sjöberg, Greg Huber, Mattia Serra, Norma Neff, Angela Oliveira Pisco, and Loïc A Royer. Zebrahub – multimodal zebrafish developmental atlas reveals the state-transition dynamics of late-vertebrate pluripotent axial progenitors. *bioRxiv*, page 2023.03.06.531398, 2023.

[5] Stephen R Quake and The Tabula Sapiens Consortium. Tabula sapiens reveals transcription factor expression, senescence effects, and sex-specific features in cell types from 28 human organs and tissues. *bioRxiv*, page 2024.12.03.626516, 2024.

[6] Florent Murat, Noe Mbengue, Sofia Boeg Winge, Timo Trefzer, Evgeny Leushkin, Mari Sepp, Margarida Cardoso-Moreira, Julia Schmidt, Celine Schneider, Katharina Mößinger, Thoomke Brüning, Francesco Lamanna, Meritxell Riera Belles, Christian Conrad, Ivanela Kondova, Ronald Bontrop, Rüdiger Behr, Philipp Khaitovich, Svante Pääbo, Tomas Marques-Bonet, Frank Grützner, Kristian Almstrup, Mikkel Heide Schierup, and Henrik Kaessmann. The molecular evolution of spermatogenesis across mammals. *Nature*, 613:308–316, 2023.

[7] The Tabula Microcebus Consortium, Camille Ezran, Shixuan Liu, Stephen Chang, Jingsi Ming, Olga Botvinnik, Lolita Penland, Alexander Tarashansky, Antoine de Morree, Kyle J Travaglini, Jia Zhao, Gefei Wang, Kazuteru Hasegawa, Hosu Sin, Rene Sit, Jennifer Okamoto, Rahul Sinha, Yue Zhang, Caitlin J Karanewsky, Jozeph L Pendleton, Maurizio Morri, Martine Perret, Fabienne Aujard, Lubert Stryer, Steven Artandi, Margaret Fuller, Irving L Weissman, Thomas A Rando, James E Ferrell, Jr, Bo Wang, Iwijn De Vlaminck, Can Yang, Kerriann M Casey, Megan A Albertelli, Angela Oliveira Pisco, Jim Karkanias, Norma Neff, Angela Ruohao Wu, Stephen R Quake, and Mark A Krasnow. Tabula microcebus: A transcriptomic cell atlas of mouse lemur, an emerging primate model organism. *bioRxiv*, page 2021.12.12.469460, 2021.

[8] James A Briggs, Caleb Weinreb, Daniel E Wagner, Sean Megason, Leonid Peshkin, Marc W Kirschner, and Allon M Klein. The dynamics of gene expression in vertebrate embryogenesis at single-cell resolution. *Science*, 360, 2018.

[9] Mengmeng Jiang, Yanyu Xiao, E Weigao, Lifeng Ma, Jingjing Wang, Haide Chen, Ce Gao, Yuan Liao, Qile Guo, Jinrong Peng, Xiaoping Han, and Guoji Guo. Characterization of the zebrafish cell landscape at single-cell resolution. *Front. Cell Dev. Biol.*, 9, 2021.

[10] Shani Levy, Anamaria Elek, Xavier Grau-Bové, Simón Menéndez-Bravo, Marta Iglesias, Amos Tanay, Tali Mass, and Arnau Sebé-Pedrós. A stony coral cell atlas illuminates the molecular and cellular basis of coral symbiosis, calcification, and immunity. *Cell*, 184:2973–2987.e18, 2021.

[11] Francesco Lamanna, Francisca Hervas-Sotomayor, A Phillip Oel, David Jandzik, Daniel Sobrido-Cameán, Gabriel N Santos-Durán, Megan L Martik, Jan Stundl, Stephen A Green, Thoomke Brüning, Katharina Mößinger, Julia Schmidt, Celine Schneider, Mari Sepp, Florent Murat, Jeramiah J Smith, Marianne E Bronner, María Celina Rodicio, Antón Barreiro-Iglesias, Daniel M Medeiros, Detlev Arendt, and Henrik Kaessmann. A lamprey neural cell type atlas illuminates the origins of the vertebrate brain. *Nat. Ecol. Evol.*, 7:1714–1728, 2023.

[12] Timothy Ting-Hsuan Wu, Kyle J Travaglini, Arjun Rustagi, Duo Xu, Yue Zhang, Leonid Andronov, Sori Jang, Astrid Gillich, Roozbeh Dehghannasiri, Giovanny J Martínez-Colón, Aimee Beck, Daniel Dan Liu, Aaron J Wilk, Maurizio Morri, Winston L Trope, Rob Bierman, Irving L Weissman, Joseph B Shrager, Stephen R Quake, Christin S Kuo, Julia Salzman, W E Moerner, Peter S Kim, Catherine A Blish, and Mark A Krasnow. Interstitial macrophages are a focus of viral takeover and inflammation in COVID-19 initiation in human lung. *J. Exp. Med.*, 221, 2024.

[13] Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Salvatore Candido, and Alexander Rives. Evolutionary-scale prediction of atomic level protein structure with a language model. *bioRxiv*, 2022.

[14] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023.

[15] Jianyuan Chen, Xuxi Ma, Weizhe Xu, and Jiashi Yang. Flex Attention: Efficient transformer training with dynamic sparse attention. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*. PMLR, 2022.

[16] Ilya Loshchilov and Frank Hutter. Fixing weight decay regularization in adam. *CoRR*, abs/1711.05101, 2017.

[17] Fabian Alexander Wolf, Philipp Angerer, and Fabian J Theis. Scanpy: large-scale single-cell gene expression data analysis. *Genome Biology*, 19(1):15, 2018.

[18] Tim Stuart, Andrew Butler, Paul Hoffman, Christoph Hafemeister, Efthymia Papalexi, William M Mauck III, Yuhan Hao, Marlon Stoeckius, Peter Smibert, and Rahul Satija. Comprehensive integration of single-cell data. *Cell*, 177(7):1888–1902.e21, 2019.

[19] Daniel Borcard, Pierre Legendre, and Pierre Drapeau. Partialling out the spatial component of ecological variation. *Ecology*, 73(3):1045–1055, 1992.

[20] Kenneth Ward Church and Patrick Hanks. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29, 1990.

[21] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300, 1995.