Transfer learning enables predictions in network biology

https://doi.org/10.1038/s41586-023-06139-9

Received: 29 March 2022

Accepted: 27 April 2023

Published online: 31 May 2023

Check for updates

Christina V. Theodoris[1,2,3,4]✉, Ling Xiao[2,5], Anant Chopra[6], Mark D. Chaffin[2], Zeina R. Al Sayed[2], Matthew C. Hill[2,5], Helene Mantineo[2,5], Elizabeth M. Brydon[6], Zexian Zeng[1,7], X. Shirley Liu[1,7,8] & Patrick T. Ellinor[2,5]✉

Mapping gene networks requires large amounts of transcriptomic data to learn the connections between genes, which impedes discoveries in settings with limited data, including rare diseases and diseases affecting clinically inaccessible tissues. Recently, transfer learning has revolutionized fields such as natural language understanding[1,2] and computer vision[3] by leveraging deep learning models pretrained on large-scale general datasets that can then be fine-tuned towards a vast array of downstream tasks with limited task-specific data. Here, we developed a context-aware, attention-based deep learning model, Geneformer, pretrained on a large-scale corpus of about 30 million single-cell transcriptomes to enable context-specific predictions in settings with limited data in network biology. During pretraining, Geneformer gained a fundamental understanding of network dynamics, encoding network hierarchy in the attention weights of the model in a completely self-supervised manner. Fine-tuning towards a diverse panel of downstream tasks relevant to chromatin and network dynamics using limited task-specific data demonstrated that Geneformer consistently boosted predictive accuracy. Applied to disease modelling with limited patient data, Geneformer identified candidate therapeutic targets for cardiomyopathy. Overall, Geneformer represents a pretrained deep learning model from which fine-tuning towards a broad range of downstream applications can be pursued to accelerate discovery of key network regulators and candidate therapeutic targets.

Mapping the gene regulatory networks that drive disease progression enables screening for molecules that correct the network by normalizing core regulatory elements, rather than targeting peripheral downstream effectors that may not be disease modifying[4,5]. However, mapping the gene network architecture requires large amounts of transcriptomic data to learn the connections between genes, which impedes network-correcting drug discovery in settings with limited data, including rare diseases and diseases affecting clinically inaccessible tissues. Although data remain limited in these settings, recent advances in sequencing technologies have driven a rapid expansion in the amount of transcriptomic data available from human tissues more broadly. Furthermore, single-cell technologies have facilitated the observation of transcriptomic states without averaging the expression of genes across multiple cells, potentially providing more precise data for inference of network interactions, especially in diseases driven by dysregulation of multiple cell types.

Recently, the concept of transfer learning has revolutionized fields such as natural language understanding[1,2] and computer vision[3] by leveraging deep learning models pretrained on large-scale general datasets that can then be fine-tuned towards a vast array of downstream tasks with limited task-specific data that would be insufficient to yield meaningful predictions when used in isolation. Unlike modelling approaches that necessitate retraining a new model from scratch for each task[6,7], this approach democratizes the fundamental knowledge learned during the large-scale pretraining phase to a multitude of downstream applications distinct from the pretraining learning objective, transferring knowledge to new tasks (Fig. 1a and Extended Data Fig. 1a,b). The advent of the self-attention mechanism[1,2] has further transformed the deep learning field by generating context-aware models that are able to pay attention to large input spaces and learn which elements are most important to focus on in each context, boosting predictions in a wide realm of applications[2,8]. Gene regulatory network architectures are highly context-dependent, and attention-based models, known as transformers, may be exceptionally suited to context-specific modelling of network dynamics.

Here, we developed a context-aware, attention-based deep learning model, Geneformer, pretrained on large-scale transcriptomic data to enable predictions in settings with limited data. We assembled a large-scale pretraining corpus, Genecorpus-30M, comprising 29.9 million human single-cell transcriptomes from a broad range of tissues from publicly available data. We then pretrained Geneformer on this corpus using a self-supervised masked learning objective to gain a fundamental

[1]Department of Data Science, Dana-Farber Cancer Institute, Boston, MA, USA. [2]Cardiovascular Disease Initiative and Precision Cardiology Laboratory, Broad Institute of MIT and Harvard, Cambridge, MA, USA. [3]Division of Genetics and Genomics, Boston Children's Hospital, Boston, MA, USA. [4]Harvard Medical School Genetics Training Program, Boston, USA. [5]Cardiovascular Research Center, Massachusetts General Hospital, Boston, MA, USA. [6]Precision Cardiology Laboratory, Bayer US LLC, Cambridge, MA, USA. [7]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, USA. [8]Center for Functional Cancer Epigenetics, Dana-Farber Cancer Institute, Boston, MA, USA. ✉e-mail: christina.theodoris@gladstone.ucsf.edu; ellinor@mgh.harvard.edu

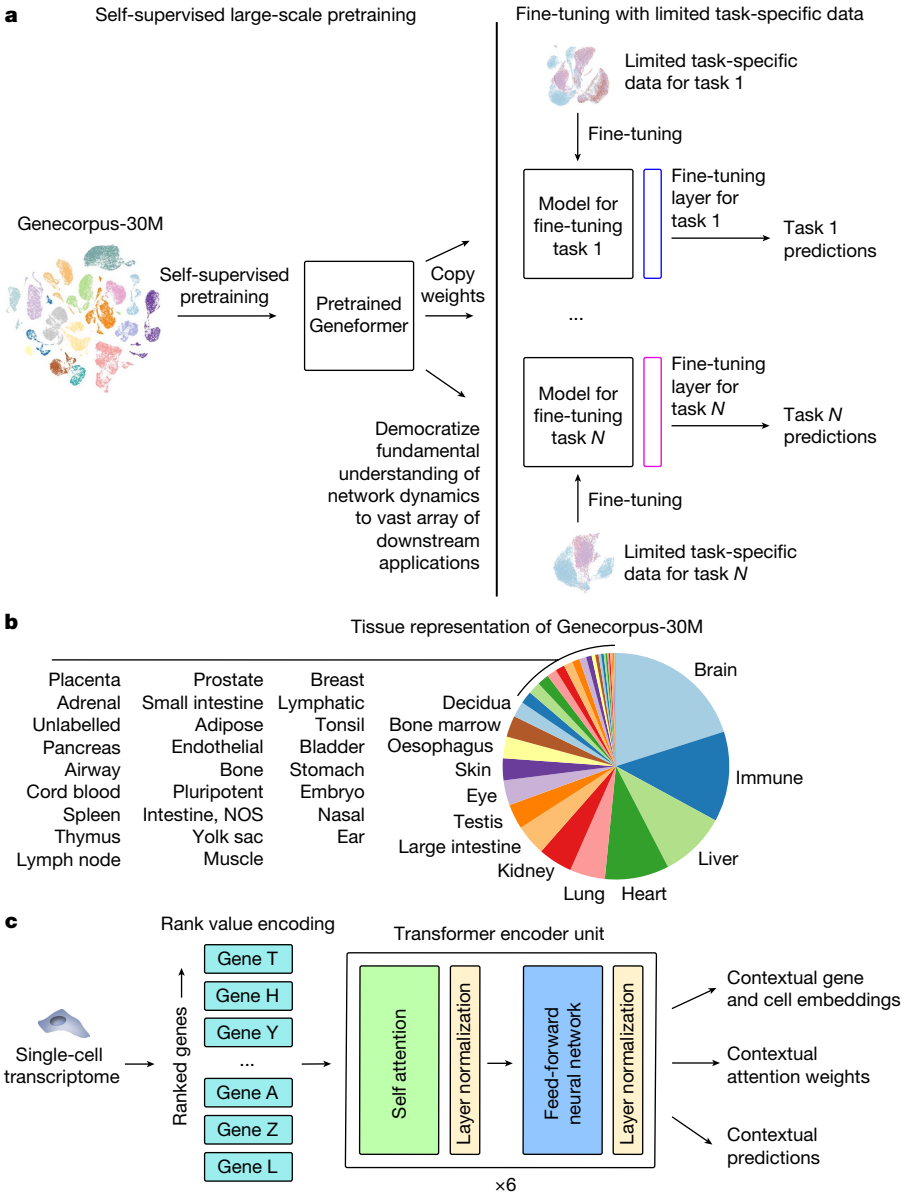


Fig. 1 | Geneformer architecture and transfer learning strategy. a, Schematic of transfer learning strategy with initial self-supervised large-scale pretraining, copying pretrained weights to models for each fine-tuning task, adding fine-tuning layer and fine-tuning with limited task-specific data towards each downstream task. Through the single initial self-supervised large-scale pretraining on a generalizable learning objective, the model gains fundamental knowledge of the learning domain that is then democratized to a multitude of downstream applications distinct from the pretraining learning objective, transferring knowledge to new tasks. **b**, Tissue representation of Genecorpus-30M. NOS, not otherwise specified. **c**, Pretrained Geneformer architecture. Each single-cell transcriptome is encoded into a rank value encoding that then proceeds through six layers of transformer encoder units with parameters as follows: input size of 2,048 (fully represents 93% of rank value encodings in Geneformer-30M), 256 embedding dimensions, four attention heads per layer and feed-forward size of 512. Geneformer uses full dense self-attention across the input size of 2,048. Extractable outputs include contextual gene and cell embeddings, contextual attention weights and contextual predictions.

understanding of network dynamics. The pretrained Geneformer accurately predicted dosage-sensitive disease genes and their downstream targets through a context-aware in silico deletion approach. Furthermore, fine-tuning Geneformer towards a diverse panel of downstream tasks relevant to chromatin and network dynamics using just a limited set of task-specific training examples demonstrated that Geneformer consistently boosted predictive accuracy. Applied to disease modelling of cardiomyopathy, Geneformer predicted candidate therapeutic targets whose experimental inhibition significantly improved cardiomyocyte contraction in an induced pluripotent stem cell (iPSC)-based model of the disease. Overall, Geneformer represents a pretrained deep learning model from which fine-tuning towards a broad range of downstream applications can be pursued to accelerate discovery of key network regulators and candidate therapeutic targets.

Geneformer architecture and pretraining

Geneformer is a context-aware, attention-based deep learning model pretrained on large-scale transcriptomic data to enable predictions in network biology with limited data through transfer learning (Fig. 1a). Geneformer harnesses the recent advent of self-attention[1,2] to maintain attention over the large input space of genes expressed in the transcriptome of each single cell and learn which genes are most important to focus on to optimize predictive accuracy within the given learning

objective. Importantly, network dynamics may vary across cell types, developmental timepoints or disease states. Accordingly, context awareness is a unique strength of Geneformer's model architecture that allows predictions specific to each cell context.

First, we assembled a large-scale pretraining corpus, Genecorpus-30M, comprising 29.9 million human single-cell transcriptomes from a broad range of tissues from publicly available data (Fig. 1b and Supplementary Table 1). We excluded cells with high mutational burdens (for example, malignant cells and immortalized cell lines) that could lead to substantial network rewiring without companion genome sequencing to facilitate interpretation, and we established metrics for scalable filtering to exclude possible doublets and/or damaged cells.

The transcriptome of each single cell is then presented to the model as a rank value encoding where genes are ranked by their expression in that cell normalized by their expression across the entire Genecorpus-30M (Fig. 1c). Although the rank-based representation has limitations including not fully taking advantage of the precise gene expression measurements provided in transcript counts, the rank value encoding provides a non-parametric representation of the transcriptome of each single cell and takes advantage of the many observations of the expression of each gene across Genecorpus-30M to prioritize genes that distinguish cell state. Specifically, this method will deprioritize ubiquitously highly expressed housekeeping genes by normalizing them to a lower rank. Conversely, genes such as transcription factors that may be expressed at low levels when they are expressed but have a high power to distinguish cell state will move to a higher rank within the encoding (Extended Data Fig. 1c). Furthermore, this rank-based approach may be more robust against technical artefacts that may systematically bias the absolute transcript counts value whereas the overall relative ranking of genes within each cell remains more stable.

The rank value encoding of the transcriptome of each single cell then proceeds through six transformer encoder units[1,2], each composed of a self-attention layer and feed forward neural network layer (Fig. 1c). Pretraining was accomplished using a masked learning objective, which has been shown in other informational fields[1,2] to improve generalizability of the foundational knowledge learned during pretraining for a wide range of downstream fine-tuning objectives. During pretraining, 15% of the genes within each transcriptome were masked, and the model was trained to predict which gene should be within each masked position in that specific cell state using the context of the remaining unmasked genes (Extended Data Fig. 1d–f). A principal strength of this approach is that it is entirely self-supervised and can be accomplished on completely unlabelled data, which allows the inclusion of large amounts of training data without being restricted to samples with accompanying labels. We implemented recent advances in distributed graphical processing unit (GPU) training[9,10] to allow efficient pretraining on the large-scale dataset.

Context awareness and batch integration

For each single-cell transcriptome presented to Geneformer, the model embeds each gene into a 256-dimensional space that encodes the characteristics of the gene specific to the context of that cell. We first tested whether the pretrained Geneformer's embedding of genes was impacted by common batch-dependent technical artefacts. We found that the gene embeddings were robust to sequencing platform[11], preservation method[12,13] and individual patient variability[14] (Extended Data Fig. 2a). However, gene embeddings were dependent on the context of other genes expressed in the cell, highlighting Geneformer's context awareness. When we in silico reprogrammed fibroblasts[15] by artificially adding *OCT4*, *SOX2*, *KLF4* and *MYC* to the front of their rank value encodings, the remaining genes in the transcriptome significantly shifted their embedding towards the iPSC state (Extended Data Fig. 2b,c). Embeddings of genes in iPSC-derived myogenic cells[16] showed similar context awareness with in silico differentiation by MYOD (Extended Data Fig. 2d,e). Furthermore, genes known to be highly context-dependent, such as NOTCH receptors, showed more variability in their embeddings across variable cell types[14] compared to the known housekeeping gene *GAPDH* (Extended Data Fig. 3).

Next, we integrated the embeddings of genes expressed in each cell to generate cell-level embeddings, which encode characteristics of the state of that single cell. Using a publicly available aortic aneurysm dataset[14] as a test case, we found that although the original data were impacted by interpatient variability, Geneformer cell embeddings clustered primarily by cell type and phenotype as opposed to individual patient (Extended Data Fig. 4a). Given that the pretrained Geneformer's cell embeddings were robust to these technical artefacts, we next tested whether fine-tuning would impact generalizability. Using a publicly available dataset[11] of iPSC differentiation to cardiomyocytes assayed in parallel on the Drop-seq (single cell) or DroNc-seq (single nucleus) platform, we tested whether fine-tuning the model to distinguish cell types using data from one platform would reduce generalizability to cells assayed on the other platform. Interestingly, the fine-tuned Geneformer's cell embeddings primarily clustered by cell types and showed improved integration of platforms compared to the original data even after batch effect removal using the ComBat[17] or Harmony[18] methods (Extended Data Fig. 4b–f).

Although Geneformer is most focused on understanding network dynamics rather than cell-level annotations, we further investigated Geneformer's performance in cell-type annotation given it is a common application for previously published models. We compared Geneformer to alternative XGBoost[7] and deep neural network-based[6] models. These methods train a new model from scratch for each separate tissue using the same supervised learning objective as is used for the final cell-type predictions in that specific tissue. Therefore, these approaches do not take advantage of the large amounts of data available more broadly that are not specifically labelled for that task. By contrast, Geneformer learns from large-scale unlabelled data during the self-supervised pretraining using a generalizable learning objective to gain fundamental knowledge that can then be transferred to a multitude of new and diverse fine-tuning tasks. Compared to these alternative methods, Geneformer boosted cell-type predictions in a variety of tissues, with the gap in performance by accuracy and macro F1 score increasing as the number of cell-type classes increased, indicating that Geneformer was robust in even increasingly complex multiclass prediction applications (Extended Data Figs. 5 and 6).

Gene dosage sensitivity predictions

We next tested whether Geneformer could boost predictions with limited data in a diverse set of downstream fine-tuning applications (Supplementary Table 2). A major challenge of interpreting copy number variants (CNVs) in genetic diagnosis is determining which genes are sensitive to changes in their dosage. Although conservation and allele frequency are commonly used to predict dosage sensitivity, these features do not vary across cell states and do not capture transcriptional dynamics that may inform contextual dosage sensitivity indicating which specific tissues would be affected by changes in the dosage of the gene. Using gene sets previously reported[19–21] to be dosage-sensitive versus dosage-insensitive, we fine-tuned Geneformer using only 10,000 random single-cell transcriptomes to distinguish dosage-sensitive versus dosage-insensitive transcription factors. The fine-tuned Geneformer significantly boosted the ability to predict dosage sensitivity compared to alternative methods (area under the receiver operating characteristic curve (AUC) 0.91) (Fig. 2a and Extended Data Fig. 7a). Notably, pretraining with larger and more diverse corpuses consistently improved the predictive power in the downstream task despite using the same amount of limited task-specific data for fine-tuning (Fig. 2b).

We then asked whether, without any further training, the fine-tuned model could predict the dosage sensitivity of a recently reported set

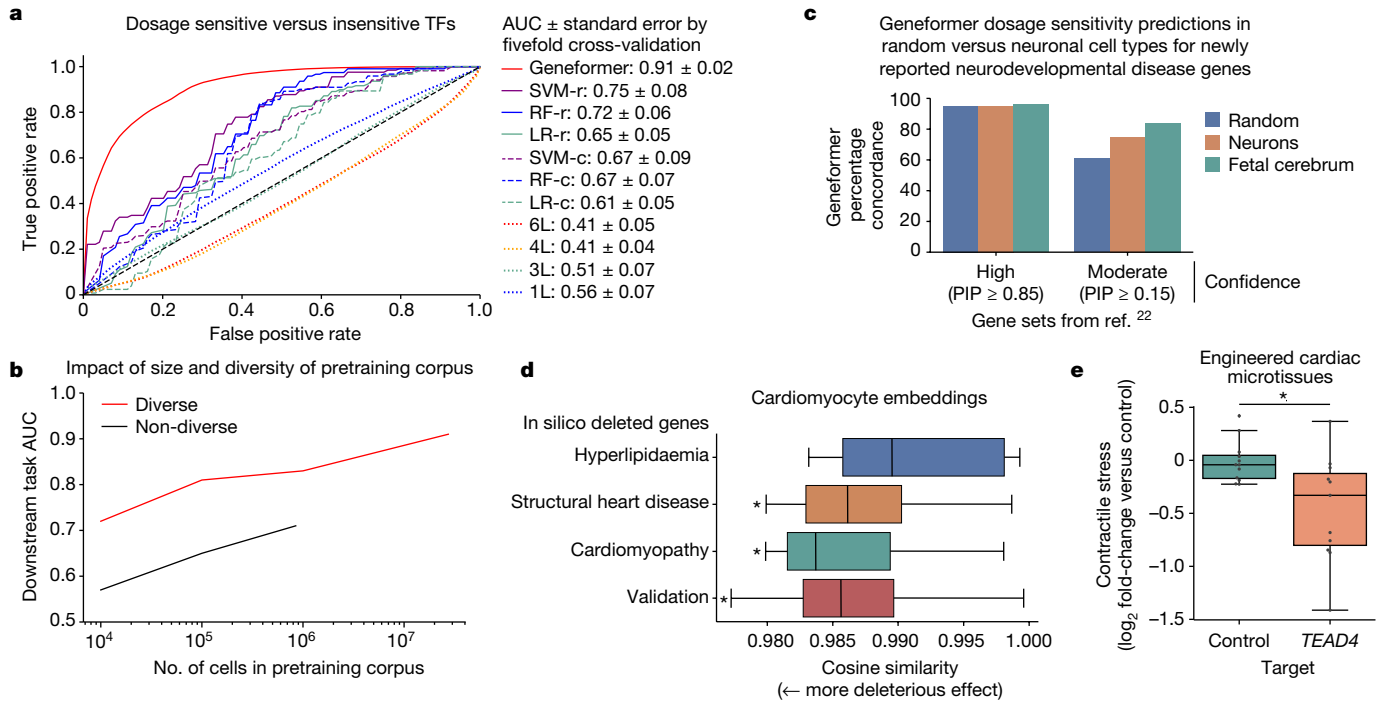


Fig. 2 | Geneformer boosted predictions of gene dosage sensitivity with limited data. a, A receiver operating characteristic curve (ROC curve) of Geneformer fine-tuned to distinguish dosage-sensitive versus dosage-insensitive transcription factors using limited data (10,000 cells) compared to alternative methods: support vector machine (SVM), random forest (RF) or logistic regression (LR) trained on gene ranks (-r) or counts (-c) or non-pretrained attention-based models with the same architecture as Geneformer (6 layers (L)) or shallower (4, 3 or 1L) with retained depth-to-width aspect ratios. **b**, Larger and more diverse pretraining corpuses improved predictive potential in downstream task of distinguishing dosage-sensitive versus dosage-insensitive transcription factors using the same limited task-specific data (10,000 cells). Diverse corpuses were randomly sampled from Genecorpus-30M, whereas non-diverse corpuses were randomly sampled from an oesophageal dataset[45]. **c**, Fine-tuned Geneformer's contextual dosage sensitivity predictions in (1) random cell types, (2) neurons (including adult) and (3) fetal cerebrum for neurodevelopmental disease genes newly reported by ref. 22. Authors reported either high- or moderate-confidence gene sets with the indicated posterior inclusion probability (PIP) scores. **d**, In silico deletion of genes associated with disease driven by cardiomyocyte pathology (cardiomyopathy and structural heart disease) had a more deleterious effect on cardiomyocyte embeddings compared to control cardiac disease genes expressed in cardiomyocytes but whose pathology occurs in non-cardiomyocyte cell types (hyperlipidaemia). Validation with experimental data from patients with cardiomyopathy (Fig. 6) demonstrated that in silico deletion of genes distinguishing the cardiomyopathy state was also predicted to be more deleterious than in silico deletion of control genes. (*$P < 0.05$ Wilcoxon, false discovery rate (FDR)-corrected). **e**, Contractile stress (force per unit area) of cardiac microtissues derived from wild-type (WT) iPSCs, exposed to either control treatment or guides promoting CRISPR-mediated knockout of Geneformer-predicted dosage-sensitive gene *TEAD4*. (Control $n = 12$, *TEAD4* $n = 11$; $P < 0.05$ Wilcoxon; points are replicates.) In **d** and **e**, centre line, median; box limits, upper and lower quartiles; whiskers, 1.5× interquartile range.

of disease genes (Fig. 2c). Collins et al. analysed CNVs from 753,994 individuals to define genes whose deletion was associated with primarily neurodevelopmental disease with either high or moderate confidence[22]. The fine-tuned Geneformer model correctly predicted the high-confidence genes to be dosage sensitive in the specific context of fetal cerebral cells with 96% concordance with the original study. The moderate-confidence genes reported by the authors were a much more permissive set (0.15–0.85 score versus high-confidence score cutoff greater than 0.85). The fine-tuned Geneformer predicted moderate-confidence genes to be dosage sensitive in fetal cerebral cells with 84% concordance with the original study. Interestingly, although the high-confidence genes, which may have a stronger effect, were predicted by Geneformer to be dosage sensitive at similar rates in fetal cerebral (96%) and other cells (95%), the predicted dosage sensitivity of the moderate-confidence genes seemed to be more context specific. The moderate-confidence genes were predicted to be dosage sensitive at a higher rate in fetal cerebral cells compared to neurons across any adult or developmental timepoint, consistent with the association of these genes with predominantly neurodevelopmental phenotypes in which adult neurons may be less relevant. They were predicted to be dosage sensitive at an even lower rate in random cells from any tissue, highlighting the context awareness of Geneformer.

We then designed an in silico deletion approach to identify genes whose deletion is predicted to have a deleterious effect in that particular cell context. We model gene deletion by removing the gene from the rank value encoding of the cell and quantifying the impact on the embeddings of the remaining genes in the encoding. To test this approach, we performed in silico deletion in fetal cardiomyocytes[23] using the pretrained Geneformer without any fine-tuning. In silico deletion of known cardiomyopathy and structural heart disease genes had a significantly larger effect than the control set of known hyperlipidaemia genes, which are expressed in cardiomyocytes and related to heart disease but whose phenotype affects cell types other than cardiomyocytes (Fig. 2d). In silico deletion of genes linked by a previous genome-wide association study[24] (GWAS) to cardiac magnetic resonance imaging (MRI) traits relevant to cardiac disease also had a larger effect compared to the control set (Extended Data Fig. 7b).

Overall, genes whose deletion was predicted to have the most deleterious effect on cardiomyocytes were significantly enriched for human phenotypes including cardiomyopathy and abnormal myocardial morphology (Supplementary Tables 3 and 4). Among the top 25 deleted genes with the most significant effect were transcription factors known to regulate myocardial development (for example, *FOXM1*; refs. 25,26) and entirely new dosage-sensitive gene candidates such as *TEAD4* (Supplementary Table 3). Experimental validation demonstrated

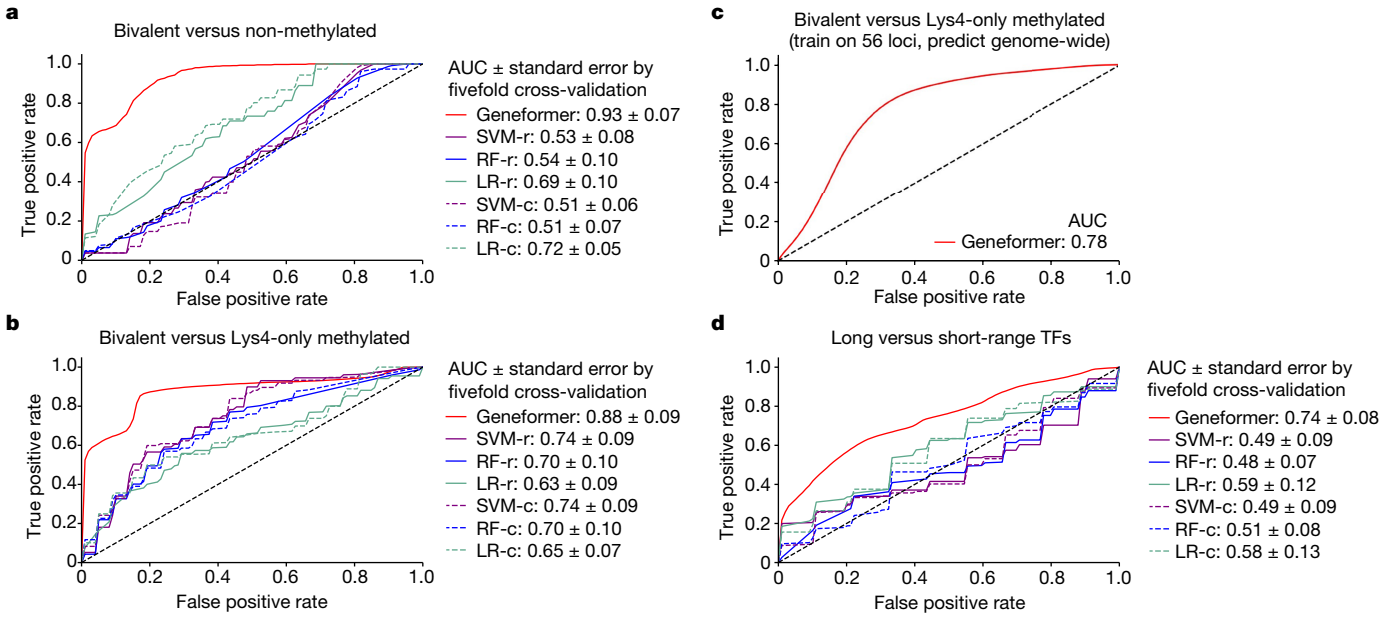


Fig. 3 | Geneformer boosted predictions of chromatin dynamics with limited data. a,**b**, ROC curve of Geneformer fine-tuned to distinguish bivalent versus non-methylated (**a**) or bivalent versus Lys4-only-methylated (**b**) genes in 56 conserved loci from ref. 28 using limited data (about 15,000 ESCs), compared to alternative methods. **c**, ROC curve of Geneformer's genome-wide predictions of bivalent versus Lys4-only-methylated genes after fine-tuning on only 56 loci as in **b**. **d**, ROC curve of Geneformer fine-tuned to distinguish long-range versus short-range transcription factors (TFs) using limited data (about 38,000 cells from iPSC to cardiomyocyte differentiation), compared to alternative methods. (Alternative methods described in Fig. 2).

that CRISPR-mediated knockout of candidate *TEAD4* in iPSC-derived cardiac microtissues caused a significant reduction in their ability to generate contractile stress (force per unit area) (Fig. 2e and Extended Data Fig. 7c). TEAD4 is a transcription factor involved in the Hippo signalling pathway[27], and future work is warranted to further examine its role in cardiac development.

Chromatin dynamics predictions

Bivalent chromatin structure is known to mark key developmental genes in embryonic stem cells (ESCs), maintaining their promoters poised for activation[28]. Bivalent domains consist of large regions of H3K27me3 harbouring smaller regions of H3K4me3. We fine-tuned Geneformer to distinguish bivalently marked genes from those whose promoters were unmethylated or marked solely by H3K4me3 using transcriptomes from about 15,000 ESCs[29]. The labelled gene set used for this fine-tuning included only genes found in 56 conserved regions of the genome, as previously reported[28]. Geneformer significantly boosted the ability to predict bivalently marked genes compared to alternative methods (AUC 0.93 and 0.88; bivalent versus unmethylated or H3K4me3-only, respectively) (Fig. 3a,b and Extended Data Fig. 7d,e). Furthermore, predictions were generalizable to the remainder of the genome that was excluded from fine-tuning (Fig. 3c and Extended Data Fig. 8a–c). Thus, by fine-tuning Geneformer using solely transcriptional data with only 56 labelled loci in about 15,000 ESCs, the model could predict the results of more recent studies[30] that included genome-wide profiling of bivalent domains.

Determining the genomic distances over which transcription factor binding influences downstream expression is valuable for interpreting regulatory variants and inferring target genes from transcription factor genome occupancy data. Others previously systematically integrated thousands of transcription factor-binding and histone-modification profiles assayed by chromatin immunoprecipitation sequencing (ChIP–seq) with thousands of gene expression profiles to identify two classes of transcription factor with distinct ranges of regulatory influence[31]. We fine-tuned Geneformer to distinguish these long- versus short-range transcription factors using only single-cell transcriptomes from about 34,000 cells undergoing iPSC to cardiomyocyte differentiation[11] with no associated ChIP–seq or genomic distance data. Again, Geneformer significantly boosted the ability to predict the regulatory range of transcription factors compared to alternative methods, whose predictions were near random (Fig. 3d and Extended Data Fig. 8d). Thus, fine-tuning the pretrained Geneformer model was able to improve predictions even for this higher-order transcription factor property of regulatory range, a particularly challenging characteristic to infer from transcriptional data alone.

Network dynamics predictions

Determining the hierarchy in gene networks enables the design of therapies targeting normalization of core regulatory elements that drive the disease process, rather than correction of peripheral downstream effectors that may not be disease modifying. We previously mapped the NOTCH1 (N1)-dependent gene network governing cardiac valve disease and identified central regulatory nodes whose correction had broad restorative impact on the network at large[4,5]. Mapping the network hierarchy required large amounts of transcriptional perturbation data from patient-specific cells with isogenic controls to learn the connections between genes.

We tested whether Geneformer could be fine-tuned to distinguish central versus peripheral factors within the N1-dependent gene network using only single-cell transcriptional data from about 30,000 normal endothelial cells (ECs) from the Heart Atlas[32] without any perturbation data. Again, Geneformer significantly boosted the ability to predict central versus peripheral factors compared to alternative methods (AUC 0.81) (Fig. 4a and Extended Data Fig. 8e). Furthermore, fine-tuning the pretrained Geneformer on the Heart Atlas ECs[32] was able to distinguish N1 downstream targets from non-targets without any perturbation data, further demonstrating the ability of the model to encode key features of gene network dynamics and again significantly boosting predictions compared to alternative methods (Fig. 4b and Extended Data Fig. 9a).

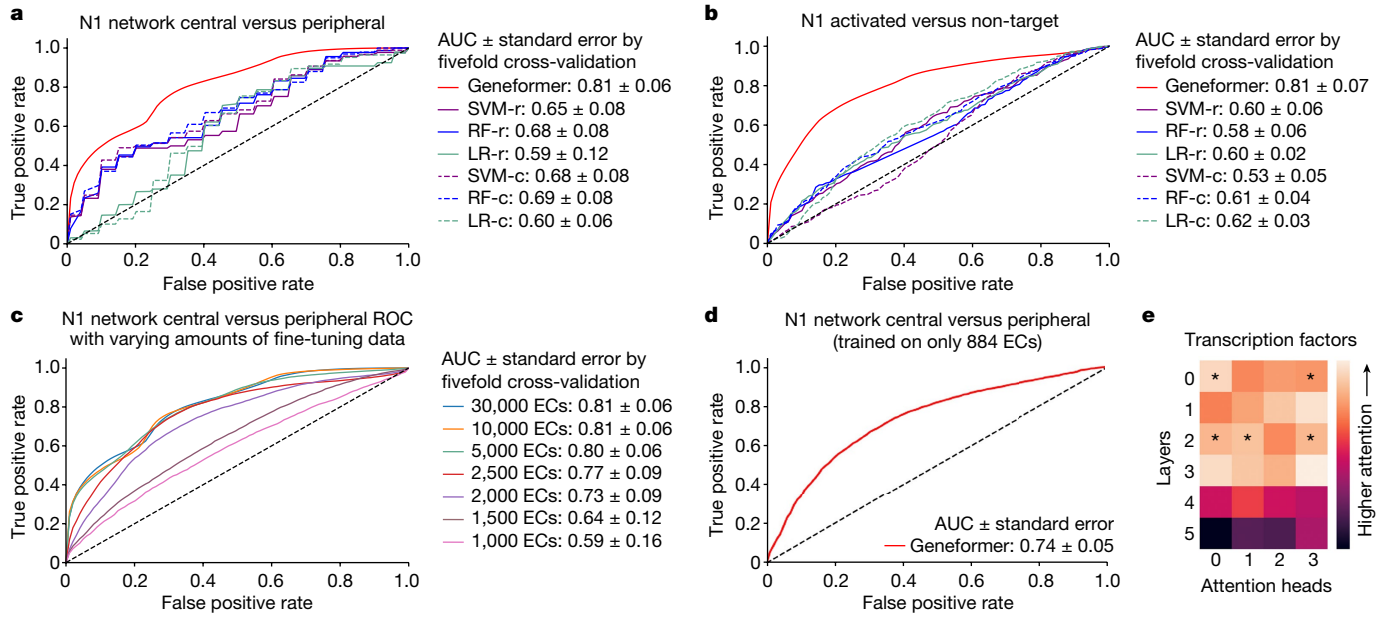


Fig. 4 | Geneformer encoded gene network hierarchy. a, ROC curve of Geneformer fine-tuned to distinguish central versus peripheral genes within the N1-dependent gene network using limited data (about 30,000 ECs), compared to alternative methods. **b**, ROC curve of Geneformer fine-tuned to distinguish N1-activated versus non-target genes using limited data (about 30,000 ECs), compared to alternative methods. **c**, ROC curve of Geneformer fine-tuned to distinguish central versus peripheral genes within the N1-dependent gene network using increasingly limited data (1,000–30,000 ECs). **d**, ROC curve of Geneformer fine-tuned to distinguish central versus peripheral genes within the N1-dependent gene network using increasingly limited but more relevant data (884 ECs from healthy or dilated aortas). AUC was higher than alternative methods trained on a larger dataset of about 30,000 ECs (Fig. 4a). **e**, Pretrained Geneformer attention weights of transcription factors indicated that the model learned in a completely self-supervised way the relative importance of transcription factors, which were more highly attended than other genes in 20% of attention heads ($P < 0.05$, Wilcoxon rank sum, FDR-corrected) and were more attended in earlier layers ($P < 0.05$, Wilcoxon rank sum). (Alternative methods described in Fig. 2).

To investigate the threshold for minimal data needed for fine-tuning, we fine-tuned the pretrained Geneformer with progressively smaller numbers of normal ECs from the Heart Atlas[32] to distinguish central versus peripheral factors within the N1-dependent gene network. We found that nearly equivalent predictive potential was retained even when reducing the fine-tuning data to only 5,000 ECs (Fig. 4c). Then, to determine whether Geneformer could generate meaningful predictions using an even more miniscule number of fine-tuning training examples when the task-specific data were more relevant to the learning objective, we fine-tuned the pretrained Geneformer using only 884 ECs from healthy versus dilated aortas[14]. Interestingly, Geneformer was able to distinguish central versus peripheral factors in the N1-dependent network with fine-tuning on this very minimal data to a better degree than the predictions of alternative methods trained on the larger dataset of about 30,000 ECs[32], demonstrating the strength of pretraining in enabling predictions from increasingly limited data (Fig. 4d and Extended Data Fig. 9b). More than twice as many general cardiac ECs were needed to gain similar predictive potential as was possible from fine-tuning with the more relevant data from healthy versus dilated aortas, suggesting that the minimum amount of fine-tuning data needed is dependent on both the specific application and relevance of the data to that task.

Pretraining encoded network hierarchy

To investigate how the model was learning network dynamics during the pretraining stage, we examined the pretrained Geneformer attention weights. The trained attention weights of the model for each gene reflect (1) which genes that gene pays attention to and (2) which genes pay attention to that gene. These attention weights are iteratively optimized during training to generate gene embeddings that best inform the correct answer for the given learning objective. Each of Geneformer's six layers has four attention heads that are meant to learn in an unsupervised manner to pay attention to distinct classes of genes to jointly improve predictions without previous knowledge of the biological function of any gene.

When examining the attention weights in aortic ECs[14], we found that 20% of attention heads significantly attended transcription factors more than other genes, indicating that specific attention heads learned, in an entirely self-supervised manner, the relative importance of transcription factors in distinguishing cell states (Fig. 4e). Furthermore, specific attention heads significantly attended central regulatory nodes to a greater degree than peripheral genes within N1-dependent network in ECs (Extended Data Fig. 9c). Concordantly, these centrality-driven attention heads consistently attended to a significantly greater degree the highest ranked genes in each cell's unique rank value encoding in aortic ECs, smooth muscle cells, T cells, and macrophage, monocyte and dendritic cells (which each have different sets of highest ranked genes on the basis of cell-type context) (Extended Data Fig. 9d).

Interestingly, attention heads in the earliest layers were consistently the most diverse in terms of gene ranks they attended, suggesting that the model initially orients to the observed cell state through a joint survey of distinct portions of the input space. The middle layers were most broad in terms of gene ranks they attended, and the final layers were dominated by centrality-driven attention heads that focused on the highest ranked genes that uniquely define each cell state (Extended Data Fig. 9c,d).

In silico gene network analysis

Given that the gene embeddings reflect the joint output of the attention weights of the network, we tested whether the pretrained Geneformer already encoded network connections between transcription factors and their targets before fine-tuning. We determined the genes whose

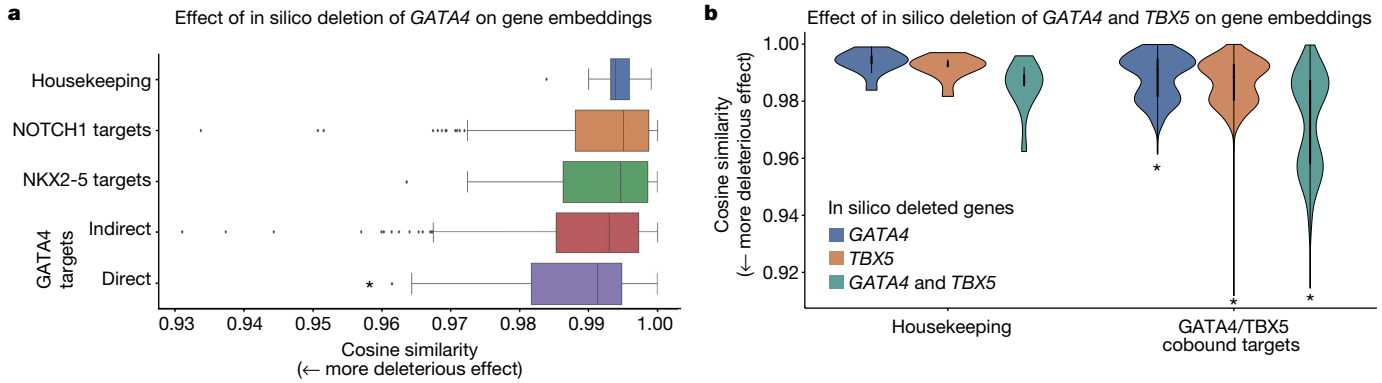


Fig. 5 | In silico deletion revealed network connections. a, In silico deletion of *GATA4* was significantly more deleterious to previously reported GATA4 direct targets[33] than to housekeeping genes, previously reported NOTCH1 targets[4], previously reported NKX2-5 targets[46] or GATA4 indirect targets[33] (*$P < 0.05$ Wilcoxon, FDR-corrected; centre line, median; box limits, upper and lower quartiles; whiskers, 1.5× interquartile range; points, outliers). **b**, In silico deletion of *GATA4* or *TBX5* alone was significantly more deleterious to previously reported GATA4/TBX5 cobound targets[33] than to housekeeping genes; in silico deletion of the combination of *GATA4* and *TBX5* was even more deleterious to cobound targets, significantly more than to housekeeping genes and significantly more than the sum of the effect of *GATA4* or *TBX5* alone on cobound targets (*$P < 0.05$ Wilcoxon, FDR-corrected).

embeddings in fetal cardiomyocytes[23] were most impacted by in silico deletion of *GATA4*, a known congenital heart disease gene. In silico deletion of *GATA4* had a significantly higher effect on genes known to be most significantly dysregulated by *GATA4* variants in a previously reported iPSC disease model of GATA4-related heart defects[33] (Extended Data Fig. 9e). Notably, direct GATA4 targets (as defined by ChIP–seq[33]) were significantly more impacted by in silico deletion of *GATA4* in fetal cardiomyocytes compared to indirect targets (Fig. 5a). Analogously, in silico deletion of *TBX5*, another known congenital heart disease gene, in fetal cardiomyocytes[23] more significantly impacted its direct targets (as defined by ChIP–seq[34]) compared to indirect targets and housekeeping genes (Extended Data Fig. 9f). These data suggest that in silico perturbation can be applied to model gene network connections.

Interestingly, the *GATA4* variant studied in the iPSC disease model disrupts the interaction of GATA4 with its binding partner, transcription factor TBX5 (ref. 33). We tested whether our in silico deletion approach could model the effect of deleting these two genes in combination (Fig. 5b). Indeed, in silico deletion of *GATA4* or *TBX5* alone had a significantly more deleterious effect on their known cobound targets[33] compared to housekeeping genes. Furthermore, in silico deletion of both *GATA4* and *TBX5* in combination had an even greater impact on their known cobound targets than the sum of their individual in silico deletion, suggesting that Geneformer recognized their cooperative action at these cobound targets.

In silico treatment analysis

We next tested whether our in silico perturbation strategy could be applied to model human disease and reveal candidate therapeutic targets (Fig. 6a). First, we fine-tuned Geneformer to distinguish cardiomyocytes[35] from non-failing hearts ($n = 9$) or hearts affected by hypertrophic ($n = 11$) or dilated ($n = 9$) cardiomyopathy with an overall out-of-sample accuracy of 90% (Fig. 6b and Extended Data Fig. 10a). We then determined the genes whose in silico deletion or activation in cardiomyocytes from non-failing hearts significantly shifted the fine-tuned Geneformer cell embeddings towards the hypertrophic or dilated cardiomyopathy states (Fig. 6c,d, Extended Data Fig. 10b,c and Supplementary Tables 5–11). Overall, the model identified 447 genes whose loss was predicted to shift cardiomyocytes towards the hypertrophic cardiomyopathy state, which were enriched for pathways including Titin binding[36] and sarcomere organization[37] known to impact hypertrophic cardiomyopathy pathogenesis. The model identified 478 genes whose loss was predicted to shift cardiomyocytes towards dilated cardiomyopathy, which were enriched for pathways involved in muscle contraction[38] and mitochondrial[39] function.

Then, we performed in silico treatment analysis in cardiomyocytes from hypertrophic or dilated cardiomyopathy patients to determine whether inhibition or activation of specific pathways would shift the cell embeddings back towards the non-failing heart state (Fig. 6e, Extended Data Fig. 10d and Supplementary Tables 12–15). Top enriched pathways for hypertrophic cardiomyopathy pointed to candidate cardiomyocyte-specific therapeutic targets including ADCY5, disruption of which is associated with longevity and protection against cardiomyopathy in mouse models[40], as well as druggable targets[41] including SRPK3, a downstream effector of MEF2 (ref. 42), which is known to play a critical role in myocardial cell hypertrophy[43].

We then performed experimental validation to determine whether inhibition of Geneformer-predicted therapeutic candidates for dilated cardiomyopathy could improve cardiomyocyte function in an experimental model of the disease. *Titin* (*TTN*) truncating mutations are the leading cause of dilated cardiomyopathy in humans and are found in about 20% of affected patients[36]. iPSC-derived cardiac microtissues harbouring a truncating variant ($TTN^{+/-}$) in the A-band are known to exhibit reduced contractile stress compared to isogenic $TTN^{+/+}$ controls[36]. Strikingly, CRISPR-mediated knockout of both Geneformer-predicted targets *GSN* and *PLN* in the $TTN^{+/-}$ cells significantly improved the contractile stress of the $TTN^{+/-}$ cardiac microtissues, validating these genes as promising candidate therapeutic targets for this disease (Fig. 6f,g and Extended Data Fig. 10e). These findings provide experimental validation in support of the utility of Geneformer as a tool for discovery of candidate therapeutic targets in human disease.

Discussion

In sum, we developed a context-aware deep learning model, Geneformer, pretrained on large-scale transcriptomic data to enable predictions in settings with limited data. Through the observation of a vast number of cell states during the pretraining process, Geneformer gained a fundamental understanding of network dynamics, encoding network hierarchy in the attention weights of the model in a completely self-supervised manner. Geneformer's ability to predict dosage-sensitive disease genes through the context-aware in silico deletion approach represents a valuable asset for interpretation of genetic variants, including prioritization of GWAS hits driving complex traits, and the specific tissues they are expected to affect. Experimental validation of a dosage-sensitive gene candidate in fetal

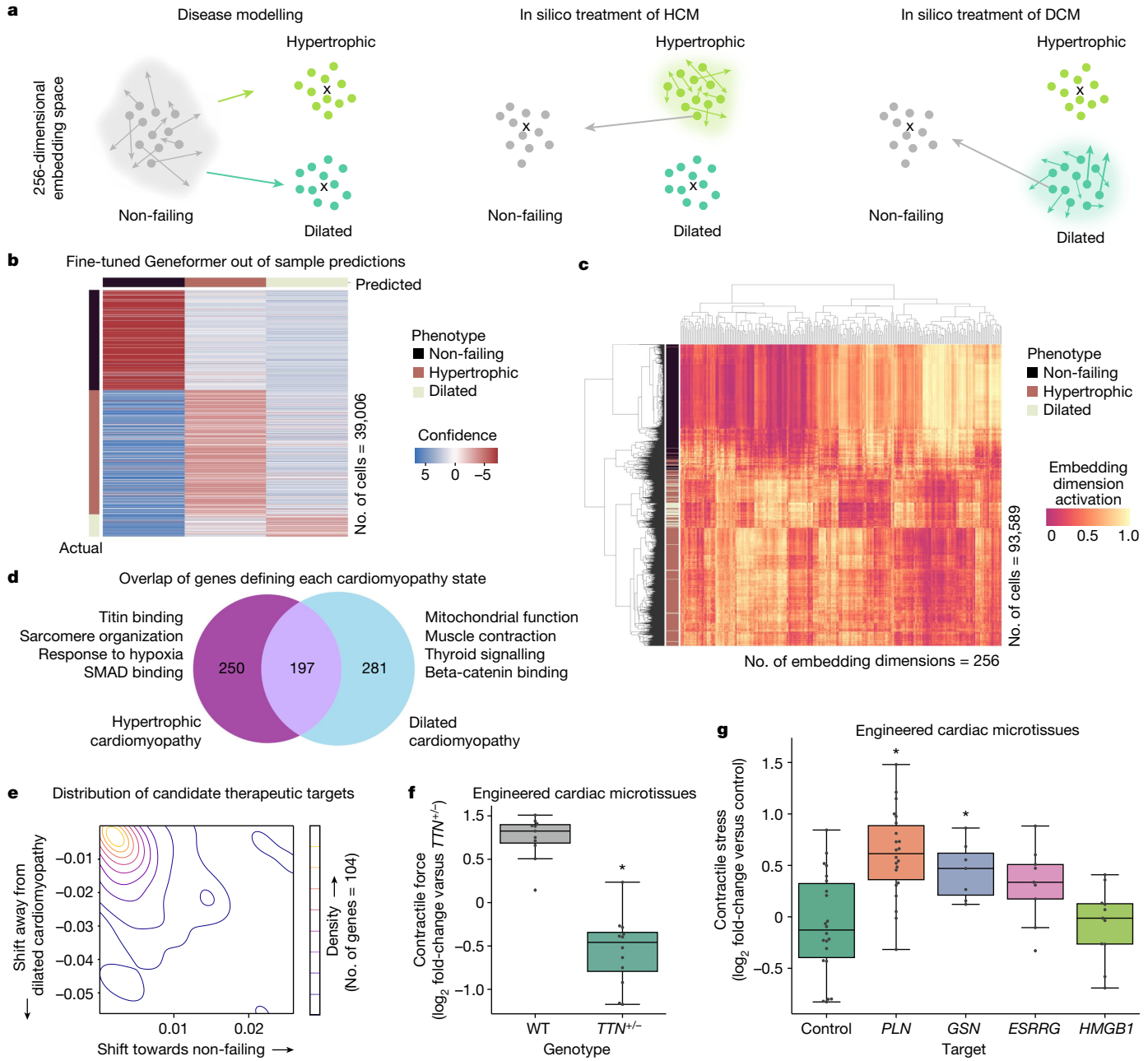


Fig. 6 | In silico treatment revealed candidate therapeutic targets. **a**, Fine-tuning Geneformer to distinguish cardiomyocytes from non-failing hearts or hearts affected by hypertrophic or dilated cardiomyopathy (HCM and DCM) defines the embedding position of each cell state. Then, disease modelling (left) can be performed by in silico deleting or activating random genes within non-failing cardiomyocytes to define the random distribution (grey cloud) and thereby identify genes whose in silico deletion or activation shifts the embedding significantly towards either the hypertrophic or dilated cardiomyopathy state. The reverse approach is taken for in silico treatment analysis (centre and right). **b**, Out-of-sample predictions of Geneformer fine-tuned to distinguish cardiomyocytes from non-failing hearts or hearts affected by hypertrophic or dilated cardiomyopathy. Accuracy 90%; precision 82%; recall 87%. (Training data: non-failing $n = 9$, hypertrophic $n = 11$, dilated $n = 9$, total 93,589 cells; out-of-sample data: non-failing $n = 4$, hypertrophic $n = 4$, dilated $n = 2$, total 39,006 cells). **c**, Hierarchical clustering of fine-tuned Geneformer cardiomyocyte cell embeddings. **d**, Overlap of genes whose in silico deletion in cardiomyocytes from non-failing hearts significantly shifted the fine-tuned Geneformer cell embeddings towards the hypertrophic or dilated cardiomyopathy states and gene ontology terms enriched for each state. **e**, Distribution of mean embedding shift in response to in silico deletion of candidate therapeutic targets in cardiomyocytes from hypertrophic cardiomyopathy ($n = 104$ genes). **f**, Contractile force of cardiac microtissues derived from WT iPSCs or iPSCs with a *TTN* truncating mutation modelling dilated cardiomyopathy (WT n = 11, $TTN^{+/-}$ $n = 12$, $^*P < 0.05$ Wilcoxon). **g**, Contractile stress (force per unit area) of cardiac microtissues derived from $TTN^{+/-}$ iPSCs exposed to either control treatment or guides promoting CRISPR-mediated knockout of Geneformer-predicted therapeutic targets. ($TTN^{+/-}$ + control treatment $n = 22$, $TTN^{+/-}$ + CRISPR guides targeting knockout of *PLN* $n = 22$, *GSN* $n = 7$, *ESRRG* $n = 9$ or *HMGB1* $n = 11$; $P < 0.05$ Wilcoxon, FDR-corrected). In **f** and **g**, centre line, median; box limits, upper and lower quartiles; whiskers, 1.5× interquartile range; points, experimental replicates.

cardiomyocytes, *TEAD4*, supports the utility of Geneformer for driving biological insights in human development. Applied to disease modelling of cardiomyopathy using a limited number of patient samples, Geneformer predicted candidate therapeutic targets whose experimental targeting in an iPSC disease model led to significant functional improvement. In silico treatment analysis using limited data may thus

enable therapeutic discovery in innumerable diseases that have been previously impeded by limited data because they are rare or affect clinically inaccessible tissue.

Furthermore, we found that pretraining with larger and more diverse corpuses consistently improved Geneformer's predictive power, in agreement with observations that large-scale pretraining allows training of deeper models that ultimately have greater predictive potential in fields including natural language understanding, computer vision and mathematical problem-solving[44]. Furthermore, exposure to hundreds of experimental datasets during pretraining also seemed to promote robustness to batch-dependent technical artefacts and individual variability that commonly impact single-cell analyses in biology. These findings suggest that as the amount of publicly available transcriptomic data continues to expand, future models pretrained on even larger-scale corpuses may open opportunities to achieve meaningful predictions in even more elusive tasks with increasingly limited task-specific data. Overall, Geneformer represents a pretrained deep learning model whose fundamental understanding of network dynamics can now be democratized to a broad range of downstream applications to accelerate discovery of key network regulators and candidate therapeutic targets in settings with limited data.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-023-06139-9.

1. Vaswani, A. et al. Attention is all you need. Preprint at https://doi.org/10.48550/arXiv.1706.03762 (2017).
2. Devlin, J., Chang, M. W., Lee, K. & Toutanova, K. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proc. 2019 Conference North American Chapter of the Association for Computational Linguistics: Human Language Technologies* Vol. 1 (eds Burstein, J. et al.) 4174–4186 (Association for Computational Linguistics, 2019).
3. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition* 770–778 (IEEE, 2016).
4. Theodoris, C. V. et al. Human disease modeling reveals integrated transcriptional and epigenetic mechanisms of NOTCH1 haploinsufficiency. *Cell* **160**, 1072–1086 (2015).
5. Theodoris, C. V. et al. Network-based screen in iPSC-derived cells reveals therapeutic candidate for heart valve disease. *Science* **371**, eabd0724 (2021).
6. Shao, X. et al. ScDeepSort: a pre-trained cell-type annotation method for single-cell transcriptomics using deep learning with a weighted graph neural network. *Nucleic Acids Res.* **49**, e122 (2021).
7. Lieberman, Y., Rokach, L. & Shay, T. CaSTLe—classification of single cells by transfer learning: harnessing the power of publicly available single cell RNA sequencing experiments to annotate new experiments. *PLoS ONE* **13**, e0205499 (2018).
8. Lin, T., Wang, Y., Liu, X. & Qiu, X. A survey of transformers. Preprint at https://doi.org/10.48550/arXiv.2106.04554 (2021).
9. Ren, J. et al. ZeRO-offload: democratizing billion-scale model training. In *Proc. 2021 USENIX Annual Technical Conference* 551–564 (USENIX, 2021).
10. Rajbhandari, S., Rasley, J., Ruwase, O. & He, Y. Zero: memory optimizations toward training trillion parameter models. In *International Conference for High Performance Computing, Networking, Storage and Analysis* 1–16 (IEEE, 2020).
11. Selewa, A. et al. Systematic comparison of high-throughput single-cell and single-nucleus transcriptomes during cardiomyocyte differentiation. *Sci. Rep.* **10**, 1535 (2020).
12. *10x Genomics Datasets* https://www.10xgenomics.com/resources/datasets/frozen-pbm-cs-donor-a-1-standard-1-1-0.
13. *10X Genomics Datasets* https://www.10xgenomics.com/resources/datasets/fresh-68-k-pbm-cs-donor-a-1-standard-1-1-0.
14. Li, Y. et al. Single-cell transcriptome analysis reveals dynamic cell populations and differential gene expression patterns in control and aneurysmal human aortic tissue. *Circulation* **142**, 1374–1388 (2020).
15. Xing, Q. R. et al. Diversification of reprogramming trajectories revealed by parallel single-cell transcriptome and chromatin accessibility sequencing. *Sci. Adv.* **6**, 463–474 (2020).
16. Guo, D. et al. iMyoblasts for ex vivo and in vivo investigations of human myogenesis and disease modeling. *eLife* **11**, e70341 (2022).
17. Zhang, Y., Parmigiani, G. & Johnson, W. E. ComBat-seq: batch effect adjustment for RNA-seq count data. *NAR Genom. Bioinform.* **2**, lqaa078 (2020).
18. Korsunsky, I. et al. Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat. Methods* **16**, 1289–1296 (2019).
19. Lek, M. et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291 (2016).
20. Shihab, H. A., Rogers, M. F., Campbell, C. & Gaunt, T. R. HIPred: an integrative approach to predicting haploinsufficient genes. *Bioinformatics* **33**, 1751–1757 (2017).
21. Ni, Z., Zhou, X. Y., Aslam, S. & Niu, D. K. Characterization of human dosage-sensitive transcription factor genes. *Front. Genet.* **10**, 1208 (2019).
22. Collins, R. L. et al. A cross-disorder dosage sensitivity map of the human genome. *Cell* **185**, 3041–3055 (2022).
23. Cao, J. et al. A human cell atlas of fetal gene expression. *Science* **370**, 808 (2020).
24. Pirruccello, J. P. et al. Analysis of cardiac magnetic resonance imaging in 36,000 individuals yields genetic insights into dilated cardiomyopathy. *Nat. Commun.* **11**, 2254 (2020).
25. Bolte, C. et al. Expression of Foxm1 transcription factor in cardiomyocytes is required for myocardial development. *PLoS ONE* **6**, e22217 (2011).
26. Bolte, C. et al. Postnatal ablation of Foxm1 from cardiomyocytes causes late onset cardiac hypertrophy and fibrosis without exacerbating pressure overload-induced cardiac remodeling. *PLoS ONE* **7**, e48713 (2012).
27. Currey, L., Thor, S. & Piper, M. TEAD family transcription factors in development and disease. *Development* **148**, dev196675 (2021).
28. Bernstein, B. E. et al. A bivalent chromatin structure marks key developmental genes in embryonic stem cells. *Cell* **125**, 315–356 (2006).
29. Franzén, O., Gan, L.-M. & Björkegren, J. L. M. PanglaoDB: a web server for exploration of mouse and human single-cell RNA sequencing data. *Database* **2019**, baz406 (2019).
30. Pan, G. et al. Whole-genome analysis of histone H3 lysine 4 and lysine 27 methylation in human embryonic stem cells. *Cell Stem Cell* **1**, 299–312 (2007).
31. Chen, C. H. et al. Determinants of transcription factor regulatory range. *Nat. Commun.* **11**, 2472 (2020).
32. Litviňuková, M. et al. Cells of the adult human heart. *Nature* **588**, 455–472 (2020).
33. Ang, Y. S. et al. Disease model of GATA4 mutation reveals transcription factor cooperativity in human cardiogenesis. *Cell* **167**, 1734–1749 (2016).
34. Kathiriya, I. S. et al. Modeling human TBX5 haploinsufficiency predicts regulatory networks for congenital heart disease. *Dev. Cell* **56**, 292–309 (2021).
35. Chaffin, M. et al. Single-nucleus profiling of human dilated and hypertrophic cardiomyopathy. *Nature* **608**, 174–180 (2022).
36. Hinson, J. T. et al. Titin mutations in iPS cells define sarcomere insufficiency as a cause of dilated cardiomyopathy. *Science* **349**, 982–986 (2015).
37. Seidman, C. E. & Seidman, J. G. Identifying sarcomere gene mutations in hypertrophic cardiomyopathy: a personal history. *Circ. Res.* **108**, 743–750 (2011).
38. Kamisago, M. et al. Mutations in sarcomere protein genes as a cause of dilated cardiomyopathy. *New Engl. J. Med.* **343**, 1688–1696 (2000).
39. Ramaccini, D. et al. Mitochondrial function and dysfunction in dilated cardiomyopathy. *Front. Cell Dev. Biol.* https://doi.org/10.3389/fcell.2020.624216 (2021).
40. Ho, D., Yan, L., Iwatsubo, K., Vatner, D. E. & Vatner, S. F. Modulation of β-adrenergic receptor signaling in heart failure and longevity: targeting adenylyl cyclase type 5. *Heart Fail. Rev.* **15**, 495–512 (2010).
41. Wagner, A. H. et al. DGIdb 2.0: mining clinically relevant drug-gene interactions. *Nucleic Acids Res.* **44**, D1036–D1044 (2016).
42. Nakagawa, O. et al. Centronuclear myopathy in mice lacking a novel muscle-specific protein kinase transcriptionally regulated by MEF2. *Genes Dev.* **19**, 2066–2077 (2005).
43. Akazawa, H. & Komuro, I. Roles of cardiac transcription factors in cardiac hypertrophy. *Circ. Res.* **92**, 1079–1088 (2003).
44. Henighan, T. et al. Scaling laws for autoregressive generative modeling. Preprint at https://doi.org/10.48550/arXiv.2010.14701 (2020).
45. Madissoon, E. et al. ScRNA-seq assessment of the human lung, spleen, and esophagus tissue stability after cold preservation. *Genome Biol.* **21**, 1 (2019).
46. Anderson, D. J. et al. NKX2-5 regulates human cardiomyogenesis via a HEY2 dependent transcriptional network. *Nat. Commun.* **9**, 1373 (2018).

Methods

Assembly and rank value encoding of transcriptomes in Genecorpus-30M

Assembly and uniform processing of single-cell transcriptomes. We assembled a large-scale pretraining corpus, Genecorpus-30M, comprising 29.9 million (29,900,531) human single-cell transcriptomes from a broad range of tissues from publicly available data (Fig. 1b and Supplementary Table 1). We excluded cells with high mutational burdens (for example, malignant cells and immortalized cell lines) that could lead to substantial network rewiring without companion genome sequencing to facilitate interpretation. We only included droplet-based sequencing platforms to assure expression value unit comparability. Overall, 561 datasets were included and stored as uniform files in the .loom HDF5 format including metadata from the original studies as row (feature) and column (cell) attributes described below.

Publicly available datasets containing raw counts were collected from the National Center for Biotechnology Information (NCBI) Gene Expression Omnibus (GEO), NCBI Sequence Read Archive (SRA), Human Cell Atlas, European Molecular Biology Laboratory-European Bioinformatics Institute (EMBL-EBI) Single Cell Expression Atlas, Broad Institute Single Cell Portal, Brotman Baty Institute (BBI)-Allen Single Cell Atlases, Tumor Immune Single-cell Hub (TISCH) (excluding malignant cells), Panglao Database, 10x Genomics, University of California, Santa Cruz Cell Browser, European Genome-phenome Archive, Synapse, Riken, Zenodo, National Institutes of Health (NIH) Figshare Archive, NCBI dbGap, Refine.bio, China National GeneBank Sequence Archive, Mendeley Data and individual communication with authors of the original studies[11,23,29,32,45,47–153]. Further resources for collecting information about suitable studies included Entrez Direct tools and the dataset summary from Database 2020 (ref. 154). Tools used in conversion of data to uniform .loom HDF5 files included loompy, scanpy[155], anndata, scipy, numpy, pandas, Cellranger and LoomExperiment.

Row feature attributes included Ensembl annotations for the gene ID, ID version (if provided by original study), name and type (for example, protein coding, microRNA, mitochondrial and so on). Annotation data were retrieved from Ensembl and MyGene[156]. Column cell attributes included a unique Genecorpus-30M cell ID comprising the dataset name, sample name and cell barcode from that dataset. The dataset and sample names were also included as separate individual attributes such that the cell barcode can be derived by subtracting these from the unique Genecorpus-30M cell ID if needed. Column cell attributes also included the principal organ included in the dataset, which we annotated as one of the following categories: adipose, adrenal, airway, bladder, bone, bone_marrow, brain, breast, cord_blood, decidua, ear, embryo, endothelial, eye, heart, immune, intestine_unspecified, kidney, large_intestine, liver, lung, lymph_node, lymphatic, muscle, nasal, oesophagus, pancreas, placenta, pluripotent_stem_cell, prostate, skin, small_intestine, spleen, stomach, testis, thymus, tonsil, various, yolk_sac. Column cell attributes also included the specific organ(s) included in the dataset on the basis of metadata provided by the original study. If the original study included cell-type annotations, we included these as a cell-type column attribute for each cell as well. We also included the sequencing platform used.

Column cell attributes also included several calculated measurements for each cell: the total number of read counts, the percentage of mitochondrial reads, the number of genes Ensembl-annotated as protein-coding or miRNA genes and whether the cell passed the quality-control metrics we established for scalable filtering of the cells to exclude possible doublets and/or damaged cells. Only cells that passed these filtering metrics were used for downstream analyses in this work. Specifically, datasets were filtered to retain cells with total read counts within 3 s.d. of the mean within that dataset and mitochondrial reads within 3 s.d. of the mean within that dataset. Ensembl-annotated protein-coding and miRNA genes were used for downstream analysis. Cells with less than seven detected Ensembl-annotated protein-coding or miRNA genes were excluded as the 15% masking used for the pretraining learning objective would not reliably mask a gene in cells with fewer detected genes. Ultimately, 27.4 million (27,406,217) cells passed the defined quality filters.

Rank value encoding of single-cell transcriptomes. We developed a rank value encoding method that provides a non-parametric representation of the transcriptome of each single cell, ranking genes by their expression within that cell normalized by their expression across the entire Genecorpus-30M (Fig. 1c). This method takes advantage of the many observations of the expression of each gene across Genecorpus-30M to prioritize genes that distinguish cell state. Specifically, this method will deprioritize ubiquitously highly expressed housekeeping genes by normalizing them to a lower rank. Conversely, genes such as transcription factors that may be expressed at low levels when they are expressed but have a high power to distinguish cell state will move to a higher rank within the encoding (Extended Data Fig. 1c). Furthermore, this rank-based approach may be more robust against technical artefacts that may systematically bias the absolute transcript counts value whereas the overall relative ranking of genes within each cell remains more stable.

To accomplish this, we first calculated the non-zero median value of expression of each detected gene across all cells passing quality filtering from the entire Genecorpus-30M. We aggregated the transcript count distribution for each gene in a memory-efficient manner by scanning through chunks of .loom data using loompy, normalizing the gene transcript counts in each cell by the total transcript count of that cell to account for varying sequencing depth and updating the normalized count distribution of the gene within the t-digest[157] data structure developed for accurate online accumulation of rank-based statistics. We then normalized the genes in each single-cell transcriptome by the non-zero median value of expression of that gene across Genecorpus-30M and ordered the genes by the rank of their normalized expression in that specific cell. Of note, we opted to use the non-zero median value of expression rather than include zeros in the distribution so as not to weight the value by tissue representation within Genecorpus-30M, assuming that a representative range of transcript values would be observed within the cells in which each gene was detected. This normalization factor for each gene is calculated once from the pretraining corpus and is used for all future datasets presented to the model. The provided tokenizer code includes this normalization procedure and should be used for tokenizing new datasets presented to Geneformer to ensure consistency of the normalization factor used for each gene.

The rank value encodings for each single-cell transcriptome were then tokenized on the basis of a total vocabulary of 25,424 protein-coding or miRNA genes detected in a median of 173,152 cells within Genecorpus-30M. The vocabulary also included two more special tokens for padding and masking. The tokenized data were stored within the Huggingface Datasets[158] structure, which is based on the Apache Arrow format that allows processing of large datasets with zero-copy reads without memory constraints. Of note, this strategy is also space-efficient as the genes are stored as ranked tokens as opposed to the exact transcript values, and we only store genes detected within each cell rather than the full sparse dataset that includes all of the undetected genes.

Geneformer architecture and pretraining

Geneformer architecture. Geneformer is composed of six transformer encoder units[1,2], each composed of a self-attention layer and feed forward neural network layer with the following parameters: input size of 2,048 (fully represents 93% of rank value encodings in Genecorpus-30M), 256 embedding dimensions, four attention heads

per layer and feed forward size of 512 (Fig. 1c). Geneformer uses full dense self-attention across the input size of 2,048. Depth was chosen on the basis of the maximum depth for which there were sufficient data to pretrain as it has been established that this approach yields the greatest predictive potential in other informational fields including natural language understanding, computer vision and mathematical problem-solving[44]. Furthermore, we maximized the amount of context (input size) considered by the model with full attention based on the number of genes standardly detected in each cell within the pretraining corpus. Further parameters were as follows: nonlinear activation function, rectified linear unit (ReLU); dropout probability for all fully connected layers, 0.02; dropout ratio for attention probabilities, 0.02; standard deviation of the initializer for weight matrices, 0.02; epsilon for layer normalization layers, 1×10^{-12}. Modelling was implemented in pytorch and using the Huggingface Transformers library[159] for model configuration, data loading and training.

Geneformer pretraining and performance optimization. Pretraining was accomplished using a masked learning objective, which has been shown in other informational fields[1,2] to improve generalizability of the foundational knowledge learned during pretraining for a wide range of downstream fine-tuning objectives. During pretraining, 15% of the genes within each transcriptome were masked and the model was trained to predict which gene should be within each masked position in that specific cell state using the context of the remaining unmasked genes. A principal strength of this approach is that it is entirely self-supervised and can be accomplished on completely unlabelled data, which allows the inclusion of large amounts of training data without being restricted to samples with accompanying labels. Pretraining hyperparameters were optimized to the following final values: max learning rate, 1×10^{-3}; learning scheduler, linear with warmup; optimizer, Adam with weight decay fix[160]; warmup steps, 10,000; weight decay, 0.001; batch size, 12. Tensorboard was used for experimentation tracking, and the model was pretrained for three epochs.

As the input size of 2,048 is considerably large for a full dense self-attention model (for example, BERT[1,2] input size is 512) and transformers have a quadratic memory and time complexity $\mathcal{O}(L^2)$ with respect to input size, we implemented measures to optimize efficiency of large-scale pretraining. The trainer from the Huggingface Transformers library[159] was used for pretraining with the substitution of a custom tokenizer to implement dynamic, length-grouped padding, which minimized computation on padding and achieved a 29.4× speedup in pretraining. This process takes a randomly sampled megabatch and then orders minibatches by their length in descending order (to ensure that any memory constraints are encountered earlier). Minibatches are then dynamically padded, minimizing the computation wasted on padding due to being length grouped. We also implemented recent advances in distributed GPU training[9,10] to allow efficient pretraining on the large-scale dataset using Deepspeed, which partitions parameters, gradients and optimizer states across available GPUs, offloads processing/memory as possible to central processing units (CPUs) to allow more to fit on GPU and reduces memory fragmentation by ensuring that long- and short-term memory allocations do not mix. Overall, pretraining was achieved in approximately 3 days distributed across three nodes each with four Nvidia V100 32GB GPUs (total 12 GPUs).

Geneformer fine-tuning

Fine-tuning of Geneformer was accomplished by initializing the model with the pretrained Geneformer weights and adding a final task-specific transformer layer. The fine-tuning objective was either gene classification or cell classification as indicated in Supplementary Table 2. The trainer from the Huggingface Transformers library[159] was used for pretraining with the substitution of a custom tokenizer as described above and a custom data collator for dynamically labelling gene or cell classes as indicated in Supplementary Table 2. To demonstrate the efficacy of the pretrained Geneformer in boosting predictive potential of downstream fine-tuning applications, we intentionally used the same fine-tuning hyperparameters for all applications. It should be noted that hyperparameter tuning for deep learning applications generally significantly enhances learning and so it is likely that the maximum predictive potential of Geneformer in these downstream applications is significantly underestimated. Hyperparameters used for fine-tuning were as follows: max learning rate, 5×10^{-5}; learning scheduler, linear with warmup; optimizer, Adam with weight decay fix[160]; warmup steps, 500; weight decay, 0.001; batch size, 12. All fine-tuning in Supplementary Table 2 was performed with a single training epoch to avoid overfitting.

The number of layers frozen from fine-tuning are indicated in Supplementary Table 2. Generally, in our experience, applications that are more relevant to the pretraining objective benefit from more layers being frozen to prevent overfitting to the limited task-specific data, whereas applications that are more distant from the pretraining objective benefit from fine-tuning of more layers to optimize performance on the new task. Fine-tuning results for gene classification applications were reported as AUCs ± standard deviation and F1 score calculated on the basis of a fivefold cross-validation strategy for which training was performed on 80% of the gene training labels and performance was tested on the 20% held-out gene training labels, repeating for five folds. Of note, because the fine-tuning applications are trained on classification objectives that are completely separate from the masked learning objective, whether or not task-specific data were included in the pretraining corpus is not relevant to the classification predictions, as demonstrated in Extended Data Fig. 1f.

We then fully fine-tuned the dosage sensitivity and bivalency classification models using all gene training labels to test their ability to generalize to out-of-sample data. We tested whether, without any further training, the model fine-tuned to distinguish dosage-sensitive versus insensitive genes could predict dosage sensitivity of a recently reported set of disease genes from ref. 22, which analysed CNVs from 753,994 individuals to define genes whose deletion was associated with primarily neurodevelopmental disease with either high (greater than 0.85 score) or moderate (0.15–0.85 score) confidence[22]. Predicted dosage sensitivity of these gene sets was tested in the context of 10,000 randomly sampled cells from Genecorpus-30M, neurons across any adult or developmental timepoint defined as TUBB3-marked cells from Genecorpus-30M or fetal cerebral cells from the Fetal Cell Atlas[23]. We also tested whether, without any further training, the model fine-tuned to distinguish bivalent versus single Lys4-marked genes by training on the 56 highly conserved loci would generalize to the genome-wide setting[30].

Geneformer gene embeddings, cell embeddings and attention weights

Gene embeddings. For each single-cell transcriptome presented to Geneformer, the model embeds each gene into a 256-dimensional space that encodes the characteristics of the gene specific to the context of that cell. Contextual Geneformer gene embeddings are extracted as the hidden state weights for the 256 embedding dimensions for each gene within the given single-cell transcriptome evaluated by forward pass through the Geneformer model. Gene embeddings analysed in this study were extracted from the second to last layer of the models as the final layer is known to encompass features more directly related to the learning objective prediction whereas the second to last layer is a more generalizable representation.

Cell embeddings. Geneformer cell embeddings, which encode characteristics of the state of that single cell, are generated by averaging the embeddings of each gene detected in that cell, resulting in a 256-dimensional embedding. We used the second to last layer

embeddings as discussed above (except for the disease modelling application as discussed in the Supplementary Methods).

Attention weights. Each of Geneformer's six layers has four attention heads that are meant to learn in an unsupervised manner to pay attention to distinct classes of genes to jointly improve predictions without previous knowledge of the biological function of any gene. Contextual Geneformer attention weights are extracted for each attention head within each self-attention layer for each gene within the given single-cell transcriptome evaluated by forward pass through the Geneformer model.

In silico perturbation

We designed an in silico perturbation approach for which the rank of given genes is perturbed to model their inhibition or activation. The effects of the in silico perturbation are measured at the cell and gene embedding level, modelling how the perturbation affects the state of the cell and the regulation of downstream genes within the gene network, respectively. In silico deletion was modelled by removing the given gene from the rank value encoding of the given single-cell transcriptome and quantifying the cosine similarity between the original and perturbed (1) cell embeddings to determine the predicted deleterious impact of deleting that gene in that cell context or (2) gene embeddings of the remaining genes in the single-cell transcriptome to determine which genes were predicted to be most sensitive to in silico deletion of the given gene. In silico deletion can be performed with a single gene or multiple genes being deleted. In silico activation was modelled by moving a given gene(s) to the front of the rank value encoding (similarly to the in silico reprogramming strategy discussed in the Supplementary Methods in which genes were artificially added to the front of the rank value encoding to model cellular reprogramming by these factors). In theory, more subtle downregulation and activation could be modelled by shifting genes up or down within the rank value encoding to a subtler degree.

Please refer to the Supplementary Methods for complete methods including analysis of context dependence and robustness to batch-dependent technical artefacts, attention weight analysis, in silico perturbation analysis, alternative modelling approaches, cell-type annotation fine-tuning application, disease modelling approach, scRNA-seq sample collection and preprocessing and experimental testing of Geneformer-predicted targets in engineered cardiac microtissues.

Reporting summary

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

Data availability

Genecorpus-30M is available on the Huggingface Dataset Hub at https://huggingface.co/datasets/ctheodoris/Genecorpus-30M.

Code availability

The pretrained Geneformer model, transcriptome tokenizer and code for pretraining and fine-tuning the model are available on the Huggingface Model Hub at https://huggingface.co/ctheodoris/Geneformer. All other code used in this study is available upon request from the corresponding authors.

47. Smillie, C. S. et al. Intra- and inter-cellular rewiring of the human colon during ulcerative colitis. *Cell* **178**, 714–730 (2019).
48. Lee, J. S. et al. Immunophenotyping of Covid-19 and influenza highlights the role of type I interferons in development of severe Covid-19. *Sci. Immunol.* **5**, eabd1554 (2020).
49. Baron, M. et al. A single-cell transcriptomic map of the human and mouse pancreas reveals inter- and intra-cell population structure. *Cell Syst.* **3**, 346–360 (2016).
50. Fang, Z. et al. Single-cell heterogeneity analysis and CRISPR screen identify key β-cell-specific disease genes. *Cell Rep.* **26**, 3132–3144 (2019).
51. Agarwal, D. et al. A single-cell atlas of the human substantia nigra reveals cell-specific pathways associated with neurological disorders. *Nat. Commun.* **11**, 4183 (2020).
52. Rasouli, J. et al. A distinct GM-CSF+ T helper cell subset requires T-bet to adopt a TH1 phenotype and promote neuroinflammation. *Sci. Immunol.* **5**, eaba9953 (2020).
53. Park, J.-E. et al. A cell atlas of human thymic development defines T cell repertoire formation. *Science* **367**, eaay3224 (2020).
54. Mende, N. et al. Quantitative and molecular differences distinguish adult human medullary and extramedullary haematopoietic stem and progenitor cell landscapes. Preprint at *BioRxiv* https://doi.org/10.1101/2020.01.26.919753 (2020).
55. Setty, M. et al. Characterization of cell fate probabilities in single-cell data with Palantir. *Nat. Biotechnol.* **37**, 451–460 (2019).
56. Popescu, D.-M. et al. Decoding human fetal liver haematopoiesis. *Nature* **574**, 365–371 (2019).
57. Vento-Tormo, R. et al. Single-cell reconstruction of the early maternal-fetal interface in humans. *Nature* **563**, 347–353 (2018).
58. Ramachandran, P. et al. Resolving the fibrotic niche of human liver cirrhosis at single-cell level. *Nature* **575**, 512–518 (2019).
59. Kinchen, J. et al. Structural remodeling of the human colonic mesenchyme in inflammatory bowel disease. *Cell* **175**, 372–386 (2018).
60. James, K. R. et al. Distinct microbial and immune niches of the human colon. *Nat. Immunol.* **21**, 343–353 (2020).
61. Zhou, L. et al. Single-cell RNA-seq analysis uncovers distinct functional human NKT cell sub-populations in peripheral blood. *Front. Cell Dev. Biol.* **8**, 384 (2020).
62. Liao, J. et al. Single-cell RNA sequencing of human kidney. *Sci. Data* **7**, 4 (2020).
63. Jäkel, S. et al. Altered human oligodendrocyte heterogeneity in multiple sclerosis. *Nature* **566**, 543–547 (2019).
64. Merrick, D. et al. Identification of a mesenchymal progenitor cell hierarchy in adipose tissue. *Science* **364**, eaav2501 (2019).
65. Habermann, A. C. et al. Single-cell RNA sequencing reveals profibrotic roles of distinct epithelial and mesenchymal lineages in pulmonary fibrosis. *Sci. Adv.* **6**, eaba1972 (2020).
66. Rosa, F. F. et al. Direct reprogramming of fibroblasts into antigen-presenting dendritic cells. *Sci. Immunol.* **3**, eaau4292 (2018).
67. Stewart, B. J. et al. Spatiotemporal immune zonation of the human kidney. *Science* **365**, 1461–1466 (2019).
68. MacParland, S. A. et al. Single cell RNA sequencing of human liver reveals distinct intrahepatic macrophage populations. *Nat. Commun.* **9**, 4383 (2018).
69. Welch, J. et al. Integrative inference of brain cell similarities and differences from single-cell genomics. Preprint at *BioRxiv* https://doi.org/10.1101/459891 (2018).
70. Ledergor, G. et al. Single cell dissection of plasma cell heterogeneity in symptomatic and asymptomatic myeloma. *Nat. Med.* **24**, 1867–1876 (2018).
71. Lukowski, S. W. et al. A single-cell transcriptome atlas of the adult human retina. *EMBO J.* **38**, e100811 (2019).
72. Kang, H. M. et al. Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. *Nat. Biotechnol.* **36**, 89–94 (2018).
73. Zirkel, A. et al. HMGB2 loss upon senescence entry disrupts genomic organization and induces CTCF clustering across cell types. *Mol. Cell* **70**, 730–744 (2018).
74. Goudot, C. et al. Aryl hydrocarbon receptor controls monocyte differentiation into dendritic cells versus macrophages. *Immunity* **47**, 582–596 (2017).
75. McCauley, K. B. et al. Single-cell transcriptomic profiling of pluripotent stem cell-derived SCGB3A2+ airway epithelium. *Stem Cell Rep.* **10**, 1579–1595 (2018).
76. Das, R. et al. Early B cell changes predict autoimmunity following combination immune checkpoint blockade. *J. Clin. Invest.* **128**, 715–720 (2018).
77. Kini Bailur, J. et al. Changes in bone marrow innate lymphoid cell subsets in monoclonal gammopathy: target for IMiD therapy. *Blood Adv.* **1**, 2343–2347 (2017).
78. Patil, V. S. et al. Precursors of human CD4+ cytotoxic T lymphocytes identified by single-cell transcriptome analysis. *Sci. Immunol.* **3**, eaan8664 (2018).
79. Wang, C. et al. Expansion of hedgehog disrupts mesenchymal identity and induces emphysema phenotype. *J. Clin. Invest.* **128**, 4343–4358 (2018).
80. Hermann, B. P. et al. The mammalian spermatogenesis single-cell transcriptome, from spermatogonial stem cells to spermatids. *Cell Rep.* **25**, 1650–1667 (2018).
81. Menon, R. et al. Single-cell analysis of progenitor cell dynamics and lineage specification in the human fetal kidney. *Development* **145**, dev164038 (2018).
82. Czerniecki, S. M. et al. High-throughput screening enhances kidney organoid differentiation from human pluripotent stem cells and enables automated multidimensional phenotyping. *Cell Stem Cell* **22**, 929–940 (2018).
83. Papa, L. et al. Ex vivo human HSC expansion requires coordination of cellular reprogramming with mitochondrial remodeling and p53 activation. *Blood Adv.* **2**, 2766–2779 (2018).
84. Schulthess, J. et al. The short chain fatty acid butyrate imprints an antimicrobial program in macrophages. *Immunity* **50**, 432–445 (2019).
85. Guo, J. et al. The adult human testis transcriptional cell atlas. *Cell Res.* **28**, 1141–1157 (2018).
86. Karow, M. et al. Direct pericyte-to-neuron reprogramming via unfolding of a neural stem cell-like program. *Nat. Neurosci.* **21**, 932–940 (2018).
87. Xin, Y. et al. Pseudotime ordering of single human β-cells reveals states of insulin production and unfolded protein response. *Diabetes* **67**, 1783–1794 (2018).
88. Phipson, B. et al. Evaluation of variability in human kidney organoids. *Nat. Methods* **16**, 79–87 (2019).
89. Balan, S. et al. Large-scale human dendritic cell differentiation revealing notch-dependent lineage bifurcation and heterogeneity. *Cell Rep.* **24**, 1902–1915 (2018).
90. Milpied, P. et al. Human germinal center transcriptional programs are de-synchronized in B cell lymphoma. *Nat. Immunol.* **19**, 1013–1024 (2018).
91. Parikh, K. et al. Colonic epithelial cell diversity in health and inflammatory bowel disease. *Nature* **567**, 49–55 (2019).

92. Habiel, D. M. et al. CCR10+ epithelial cells from idiopathic pulmonary fibrosis lungs drive remodeling. *JCI Insight* **3**, e122211 (2018).
93. Paik, D. T. et al. Large-scale single-cell RNA-seq reveals molecular signatures of heterogeneous populations of human induced pluripotent stem cell-derived endothelial cells. *Circ. Res.* **123**, 443–450 (2018).
94. Martin, J. C. et al. Single-cell analysis of Crohn's disease lesions identifies a pathogenic cellular module associated with resistance to anti-TNF therapy. *Cell* **178**, 1493–1508 (2019).
95. Zheng, Y. et al. A human circulating immune cell landscape in aging and COVID-19. *Protein Cell* **11**, 740–770 (2020).
96. Hochane, M. et al. Single-cell transcriptomics reveals gene expression dynamics of human fetal kidney development. *PLoS Biol.* **17**, e3000152 (2019).
97. Sohni, A. et al. The neonatal and adult human testis defined at the single-cell level. *Cell Rep.* **26**, 1501–1517 (2019).
98. Tran, T. et al. In vivo developmental trajectories of human podocyte inform in vitro differentiation of pluripotent stem cell-derived podocytes. *Dev. Cell* **50**, 102–116 (2019).
99. Wang, Y. et al. Single-cell transcriptome analysis reveals differential nutrient absorption functions in human intestine. *J. Exp. Med.* **217**, e20191130 (2020).
100. Vieira Braga, F. A. et al. A cellular census of human lungs identifies novel cell states in health and in asthma. *Nat. Med.* **25**, 1153–1163 (2019).
101. Guo, J. et al. The dynamic transcriptional cell atlas of testis development during human puberty. *Cell Stem Cell* **26**, 262–276 (2020).
102. Voigt, A. P. et al. Single-cell transcriptomics of the human retinal pigment epithelium and choroid in health and macular degeneration. *Proc. Natl Acad. Sci. USA* **116**, 24100–24107 (2019).
103. Menon, M. et al. Single-cell transcriptomic atlas of the human retina identifies cell types associated with age-related macular degeneration. *Nat. Commun.* **10**, 4902 (2019).
104. Wilk, A. J. et al. A single-cell atlas of the peripheral immune response in patients with severe COVID-19. *Nat. Med.* **26**, 1070–1076 (2020).
105. Li, B. et al. Cumulus provides cloud-based data analysis for large-scale single-cell and single-nucleus RNA-seq. *Nat. Methods* **17**, 793–798 (2020).
106. Daniszewski, M. et al. Single cell RNA sequencing of stem cell-derived retinal ganglion cells. *Sci. Data* **5**, 180013 (2018).
107. Goveia, J. et al. An integrated gene expression landscape profiling approach to identify lung tumor endothelial cell heterogeneity and angiogenic candidates. *Cancer Cell* **37**, 21–36 (2020).
108. Norelli, M. et al. Monocyte-derived IL-1 and IL-6 are differentially required for cytokine-release syndrome and neurotoxicity due to CAR T cells. *Nat. Med.* **24**, 739–748 (2018).
109. Daniszewski, M. et al. Single-cell profiling identifies key pathways expressed by iPSCs cultured in different commercial media. *iScience* **7**, 30–39 (2018).
110. Miller, A. J. et al. In vitro and in vivo development of the human airway at single-cell resolution. *Dev. Cell* **53**, 117–128 (2020).
111. Silvin, A. et al. Elevated calprotectin and abnormal myeloid cell subsets discriminate severe from mild COVID-19. *Cell* **182**, 1401–1418 (2020).
112. Deprez, M. et al. A single-cell atlas of the human healthy airways. *Am. J. Resp. Crit. Care Med.* **202**, 1636–1645 (2020).
113. Sridhar, A. et al. Single-cell transcriptomic comparison of human fetal retina, hPSC-derived retinal organoids, and long-term retinal cultures. *Cell Rep.* **30**, 1644–1659 (2020).
114. Wu, H. et al. Comparative analysis and refinement of human PSC-derived kidney organoid differentiation with single-cell transcriptomics. *Cell Stem Cell* **23**, 869–881 (2018).
115. Vijay, J. et al. Single-cell analysis of human adipose tissue identifies depot and disease specific cell types. *Nat. Metab.* **2**, 97–109 (2020).
116. Solé-Boldo, L. et al. Single-cell transcriptomes of the human skin reveal age-related loss of fibroblast priming. *Commun. Biol.* **3**, 188 (2020).
117. Adams, T. S. et al. Single-cell RNA-seq reveals ectopic and aberrant lung-resident cell populations in idiopathic pulmonary fibrosis. *Sci. Adv.* **6**, eaba1983 (2020).
118. Moreira, L. M. et al. Paracrine signalling by cardiac calcitonin controls atrial fibrogenesis and arrhythmia. *Nature* **587**, 460–465 (2020).
119. Ren, X. et al. COVID-19 immune features revealed by a large-scale single-cell transcriptome atlas. *Cell* **184**, 1895–1913 (2021).
120. Bunis, D. G. et al. Single-cell mapping of progressive fetal-to-adult transition in human naive T cells. *Cell Rep.* **34**, 108573 (2021).
121. Plasschaert, L. W. et al. A single-cell atlas of the airway epithelium reveals the CFTR-rich pulmonary ionocyte. *Nature* **560**, 377–381 (2018).
122. Takeda, A. et al. Single-cell survey of human lymphatics unveils marked endothelial cell heterogeneity and mechanisms of homing for neutrophils. *Immunity* **51**, 561–572 (2019).
123. Frumm, S. M. et al. A hierarchy of proliferative and migratory keratinocytes maintains the tympanic membrane. *Cell Stem Cell* **28**, 315–330 (2021).
124. Yu, Z. et al. Single-cell transcriptomic map of the human and mouse bladders. *J. Am. Soc. Nephrol.* **30**, 2159–2176 (2019).
125. Rubenstein, A. B. et al. Single-cell transcriptional profiles in human skeletal muscle. *Sci. Rep.* **10**, 229 (2020).
126. McCracken, I. R. et al. Transcriptional dynamics of pluripotent stem cell-derived endothelial cell differentiation revealed by single-cell RNA sequencing. *Eur. Heart J.* **41**, 1024–1036 (2020).
127. Hua, P. et al. Single-cell analysis of bone marrow-derived CD34+ cells from children with sickle cell disease and thalassemia. *Blood* **134**, 2111–2115 (2019).
128. Orozco, L. D. et al. Integration of eQTL and a single-cell atlas in the human eye identifies causal genes for age-related macular degeneration. *Cell Rep.* **30**, 1246–1259 (2020).
129. Hurley, K. et al. Reconstructed single-cell fate trajectories define lineage plasticity windows during differentiation of human PSC-derived distal lung progenitors. *Cell Stem Cell* **26**, 593–608 (2020).
130. Schafflick, D. et al. Integrated single cell analysis of blood and cerebrospinal fluid leukocytes in multiple sclerosis. *Nat. Commun.* **11**, 247 (2020).
131. Su, C. et al. Single-cell RNA sequencing in multiple pathologic types of renal cell carcinoma revealed novel potential tumor-specific markers. *Front. Oncol.* **11**, 719564 (2021).
132. He, J. et al. Dissecting human embryonic skeletal stem cell ontogeny by single-cell transcriptomic and functional analyses. *Cell Res.* **31**, 742–757 (2021).
133. Liao, M. et al. Single-cell landscape of bronchoalveolar immune cells in patients with COVID-19. *Nat. Med.* **26**, 842–844 (2020).
134. Liu, X. et al. Reprogramming roadmap reveals route to human induced trophoblast stem cells. *Nature* **586**, 101–107 (2020).
135. He, S. et al. Single-cell transcriptome profiling of an adult human cell atlas of 15 major organs. *Genome Biol.* **21**, 294 (2020).
136. Wu, C.-L. et al. Single cell transcriptomic analysis of human pluripotent stem cell chondrogenesis. *Nat. Commun.* **12**, 362 (2021).
137. Cowan, C. S. et al. Cell types of the human retina and its organoids at single-cell resolution. *Cell* **182**, 1623–1640 (2020).
138. Savas, P. et al. Single-cell profiling of breast cancer T cells reveals a tissue-resident memory subset associated with improved prognosis. *Nat. Med.* **24**, 986–993 (2018).
139. Wang, L. et al. Single-cell map of diverse immune phenotypes in the metastatic brain tumor microenvironment of non small cell lung cancer. Preprint at *BioRxiv* https://doi.org/10.1101/2019.12.30.890517 (2019).
140. Lu, Y.-C. et al. Single-cell transcriptome analysis reveals gene signatures associated with T-cell persistence following adoptive cell therapy. *Cancer Immunol. Res.* **7**, 1824–1836 (2019).
141. Wang, L. et al. The phenotypes of proliferating glioblastoma cells reside on a single axis of variation. *Cancer Discov.* **9**, 1708–1719 (2019).
142. Wang, R. et al. Adult human glioblastomas harbor radial glia-like cells. *Stem Cell Rep.* **14**, 338–350 (2020).
143. Wang, L., Catalan, F., Shamardani, K., Babikir, H. & Diaz, A. Ensemble learning for classifying single-cell data and projection across reference atlases. *Bioinformatics* **36**, 3585–3587 (2020).
144. Ruffin, A. T. et al. B cell signatures and tertiary lymphoid structures contribute to outcome in head and neck squamous cell carcinoma. *Nat. Commun.* **12**, 3349 (2021).
145. Zhang, Q. et al. Landscape and dynamics of single immune cells in hepatocellular carcinoma. *Cell* **179**, 829–845 (2019).
146. Song, Q. et al. Dissecting intratumoral myeloid cell plasticity by single cell RNA-seq. *Cancer Med.* **8**, 3072–3085 (2019).
147. Kim, N. et al. Single-cell RNA sequencing demonstrates the molecular and cellular reprogramming of metastatic lung adenocarcinoma. *Nat. Commun.* **11**, 2285 (2020).
148. Tang-Huau, T.-L. et al. Human in vivo-generated monocyte-derived dendritic cells and macrophages cross-present antigens through a vacuolar pathway. *Nat. Commun.* **9**, 2570 (2018).
149. Peng, J. et al. Single-cell RNA-seq highlights intra-tumoral heterogeneity and malignant progression in pancreatic ductal adenocarcinoma. *Cell Res.* **29**, 725–738 (2019).
150. 10x Genomics Datasets: Single Cell Gene Expression. *10x Genomics* https://www.10xgenomics.com/resources/datasets?menu%5Bproducts.name%5D=Single%20Cell%20Gene%20Expression&query=&page=1&configure%5Bfacets%5D%5B0%5D=chemistryVersionAndThroughput&configure%5Bfacets%5D%5B1%5D=pipeline.version&configure%5BhitsPerPage%5D=500.
151. de Andrade, L. F. et al. Discovery of specialized NK cell populations infiltrating human melanoma metastases. *JCI Insight* **4**, e133103 (2019).
152. Zhang, P. et al. Dissecting the single-cell transcriptome network underlying gastric premalignant lesions and early gastric cancer. *Cell Rep.* **27**, 1934–1947 (2019).
153. Durante, M. A. et al. Single-cell analysis reveals new evolutionary complexity in uveal melanoma. *Nat. Commun.* **11**, 496 (2020).
154. Svensson, V., da Veiga Beltrame, E. & Pachter, L. A curated database reveals trends in single-cell transcriptomics. *Database* **2020**, baaa073 (2020).
155. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.* **19**, 15 (2018).
156. Xin, J. et al. High-performance web services for querying gene and variant annotation. *Genome Biol.* **17**, 91 (2016).
157. Dunning, T. The t-digest: efficient estimates of distributions. *Softw. Impacts* **7**, 100049 (2021).
158. Lhoest, Q. et al. Datasets: a community library for natural language processing. Preprint at https://doi.org/10.48550/arXiv.2109.02846 (2021).
159. Wolf, T. et al. HuggingFace's transformers: state-of-the-art natural language processing. Preprint at https://doi.org/10.48550/arXiv.1910.03771 (2019).
160. Loshchilov, I. & Hutter, F. Decoupled weight decay regularization. Preprint at https://doi.org/10.48550/arXiv.1711.05101 (2017).

Acknowledgements We thank J. Rae for helpful scientific discussions and Google Research for providing tensor processing unit (TPU) resources for experimentation. P.T.E. was supported by grants from the National Institutes of Health (NIH) (1RO1HL092577, 1R01HL157635 and 5R01HL139731), American Heart Association Strategically Focused Research Networks (18SFRN34110082) and European Union (MAESTRIA 965286). C.V.T. was supported by NIH T32GM007748 and the Helen Hay Whitney Foundation Postdoctoral Fellowship. L.X. was supported by the American Heart Association (20CDA35260081).

Author contributions C.V.T. conceived of the work, developed Geneformer, assembled Genecorpus-30M and designed and performed computational analyses. L.X., A.C., Z.R.A.S., M.C.H., H.M. and E.M.B. performed experimental validation in engineered cardiac microtissues.

M.D.C. performed preprocessing, cell annotation and differential expression analysis of the cardiomyopathy dataset. Z.Z. provided data from the TISCH database for inclusion in Genecorpus-30M. X.S.L. and P.T.E. designed analyses and supervised the work. C.V.T., X.S.L. and P.T.E. wrote the manuscript. All authors edited the manuscript.

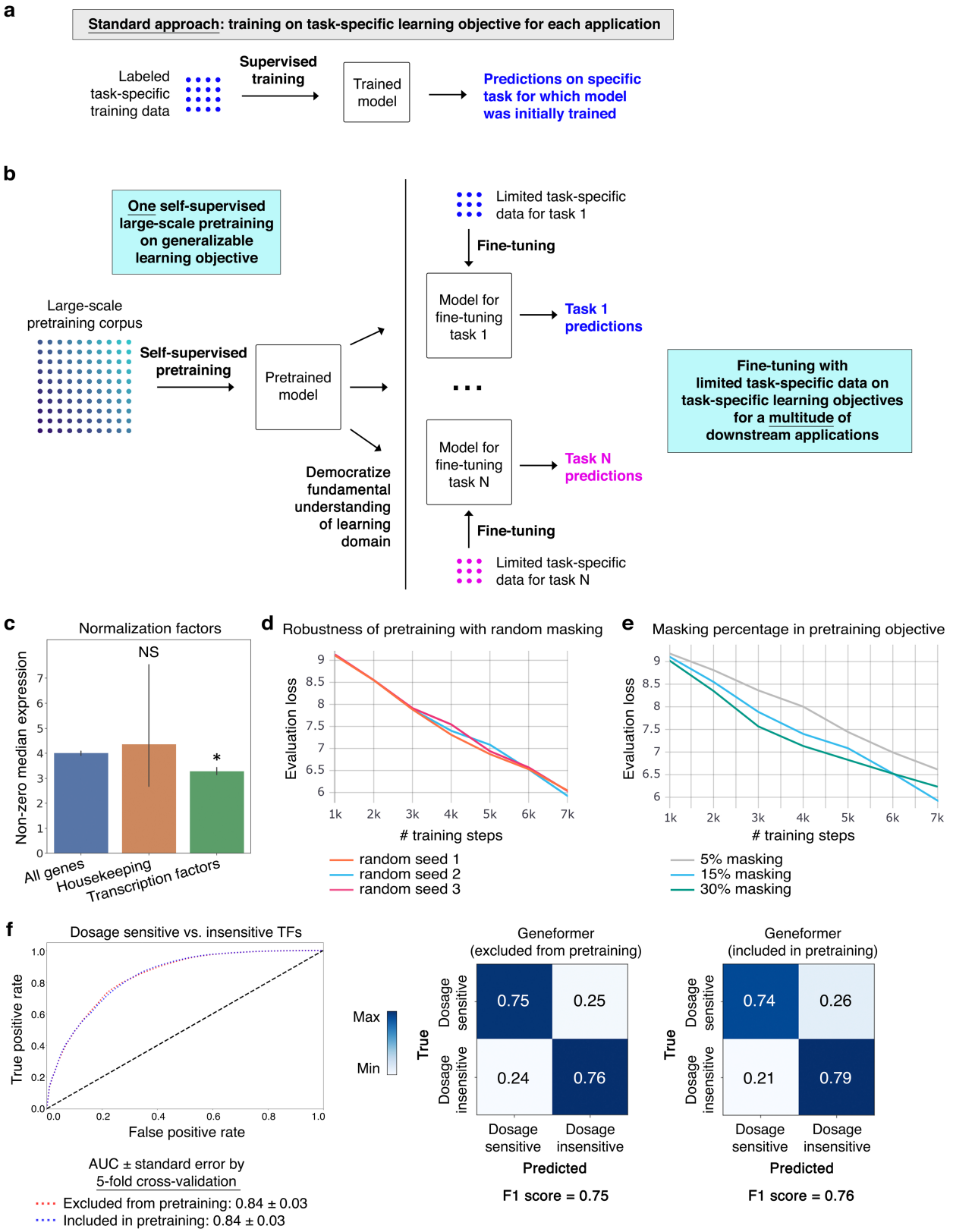
Competing interests X.S.L. conducted this work while on faculty at Dana-Farber Cancer Institute and is now a board member and CEO of GV20 Therapeutics. P.T.E. has received sponsored research support from Bayer AG, IBM Research, Bristol Myers Squibb and Pfizer. P.T.E. has also served on advisory boards or consulted for Bayer AG, MyoKardia and Novartis. A.C. is an employee of Bayer US LLC (a subsidiary of Bayer AG) and may own stock in Bayer AG. E.M.B. was a full-time employee of Bayer when this work was performed. The remaining authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41586-023-06139-9.

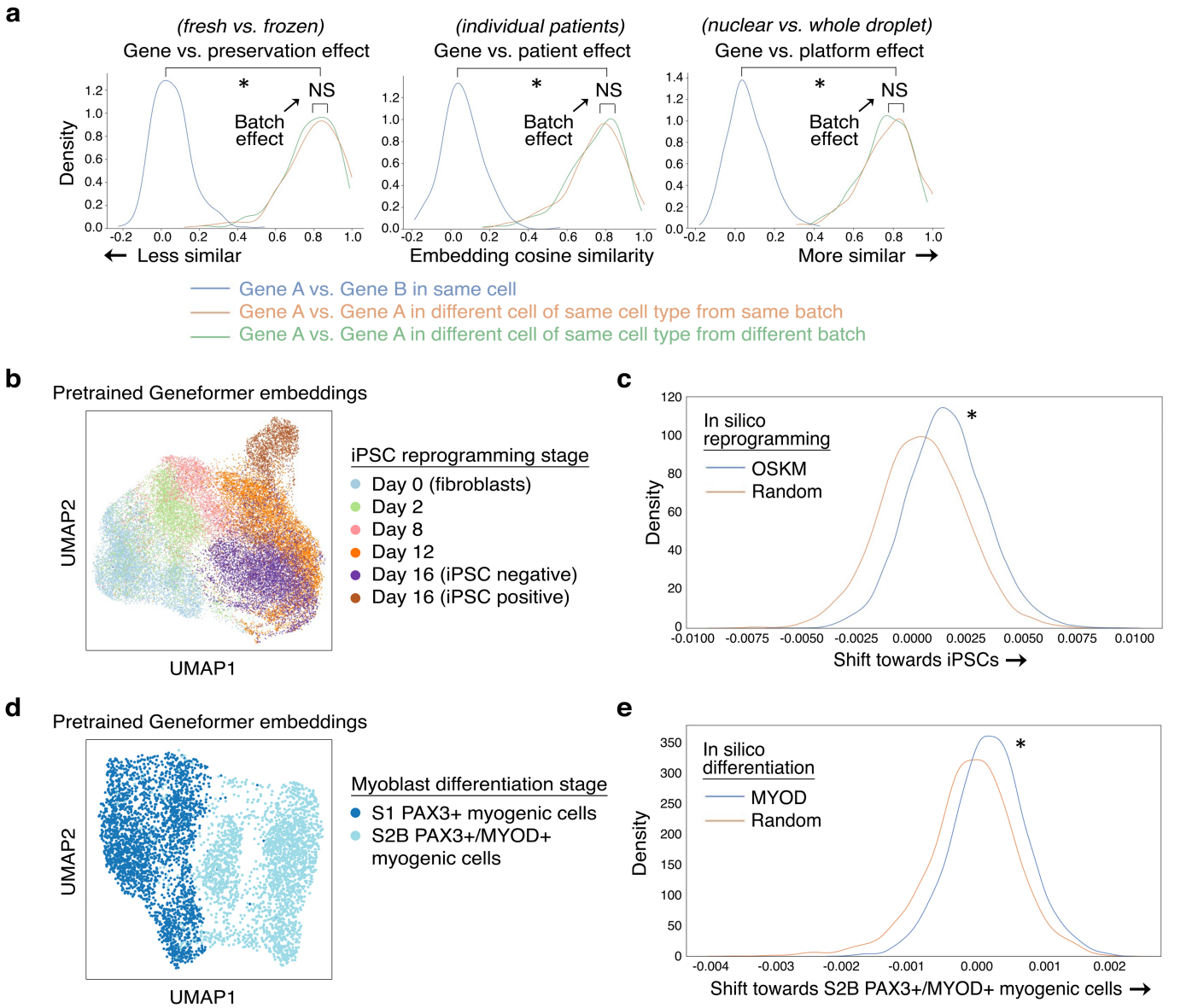
Correspondence and requests for materials should be addressed to Christina V. Theodoris or Patrick T. Ellinor.

Peer review information *Nature* thanks Amir Bashan, Natasa Przulj and Nathan Palpant for their contribution to the peer review of this work. Peer reviewer reports are available.

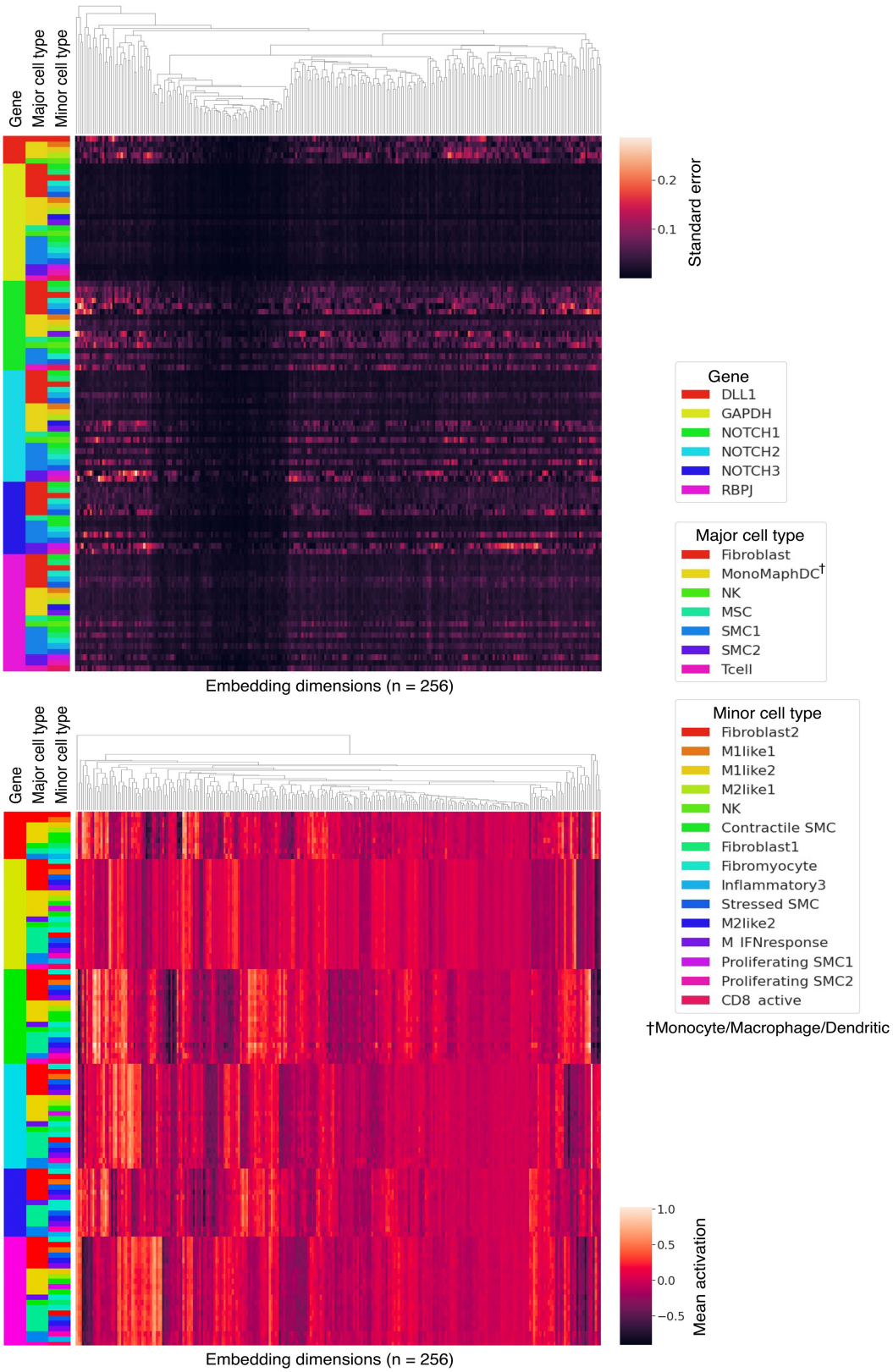


Extended Data Fig. 1 | See next page for caption.

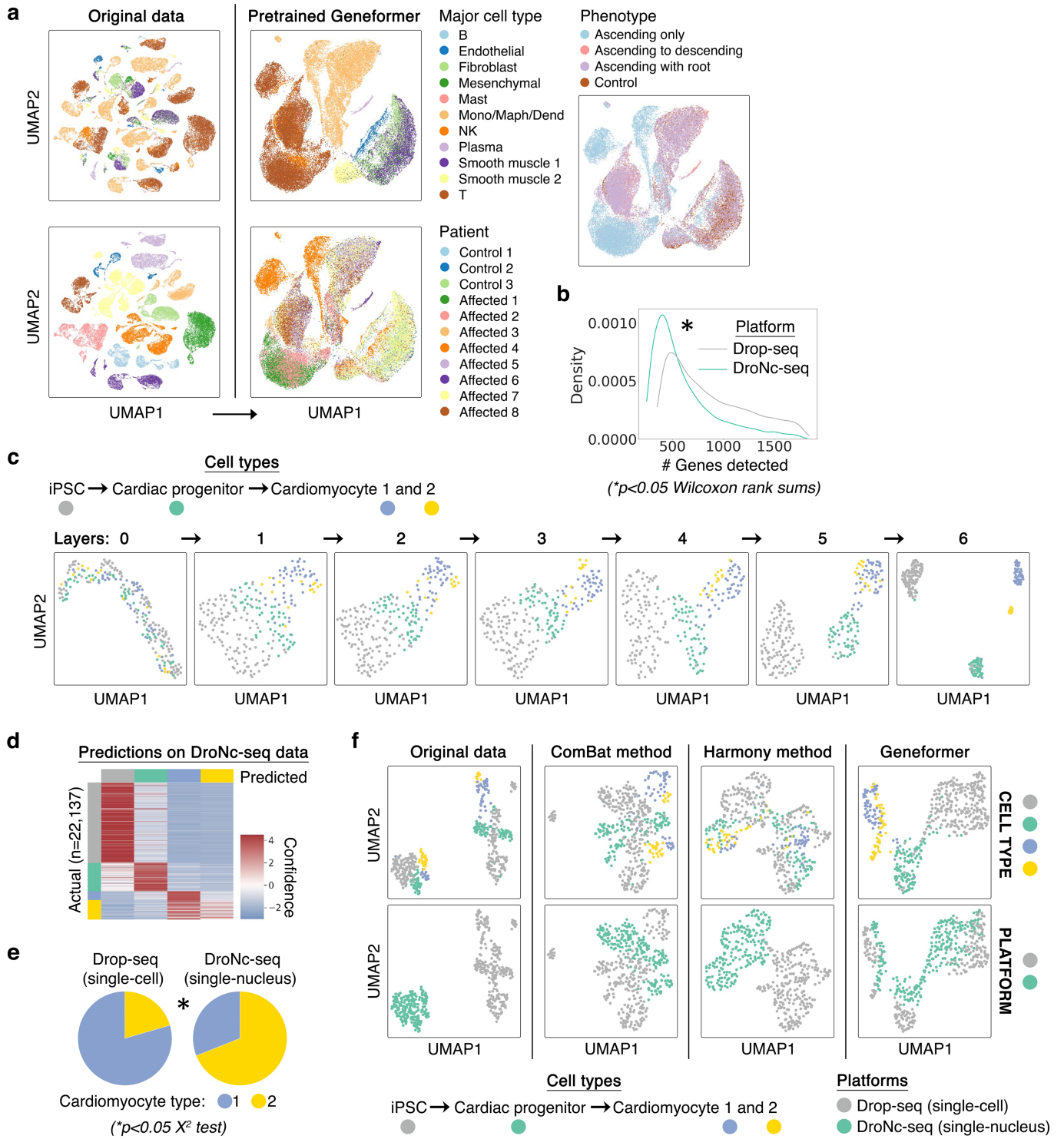
Extended Data Fig. 1 | Geneformer transfer learning strategy. a, Schematic of standard modelling approach, which necessitates retraining a new model from scratch for each new task. **b**, Schematic of transfer learning strategy. Through a single initial self-supervised large-scale pretraining on a generalizable learning objective, the model gains fundamental knowledge of the learning domain that is then democratized to a multitude of downstream applications distinct from the pretraining learning objective, transferring knowledge to new tasks. **c**, Transcription factors are normalized by a statistically significantly lower factor (resulting in higher prioritization in the rank value encoding) compared to all genes. Housekeeping genes on average show a trend of a higher normalization factor (resulting in deprioritization in the rank value encoding) compared to all genes (*$p < 0.05$ by Wilcoxon, FDR-corrected; all genes n = 17,903, housekeeping genes n = 11, transcription factors n = 1,384; error bars = standard deviation). **d**, Pretraining was performed with a randomly subsampled corpus of 100,000 cells, holding out 10,000 cells for evaluation, with 3 different random seeds. Evaluation loss was essentially equivalent in the 3 trials, indicating robustness to the set of genes randomly masked for each cell during the pretraining. **e**, Pretraining was performed with a randomly subsampled corpus of 100,000 cells, holding out 10,000 cells for evaluation, with 3 different masking percentages. 15% masking had marginally lower evaluation loss compared to 5% or 30% masking. **f**, Pretraining was performed with a randomly subsampled corpus of 90,000 cells and the model was then fine-tuned to distinguish dosage-sensitive vs. -insensitive transcription factors using 10,000 cells that were either included in or excluded from the 90,000 cell pretraining corpus. Predictive potential on the downstream fine-tuning task was measured by fivefold cross-validation with these 10,000 cells, demonstrating essentially equivalent results by AUC, confusion matrices, and F1 score. Because the fine-tuning applications are trained on classification objectives that are completely separate from the masked learning objective, whether or not task-specific data was included in the pretraining corpus is not relevant to the downstream classification predictions.



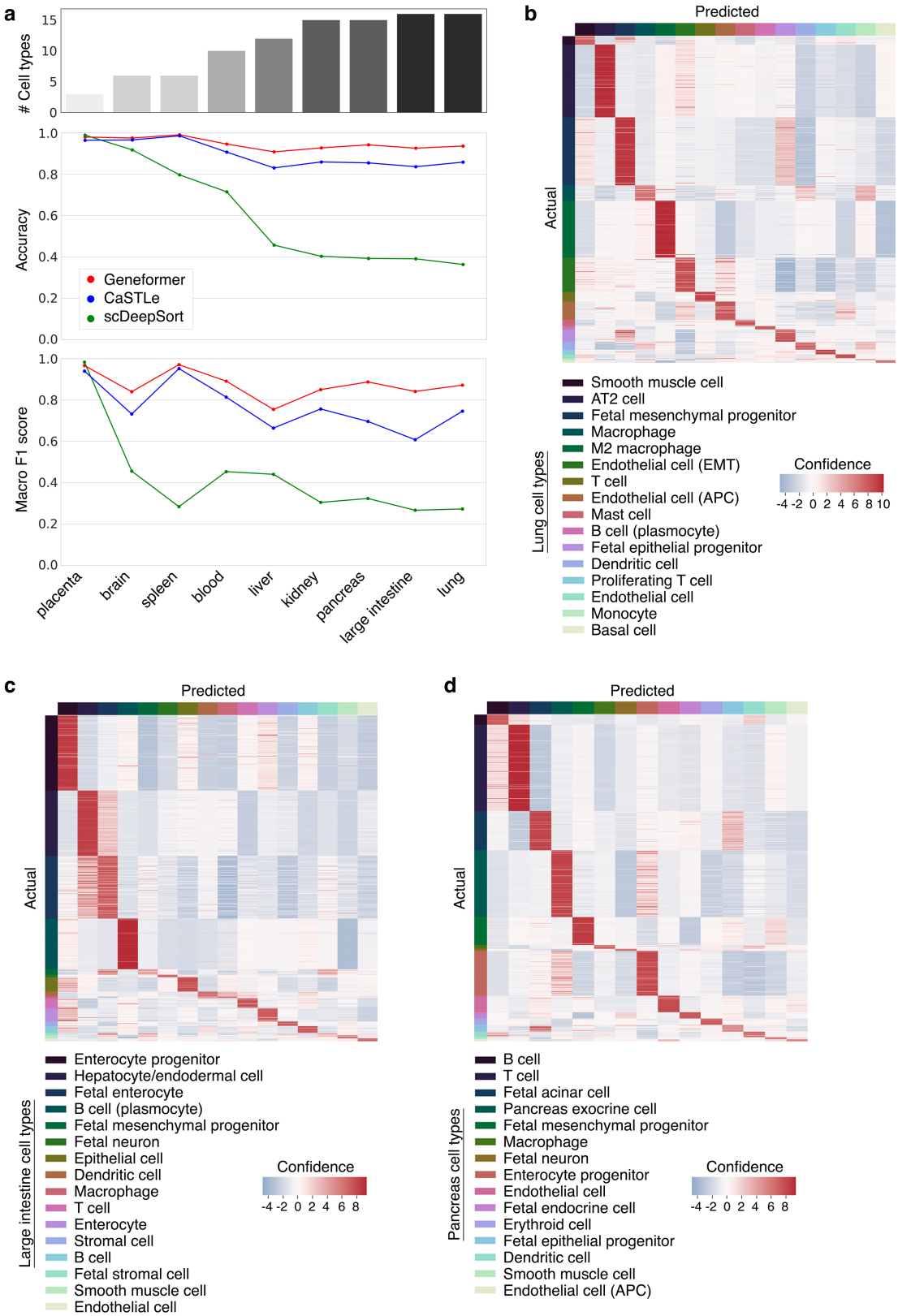
Extended Data Fig. 2 | Geneformer was context-aware and robust to batch-dependent technical artefacts. a, Effect of gene versus the indicated batch-dependent technical artefact on pretrained Geneformer gene embeddings (*p < 0.05 by Wilcoxon, FDR-corrected; NS: non-significant). We found that the gene embeddings were robust to sequencing platform[11], preservation method[12,13], and individual patient variability[14]. **b**, UMAP of pretrained Geneformer cell embeddings of cells undergoing iPSC reprogramming appropriately captured temporal trajectory of reprogramming (cell types as annotated by original study[15]; iPSC negative or positive refers to expression of marker TRA-1-60). Cell embeddings suggested that cells which do not progress to the iPSC state bifurcate into an alternative fate compared to cells that progress to the iPSC state after the day 12 stage. **c**, Compared to in silico reprogramming with random genes, in silico reprogramming of fibroblasts by artificially adding *OCT4*, *SOX2*, *KLF4*, and *MYC* (*OSKM*) to the front of their rank value encodings significantly shifted the gene embeddings from their initial fibroblast state to the embedding of that gene in the iPSC state (*p < 0.05 by Wilcoxon). **d**, UMAP of pretrained Geneformer cell embeddings of cells undergoing iPSC to myoblast differentiation at the earlier S1 (PAX3+) and later S2B (PAX3+/MYOD+) stages (cell types as annotated by original study[16]). **e**, Compared to in silico differentiation with random genes, in silico differentiation of the early-stage myogenic cells by artificially adding *MYOD* to the front of their rank value encodings significantly shifted the gene embeddings from their earlier state to the embedding of that gene in the later MYOD+ myogenic state (*p < 0.05 by Wilcoxon).



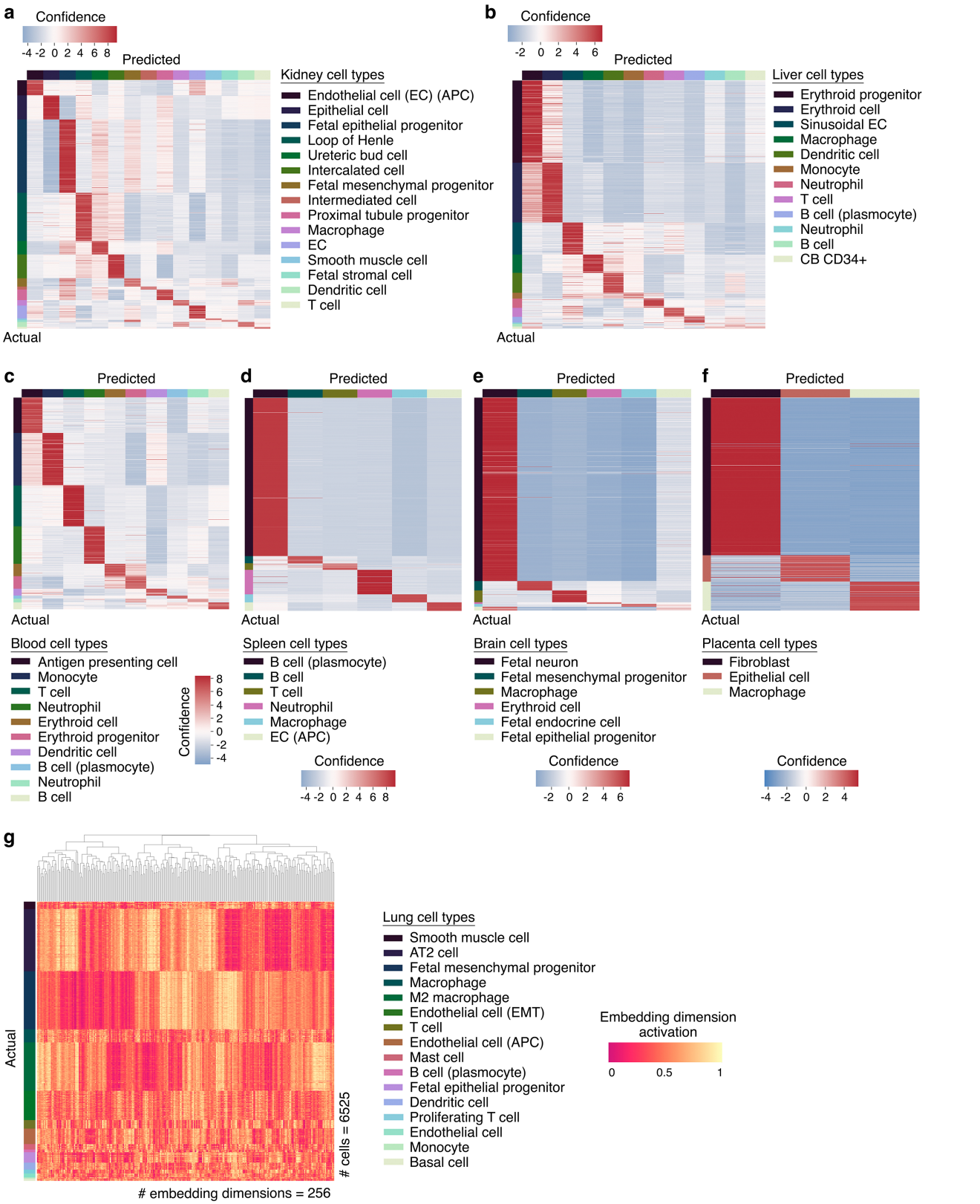
Extended Data Fig. 3 | Geneformer encoded context-specificity of key NOTCH pathway genes. Known context-dependent *NOTCH* genes showed higher variance in their contextual embeddings across variable aortic cell types compared to housekeeping gene *GAPDH*.



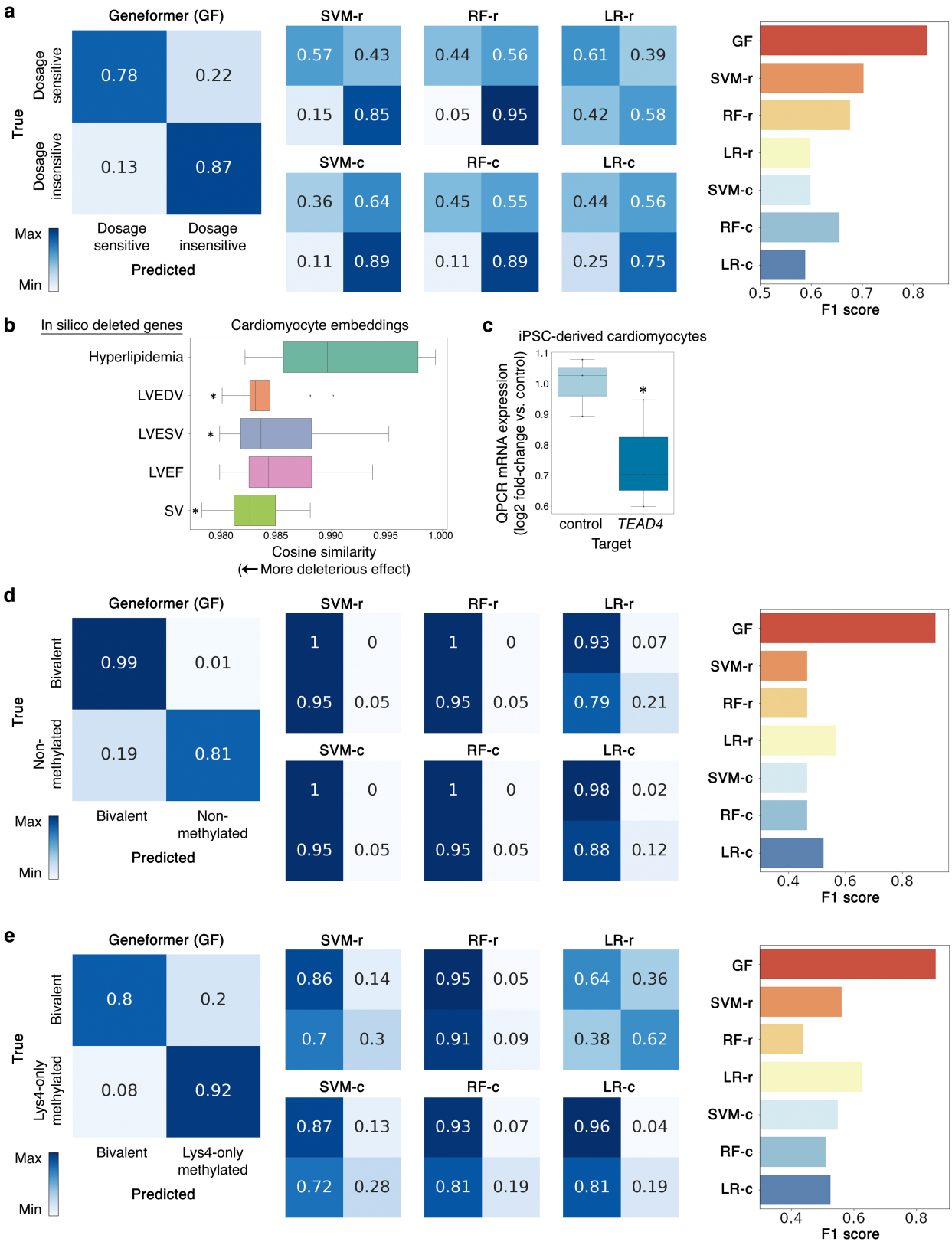
Extended Data Fig. 4 | Geneformer pretrained and fine-tuned cell embeddings were robust to batch-dependent technical artefacts. a, While original data (left) was highly affected by patient batch effect, cell embeddings generated by pretrained Geneformer (right) (without fine-tuning) clustered primarily by cell type and phenotype. Of note, affected individuals 1, 2, and 4 had the phenotype of ascending only aortic aneurysm, which is a different phenotype than aortic aneurysm that includes the root. **b**, Imbalance in the number of genes detected in each of the two platforms (single-cell Drop-seq versus single-nucleus DroNc-seq), which may result in batch-dependent technical artefacts. **c**, Cell embeddings from each layer of the Geneformer model fine-tuned to distinguish the indicated cell types (as annotated by original study[11]) using only the Drop-seq data. As the cells pass through each layer, the model successively extrudes them from each other to derive separable embeddings that distinguish the cell types. **d**, Cell type predictions on the DroNc-seq data by the model fine-tuned only on the Drop-seq data (out of sample accuracy 84%). Of note, inaccurate predictions were predominantly in predicting that cardiomyocyte type 2 was type 1, as expected given the minimal examples of cardiomyocyte type 2 in the Drop-seq data. **e**, The imbalance of cardiomyocyte type 1 and 2 between the platforms also suggests that these cellular subtypes may be an artefact of variable gene detection between the two platforms. **f**, Geneformer fine-tuned with only Drop-seq data automatically integrated DroNc-seq data such that the fine-tuned Geneformer cell embeddings primarily clustered by cell types and showed improved integration of platforms compared to the original data even after batch effect removal using the ComBat[17] or Harmony[18] methods.



Extended Data Fig. 5 | Geneformer boosted predictions in multiclass cell type annotation. a, Predictive potential (as measured by accuracy and macro F1 score) of Geneformer fine-tuned for cell type annotation in the indicated human tissues as compared to XGBoost (CaSTLe) and deep neural network-based (scDeepSort) methods. The top bar graph indicates the number of cell type classes for each tissue; the gap in performance of Geneformer compared to alternatives increased as the number of cell type classes increased, indicating that Geneformer was robust in even increasingly complex multiclass prediction applications. **b**, Lung, **c**, large intestine, or **d**, pancreas out of sample predictions by Geneformer fine-tuned to distinguish cell types in each tissue (training on 80% of cells, predictions on held-out 20% of cells).

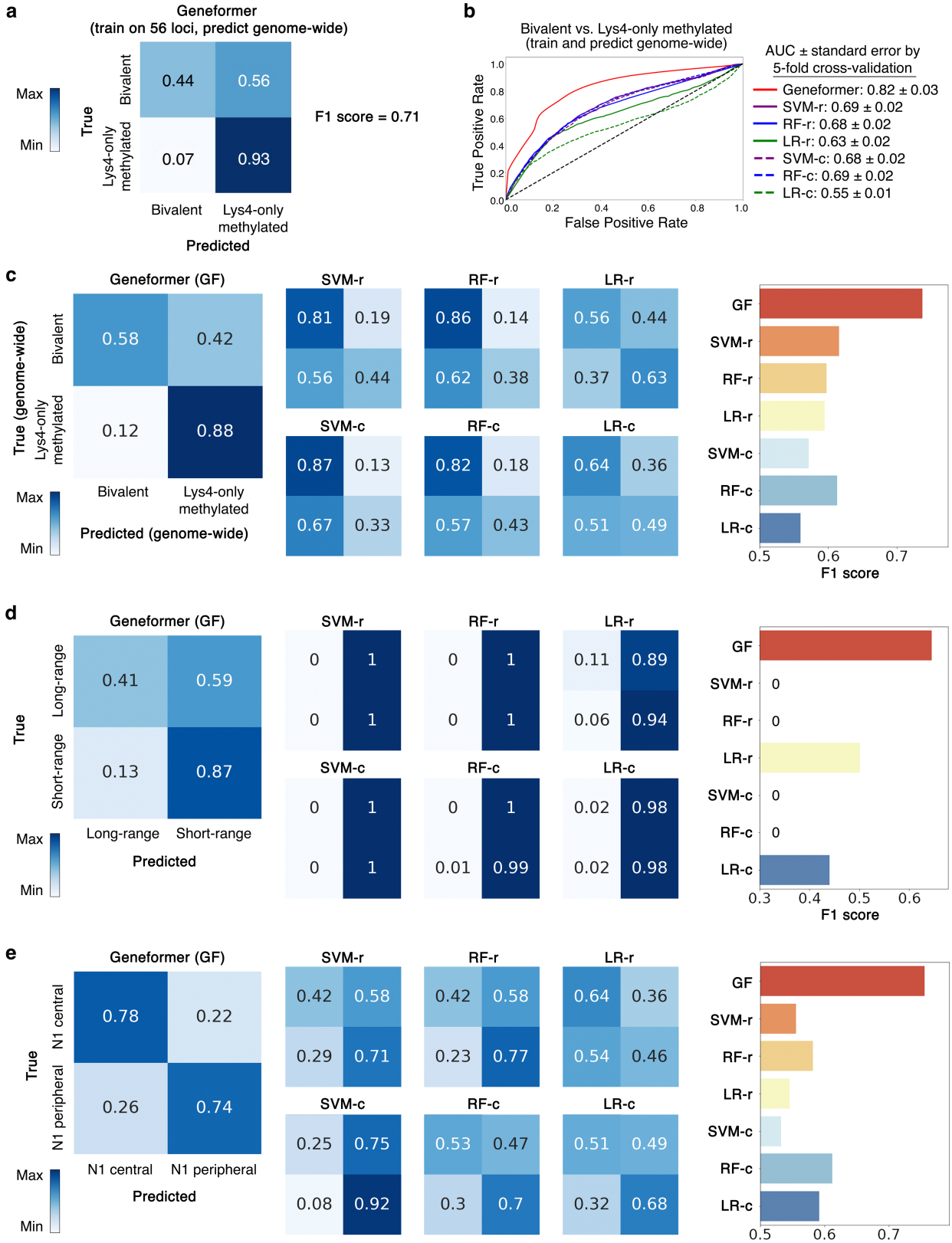


Extended Data Fig. 6 | Embedding dimension activations distinguish cell types in fine-tuned Geneformer model. a, Kidney, **b**, liver, **c**, blood, **d**, spleen, **e**, brain, or **f**, placenta out of sample predictions by Geneformer fine-tuned to distinguish cell types in each tissue (training on 80% of cells, predictions on held-out 20% of cells). **g**, Specific embedding dimension activations distinguish each lung cell type in the fine-tuned model.

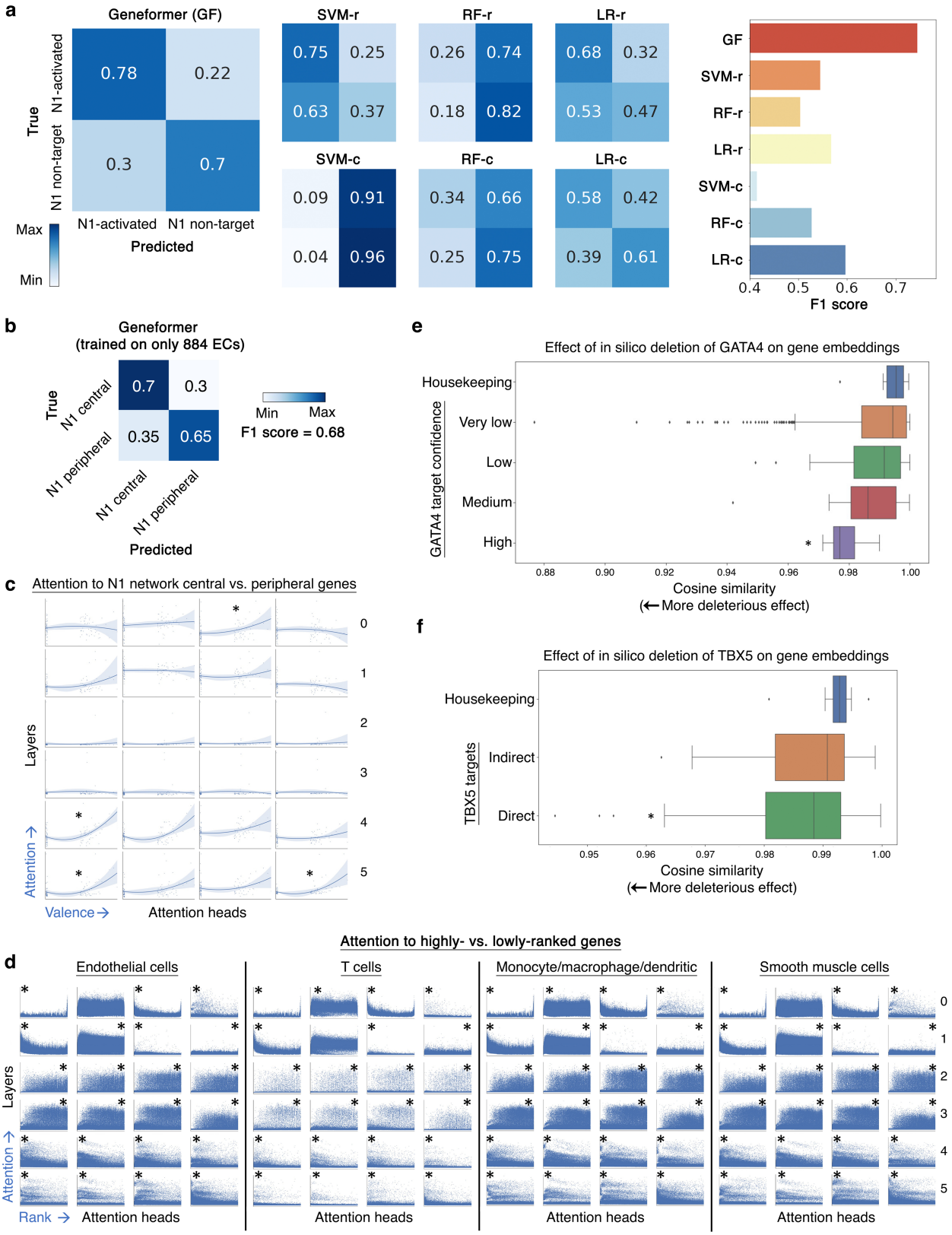


Extended Data Fig. 7 | See next page for caption.

Extended Data Fig. 7 | Geneformer boosted predictions in a diverse panel of downstream tasks. a, Confusion matrices and F1 score for Geneformer predictions vs. alternative methods (as described in Fig. 2a) for downstream task of distinguishing dosage-sensitive vs. insensitive transcription factors. **b**, Effect on cardiomyocyte embeddings from in silico deletion of genes linked by prior transcriptome-wide association study (TWAS)-prioritized GWAS[24] to cardiac MRI traits relevant to cardiac pathology (left ventricular (LV) end diastolic volume (EDV), LV end systolic volume (LVESV), LV ejection fraction (LVEF), and stroke volume (SV)) compared to in silico deletion of control cardiac disease genes expressed in cardiomyocytes but whose pathology occurs in non-cardiomyocyte cell types (hyperlipidemia). (*$p < 0.05$ by Wilcoxon, FDR-corrected; centre line = median, box limits = upper and lower quartiles, whiskers = 1.5x interquartile range, points = outliers). **c**, Quantitative PCR (QPCR) data of CRISPR-mediated knockout of *TEAD4* in iPSC-derived cardiomyocytes (n = 3, *$p < 0.05$ by t-test; centre line = median, box limits = upper and lower quartiles, whiskers = 1.5× interquartile range, points = experimental replicates). **d**, Confusion matrices and F1 score for Geneformer predictions vs. alternative methods for downstream task of distinguishing bivalent vs. non-methylated genes (56 highly conserved loci[28]). **e**, Confusion matrices and F1 score for Geneformer predictions vs. alternative methods for downstream task of distinguishing bivalent vs. Lys4-only methylated genes (56 highly conserved loci[28]).

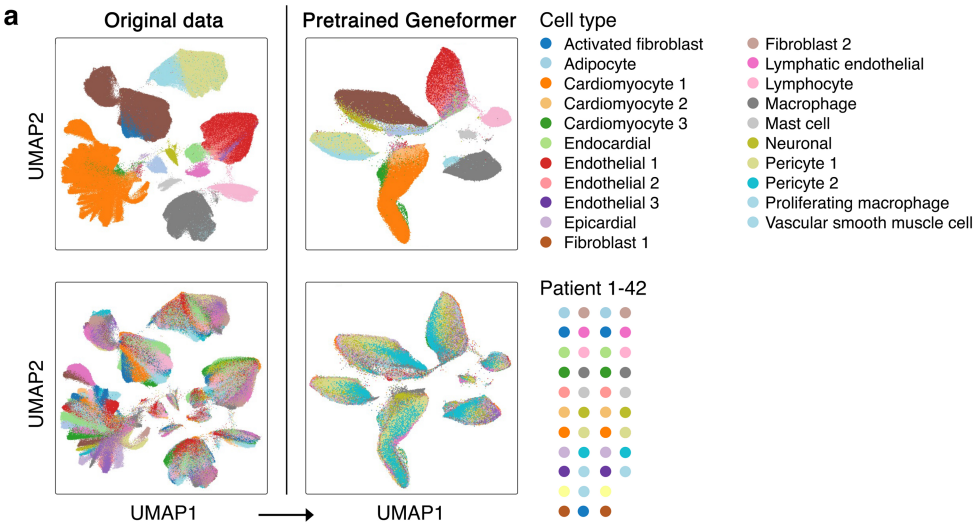


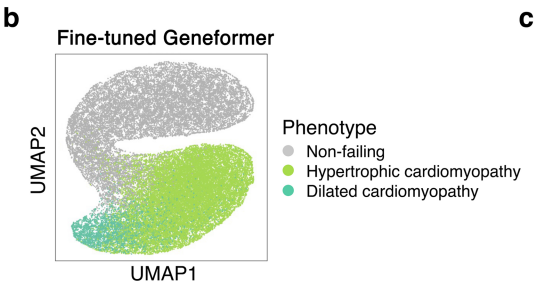
Extended Data Fig. 8 | Geneformer boosted predictions in a diverse panel of downstream tasks. a, Confusion matrix and F1 score for Geneformer predictions vs. alternative methods (as described in Fig. 2a) for downstream task of distinguishing genome-wide[30] bivalent vs. Lys4-only methylated genes with model fine-tuned only on 56 highly conserved loci[28]. **b**, ROC curve of Geneformer fine-tuned to distinguish genome-wide bivalent vs. Lys4-only-methylated genes using limited data (about 15K ESCs), compared to alternative methods. **c**, Confusion matrices and F1 score for Geneformer predictions vs. alternative methods for downstream task of distinguishing genome-wide bivalent vs. non-methylated genes with model fine-tuned on 80% of genome-wide loci and predicting on 20% of out of sample loci. **d**, Confusion matrices and F1 score for Geneformer predictions vs. alternative methods for downstream task of distinguishing long- vs. short-range transcription factors. **e**, Confusion matrices and F1 score for Geneformer predictions vs. alternative methods for downstream task of distinguishing central vs. peripheral genes within the N1-dependent network in endothelial cells.

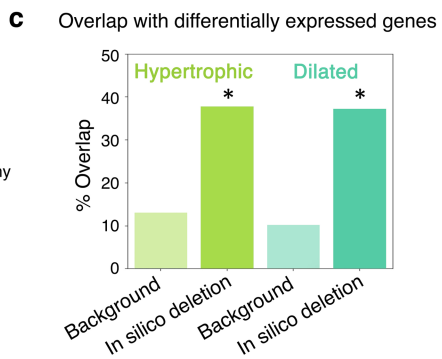


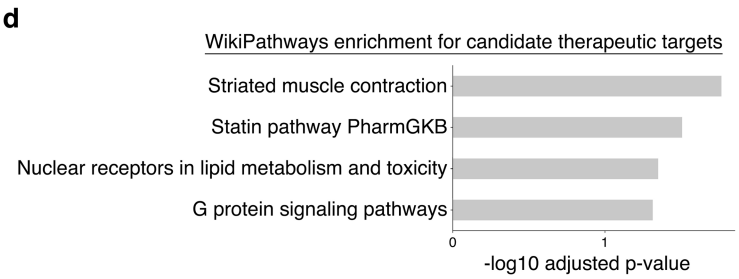
Extended Data Fig. 9 | See next page for caption.

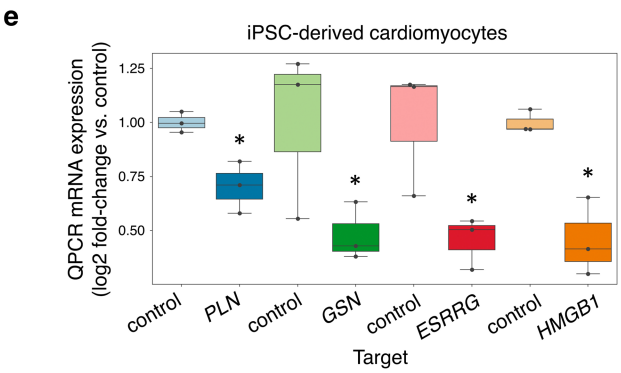
Extended Data Fig. 9 | In silico deletion strategy revealed network connectivity. a, Confusion matrices and F1 score for Geneformer predictions vs. alternative methods (as described in Fig. 2a) for downstream task of distinguishing N1-activated vs. non-targets. **b**, Confusion matrix and F1 score of Geneformer predictions of central vs. peripheral genes within the N1-dependent network in endothelial cells (ECs) with model fine-tuned only on 884 ECs from healthy or dilated aortas[14]. **c**, Pretrained Geneformer attention weights in aortic ECs demonstrated that specific attention heads learned in a completely self-supervised way the relative centrality of the top most central versus most peripheral genes in the N1-dependent gene network (higher valence = more central) (*p < 0.05 Wilcoxon, FDR-corrected). **d**, Pretrained Geneformer contextual attention versus gene rank in rank value encoding in the indicated aortic cell types, which each have different sets of highest ranked genes based on cell type context (higher rank is leftward on x axis) (*p < 0.05 by Wilcoxon, FDR-corrected, * position = side with higher attention). All cells used for analysis had the same number of genes so that the rank values would be comparable. **e**, In silico deletion of *GATA4* was significantly more deleterious to the previously reported highest confidence GATA4 targets[33] than to housekeeping genes. **f**, In silico deletion of *TBX5* was significantly more deleterious to previously reported TBX5 direct targets[34] than to housekeeping genes or TBX5 indirect targets. In (e–f): *p < 0.05 by Wilcoxon, FDR-corrected; centre line = median, box limits = upper and lower quartiles, whiskers = 1.5× interquartile range, points = outliers.











Extended Data Fig. 10 | See next page for caption.

Extended Data Fig. 10 | Geneformer fine-tuned cardiomyocyte embeddings clustered by phenotype. a, While original data (left) was highly affected by patient batch effect, cell embeddings generated by pretrained Geneformer (right) (without fine-tuning) clustered primarily by cell type. **b**, UMAP of cardiomyocyte embeddings from the model fine-tuned to distinguish cardiomyocytes in non-failing hearts from cardiomyocytes in patients with hypertrophic or dilated cardiomyopathy. **c**, Gene sets significantly associated with hypertrophic or dilated cardiomyopathy states by Geneformer in silico deletion disease modelling significantly overlapped with genes differentially expressed in those respective disease states (differentially expressed vs. non-failing) compared to the overlap of those differentially expressed genes with background genes (the remainder of the genes detected in cardiomyocytes that were not significantly associated with hypertrophic or dilated cardiomyopathy by Geneformer disease modelling) (*$p < 0.05$ by X^2 test, FDR-corrected). **d**, Pathway enrichment for genes whose in silico deletion in cardiomyocytes from hypertrophic cardiomyopathy patients significantly shifted embeddings towards the non-failing state and away from the dilated cardiomyopathy state, suggesting candidate therapeutic targets. **e**, QPCR data of CRISPR-mediated knockout of indicated genes in *TTN*$^{+/-}$ iPSC-derived cardiomyocytes ($n = 3$, *$p < 0.05$ by t-test). Centre line = median, box limits = upper and lower quartiles, whiskers = 1.5× interquartile range, points = experimental replicates.

nature portfolio

Corresponding author(s): Patrick Ellinor

Last updated by author(s): Oct 19, 2022

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	mygene (v3.2.1): https://github.com/SuLab/mygene.info requests (v2.25.1): https://github.com/psf/requests
Data analysis	The pretrained Geneformer model and transcriptome tokenizer are available on the Huggingface Model Hub (https://huggingface.co/ctheodoris/Geneformer). All other code used in this study is available upon request from the corresponding author. anndata (v0.7.6): https://github.com/scverse/anndata cudatoolkit (v11.2.2): https://developer.nvidia.com/cuda-toolkit cudnn (v8.1.0.77): https://developer.nvidia.com/cudnn datasets (v1.6.2): https://github.com/huggingface/datasets deepspeed (v0.3.10): https://github.com/microsoft/DeepSpeed gseapy (v0.10.8): https://github.com/zqfang/GSEApy harmonypy (v0.0.5): https://github.com/nasa/harmony-py hyperopt (v0.2.5): https://github.com/hyperopt/hyperopt loompy (v3.0.6): https://github.com/linnarsson-lab/loompy matplotlib (v3.4.2): https://github.com/matplotlib/matplotlib mpi4py (v3.0.3): https://github.com/mpi4py/mpi4py multiprocess (v0.70.11.1): https://github.com/uqfoundation/multiprocess nccl (v2.8.4.1): https://developer.nvidia.com/nccl networkx (v2.5): https://github.com/networkx numpy (v1.20.2): https://github.com/numpy/numpy openmpi (v4.0.5): https://github.com/open-mpi pandas (v1.2.4): https://github.com/pandas-dev/pandas parquet-cpp (v1.5.1): https://github.com/apache/parquet-cpp

pyarrow (v4.0.0): https://github.com/apache/arrow/tree/master/python/pyarrow
python (v3.8.5): https://www.python.org/downloads/
pytorch (v1.8.0): https://github.com/pytorch/pytorch
ray-tune (v1.9.2): https://github.com/ray-project/ray/tree/master/python/ray/tune
scanpy (v1.7.2): https://github.com/scverse/scanpy
scikit-learn (v0.24.2): https://github.com/scikit-learn/scikit-learn
scipy (v1.6.3): https://github.com/scipy/scipy
seaborn (v0.11.1): https://github.com/mwaskom/seaborn
statsmodels (v0.12.2): https://github.com/statsmodels/statsmodels
tdigest (v0.5.2.2): https://github.com/tdunning/t-digest
tensorboard (v2.4.1): https://github.com/tensorflow/tensorboard
tokenizers (v0.10.2): https://github.com/huggingface/tokenizers
torchmetrics (v0.3.2): https://github.com/PyTorchLightning/metrics
transformers (v4.6.0): https://github.com/huggingface/transformers

R (v4.0.5): https://www.r-project.org/
LoomExperiment (v1.8.0): https://bioconductor.org/packages/release/bioc/html/LoomExperiment.html
CellRanger (v6.0.1): https://support.10xgenomics.com/single-cell-gene-expression/software

For preprocessing and differential expression analysis of the cardiomyopathy dataset:
CellRanger (v4.0.0): https://support.10xgenomics.com/single-cell-gene-expression/software
cutadapt (v1.18): https://cutadapt.readthedocs.io/en/stable/
CellBender remove-background (v0.2): https://github.com/broadinstitute/CellBender
scR-Invex (sha1:4a067c5): https://github.com/broadinstitute/scrinvex
Python (v3.7): https://www.python.org/
scanpy (v1.6.0): https://github.com/theislab/scanpy
harmony-pytorch (v0.1.4): https://github.com/lilab-bcb/harmony-pytorch
ndd (v1.6.3): https://github.com/simomarsili/ndd
Scrublet (v0.2.1): https://github.com/AllonKleinLab/scrublet
R (v3.5.0, v3.6.0): https://www.r-project.org/
DESeq2 (v1.20.0): https://bioconductor.org/packages/release/bioc/html/DESeq2.html
limma (v3.36.5): https://bioconductor.org/packages/release/bioc/html/limma.html
edgeR (v3.22.5): https://bioconductor.org/packages/release/bioc/html/edgeR.html
GOstats (v2.46.0): https://bioconductor.org/packages/release/bioc/html/GOstats.html

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

Genecorpus-30M is available on the Huggingface Dataset Hub (https://huggingface.co/datasets/ctheodoris/Genecorpus-30M).
Processed single-nuclei transcriptomic data from non-failing hearts and hearts affected by hypertrophic cardiomyopathy or dilated cardiomyopathy are available through the Broad Institute's Single Cell Portal (https://singlecell.broadinstitute.org/single_cell) under project ID SCP1303 (https://singlecell.broadinstitute.org/single_cell/study/SCP1303/). Raw sequence data are available to authorized users through dbGaP (the database of Genotypes and Phenotypes) accession number phs001539.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

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

Sample size	For pretraining, we collected as much data as was publicly available and contributing towards a breadth of represented human tissues. For fine-tuning, we selected the largest sample set available within single datasets relevant to each application to test the efficacy of fine-tuning the model with limited data. For the cardiomyopathy disease modeling, we selected the largest sample set available to us of non-failing (n=13), hypertrophic cardiomyopathy (n=15), and dilated cardiomyopathy (n=11) left ventricles for single-nuclei RNA-sequencing. No statistical methods were used to predetermine sample size.
Data exclusions	For all pretraining and fine-tuning, single cells were excluded if single cell RNA-sequencing was low quality based on 1) total read counts not

Data exclusions	within three standard deviations of the mean within that dataset, 2) mitochondrial reads not within three standard deviations of the mean within that dataset, or 3) less than seven detected Ensembl-annotated protein-coding or miRNA genes. The publicly available data used in this study were collected as count matrices so initial filtering for sequencing quality control had already been performed by the original authors. However, for the cardiomyopathy disease modeling, samples were excluded if single-nuclei RNA-sequencing was low quality based on 1) <50% of reads in cells, 2) < 65% of reads confidently mapping to transcriptome, 3) < 90% valid barcodes, 4) abnormally low Q30, or 5) no ambient plateau in the unique molecular identifier (UMI) decay curve. Additionally, we only included non-failing heart samples that had a documented normal ejection fraction. Nuclei were removed if they 1) were enriched for mitochondrial reads, 2) were enriched for the proportion of reads mapping exclusively to exonic regions, 3) had a high prediction for being a doublet, 4) had extremely high or low UMI, 5) had extremely high or low number of genes detected, or 6) had low entropy. For experimental validation in engineered cardiac microtissues, only samples with confirmed CRISPR-mediated knockout of the intended target were analyzed.
Replication	For all AUC evaluations, 5 replicates were performed for each model by 5-fold cross-validation. For cardiomyopathy disease modeling, we performed single-nuclei RNA-seq in duplicate on left ventricles from 44 hearts. At least 1 replicate was retained for 39 samples after quality control. For experimental validation in engineered cardiac microtissues, at least 7 replicates were included for each condition (see figure legends for exact number of replicates for each experiment).
Randomization	For gene classification fine-tuning, labeled genes were randomized into training data (80%) and test data (20%), repeating for 5 cross-validation folds. For cardiomyopathy disease modeling, patients were randomized into training data (non-failing n=9, hypertrophic cardiomyopathy n=11, dilated cardiomyopathy n=9) and test data (non-failing n=4, hypertrophic cardiomyopathy n=4, dilated cardiomyopathy n=2).
Blinding	Blinding was not relevant to our study. Given the nature of the study, we sought to compare the predictive potential of the pretrained Geneformer model to alternative modeling approaches and to training with smaller and less diverse pretraining corpuses. For cardiomyopathy disease modeling, we sought to compare the transcriptional networks of cardiomyopathy and non-failing left ventricles. For experimental validation in engineered cardiac microtissues, measurements were quantified through software that did not involve qualitative assessment.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

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

Methods

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

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)	Patrick Ellinor Lab
Authentication	Cells were confirmed by genotyping.
Mycoplasma contamination	Cells are tested for mycoplasma contamination prior to use in stem cell culture facility.
Commonly misidentified lines (See ICLAC register)	No commonly misidentified lines were used in this study.

Human research participants

Policy information about studies involving human research participants

Population characteristics	All samples came from the Myocardial Applied Genomics Network (MAGNet) repository. Left ventricle tissue was obtained from a total of 44 individuals, including 12 dilated cardiomyopathy donors, 16 hypertrophic cardiomyopathy donors, and 16 non-failing donors. All samples were of European ancestry. 5/12 dilated cardiomyopathy donors were female, 5/16 hypertrophic cardiomyopathy donors were female, and 10/16 non-failing donors were female. Average age of dilated cardiomyopathy patients was 55, hypertrophic cardiomyopathy 49, and non-failing 57.
Recruitment	Dilated and hypertrophic cardiomyopathy left ventricle tissue was collected at time of heart transplantation. Non-failing left ventricle tissue was collected from deceased donors with no overt cardiovascular disease.
Ethics oversight	Written informed consent for research use of donated tissue was obtained from next of kin in all cases. Research use of

Ethics oversight	tissues were approved by the relevant institutional review boards at the Gift-of-Life Donor Program, the University of Pennsylvania, Massachusetts General Hospital, and the Broad Institute.

Note that full information on the approval of the study protocol must also be provided in the manuscript.