nature biomedical engineering

Article

https://doi.org/10.1038/s41551-024-01284-6

Simple and effective embedding model for single-cell biology built from ChatGPT

Received: 29 October 2023

Accepted: 16 October 2024

Published online: 6 December 2024

Check for updates

Yiqun Chen[1] & James Zou[1,2,3]

Large-scale gene-expression data are being leveraged to pretrain models that implicitly learn gene and cellular functions. However, such models require extensive data curation and training. Here we explore a much simpler alternative: leveraging ChatGPT embeddings of genes based on the literature. We used GPT-3.5 to generate gene embeddings from text descriptions of individual genes and to then generate single-cell embeddings by averaging the gene embeddings weighted by each gene's expression level. We also created a sentence embedding for each cell by using only the gene names ordered by their expression level. On many downstream tasks used to evaluate pretrained single-cell embedding models—particularly, tasks of gene-property and cell-type classifications—our model, which we named GenePT, achieved comparable or better performance than models pretrained from gene-expression profiles of millions of cells. GenePT shows that large-language-model embeddings of the literature provide a simple and effective path to encoding single-cell biological knowledge.

Recently, the field of single-cell biology has seen a surge in interest and efforts to develop 'foundation models', that is, models designed to learn embeddings of genes and cells to facilitate various downstream analyses. Several methods, such as scBERT[1], Geneformer[2] and scGPT[3], have been recently proposed to tackle this challenge. At a conceptual level, they adopt similar recipes that consist of the following steps:

1. Adopt a deep learning architecture (often from the transformer family[4]).
2. Gather large-scale single-cell gene-expression datasets for pretraining the model in a self-supervised manner (that is, by imputing some masked-out expression values). The trained encoder maps input genes and cells to a high-dimensional embedding vector encapsulating the underlying biology.
3. For downstream tasks, one can optionally utilize a modest amount of task-specific data to fine-tune the model, boosting its predictive capabilities.

Notably, the approach outlined above derives embeddings only from gene-expression datasets, without making any use of the literature and pre-existing knowledge about a gene. While this strategy has shown some success in applications to single-cell transcriptomics data and tasks, it has several limitations. First, it takes extensive effort to collect, process and train on the large-scale single-cell transcriptomics data. Furthermore, the signals from extracted embeddings are heavily dependent on the gene-expression data used in step 2, which does not take advantage of the vast research and literature summarizing the functionalities of a gene, potentially leading to sample inefficiency and suboptimal results in certain applications. Therefore, in this study, we introduced GenePT—an alternative, complementary approach that represents genes and cells by utilizing OpenAI's ChatGPT text embedding models[5] and investigated the feasibility of encoding the biology of genes and cells using natural language (see an overview in Fig. 1a–c).

The intuition for our approach is as follows: large language models (LLMs) such as GPT-3.5 and GPT-4 have been trained with substantial computational resources on extensive text corpus[6], including biomedical literature, and have demonstrated remarkable ability in understanding, reasoning and even generating biomedical text[7–10]. Consequently, we hypothesize that LLM-derived embeddings of gene summaries and functionalities—which often are curated from a broad spectrum of experiments and studies—could more directly capture the underlying biology.

[1]Department of Biomedical Data Science, Stanford University, Stanford, CA, USA. [2]Department of Electrical Engineering, Stanford University, Stanford, CA, USA. [3]Department of Computer Science, Stanford University, Stanford, CA, USA. e-mail: jamesz@stanford.edu

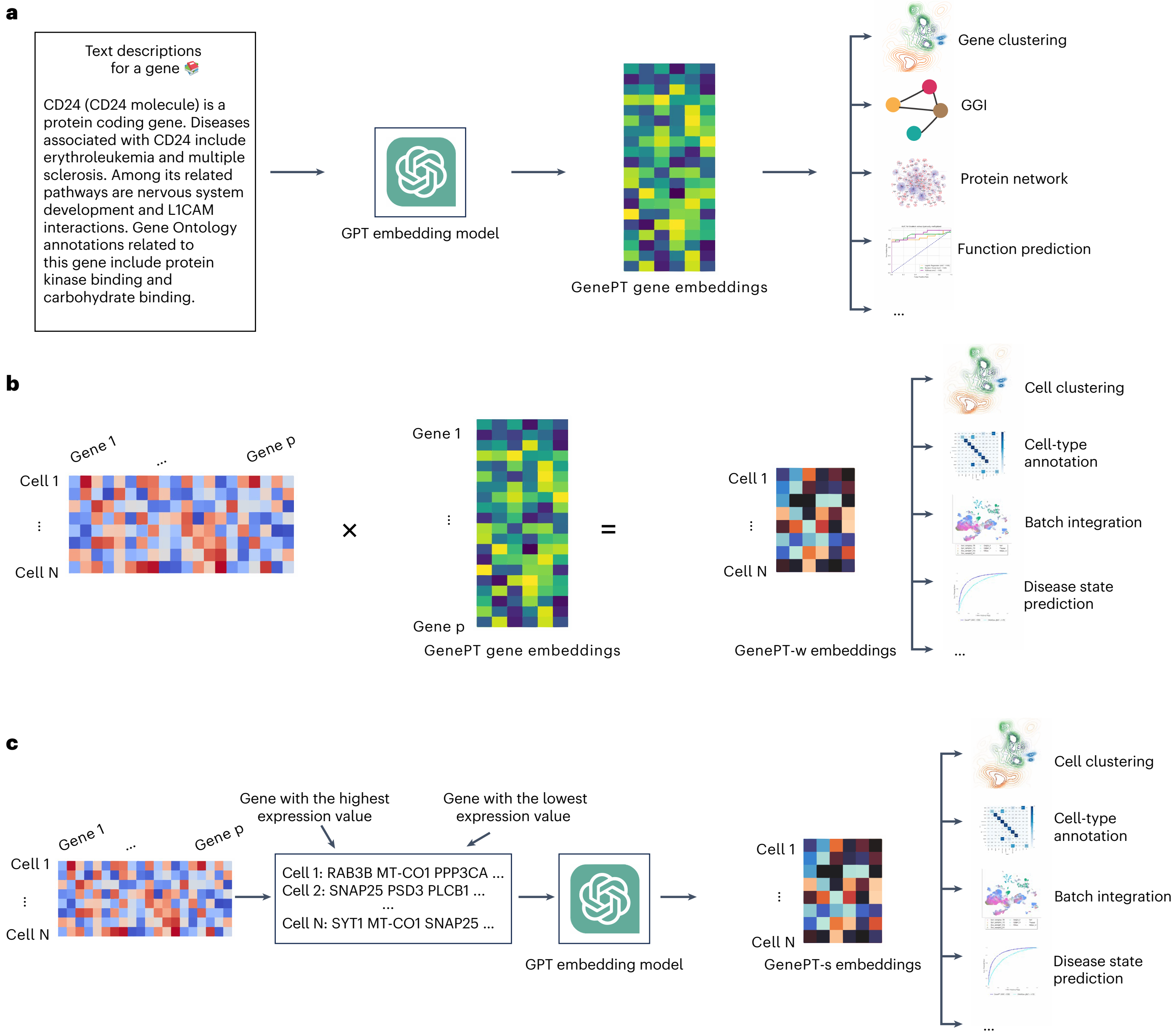

a
Text descriptions for a gene
CD24 (CD24 molecule) is a protein coding gene. Diseases associated with CD24 include erythroleukemia and multiple sclerosis. Among its related pathways are nervous system development and L1CAM interactions. Gene Ontology annotations related to this gene include protein kinase binding and carbohydrate binding.
GPT embedding model
GenePT gene embeddings
Gene clustering
GGI
Protein network
Function prediction
...
b
Gene 1
...
Gene p
Cell 1
⋮
Cell N
×
Gene 1
⋮
Gene p
GenePT gene embeddings
=
Cell 1
⋮
Cell N
GenePT-w embeddings
Cell clustering
Cell-type annotation
Batch integration
Disease state prediction
...
c
Gene 1
...
Gene p
Cell 1
⋮
Cell N
Gene with the highest expression value
Gene with the lowest expression value
Cell 1: RAB3B MT-CO1 PPP3CA ...
Cell 2: SNAP25 PSD3 PLCB1 ...
...
Cell N: SYT1 MT-CO1 SNAP25 ...
GPT embedding model
Cell 1
⋮
Cell N
GenePT-s embeddings
Cell clustering
Cell-type annotation
Batch integration
Disease state prediction
...

Fig. 1 | An overview of the GenePT framework. **a**, For each gene, we extract its corresponding text summary from NCBI and use GPT-3.5 text embedding as its representation. **b**, In the GenePT-w cell embeddings framework, we average gene embeddings from step a, weighted by their cell expression levels, and normalize these cell embeddings to a unit ℓ_2 norm. **c**, In the GenePT-s cell embeddings framework, each cell from the input single-cell data is translated into a natural language sentence based on ranked gene expressions, and the GPT-3.5 embedding of the entire sentence is used to represent the cell.

We evaluated the generated embeddings on several biologically driven tasks; our findings reveal that GenePT exhibits performance comparable to, and sometimes surpassing, specially designed models such as Geneformer across a diverse set of downstream tasks. GenePT offers several advantages to existing embedding models for single-cell RNA-seq data: (i) it performs better on several biological tasks; (ii) it does not require extensive dataset curation processing, or additional pretraining on the genomics data; and (iii) it is very simple to use and to generate gene and cell embeddings. In particular, GenePT uses LLM-based embeddings, an orthogonal source of information compared with the expression-based representations, and our findings suggest a promising new direction of combining these two approaches.

Foundation models, trained on broad data and applicable across a wide range of use cases, have revolutionized the fields such as natural language processing and computer vision by learning informative representations known as embeddings of the input[11]. The impressive results have motivated efforts to adapt such models to tackle tasks in biology, especially in the field of single-cell transcriptomics[2,3,12–14]. The aspiration behind these models is to craft an analogous 'foundational model' for single-cell transcriptomics, with the hope that the resulting model could display broad capabilities across an array of biological tasks rather than just one specific task.

Promising applications of these models include cell-type annotation (where a cell is labelled based on its biological identity[1,15]); gene functional and regulatory network inference (where the functionality of individual genes and clustered gene groups is examined[3,16]); sample integration[15] (where the goal is to account for cases where transcript abundance is influenced primarily by technical noise instead of underlying biology); and tissue drug response prediction[14]. For instance, using a transformer architecture, Geneformer[2] was built with extensive pretraining on the ranks of gene-expression levels through masked token prediction across 30 million cells collected from Gene Expression Omnibus[17]. This model excels in tasks ranging from understanding network dynamics to deciphering network hierarchy. Similarly, scGPT[3] used generative pretraining (with normalized gene-expression prediction as the task) on 33 million cells from the CELLxGENE collection for training[18]. scGPT's effectiveness is evident in its downstream applications, such as perturbation prediction, batch integration and cell-type annotation. Other innovative endeavours include scBERT[1], where the authors applied a BERT model[19] to represent single-cell RNA-seq data and demonstrated state-of-the-art performance in cell-type annotation and novel cell-type discovery, as well as scFoundation[14], where the authors expanded the training to 50 million cells and utilized an asymmetric encoder–decoder architecture to accommodate for the highly sparse nature of the data.

Pioneering work in applying natural language processing techniques to gene and cell biology aims to represent the semantics of biomedical terms by training co-occurrence-based neural network embeddings that map individual terms (that is, gene names) to vectors[20–22]. Recently, researchers have begun exploring the use of LLMs for biomedically focused tasks, leveraging their capability to encode information from the entire input text. This approach allows for more nuanced and dynamic representations. For example, Hou and Ji[23] used ChatGPT for cell-type annotation; Wysocki et al.[24] investigated biomedical meanings encoded by BioBERT and BioMegatron embeddings; and Ye et al.[25] utilized instruction fine-tuning to achieve competitive results on graph data task benchmarks with an LLM. Our proposal, GenePT, is directly inspired by the extensive previous work in the biomedical natural language processing community and the demonstrated effectiveness of ChatGPT-family models in biomedical tasks. Compared with previous works that directly query LLMs for biological tasks, our method instead utilizes the input descriptions of each gene, which can be sourced from high-quality databases such as the National Center for Biotechnology Information (NCBI)[26], and the embedding model of LLMs, which suffers less from problems such as hallucination. While our paper is under preparation, Levine et al.[27] have independently embarked on a conceptually related approach to ours, where each cell is transformed into a sequence of gene names, ranked by expression level and truncated at top 100 genes. The emphasis of their paper, however, is on generating new cells conditional on cell types.

Our work makes the following contribution to the literature: we demonstrate that a simple approach, using natural language embeddings of gene function descriptions—such as summaries readily available from sources like the NCBI gene database[28]—can encapsulate the underlying biological relationships and insights associated with genes when evaluated on biologically relevant prediction tasks. The strength of our GenePT approach lies in its simplicity, effectiveness and broad utility.

Results

GenePT embeddings capture underlying gene functionality

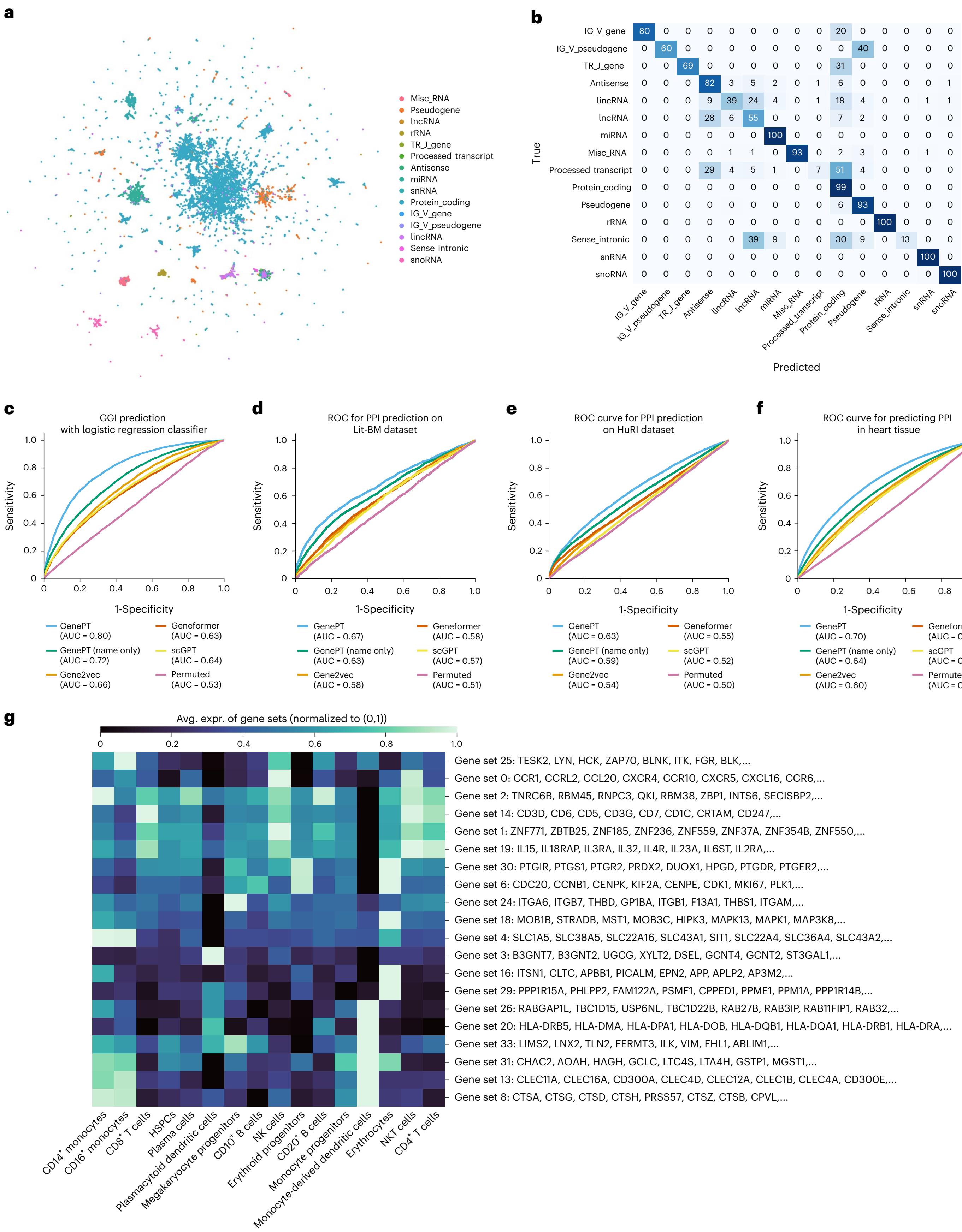
In Fig. 2a, we show a 2D UMAP of the GenePT embeddings for over 34,000 genes using the `text-embedding-ada-002` model. These genes belong to the top 15 most prevalent functional classes (see online Supplementary Information for a detailed class breakdown). The UMAP reveals distinct clusters when coloured by various gene functionality groups, implying that GenePT embeddings encode the functions of the genes. This confirms that language model embedding retains key biological information, as functionality is frequently found in NCBI gene summaries. To evaluate the observations in Fig. 2a more quantitatively, we further divided the genes into a 70%/30% train/test split and evaluated the prediction accuracy using an ℓ_2-regularized logistic regression on the 15 classes. The predicted functional class aligns with the true annotation well, with an overall accuracy of 96% and high class-specific accuracies and only minor misclassifications between closely related functional groups such as lincRNA, lncRNA and processed transcripts (Fig. 2b).

We further assessed the efficacy of GenePT embeddings in predicting gene–gene interactions (GGIs) in Fig. 2c. We compared the receiver operating characteristic area under the curve (ROC-AUC) for three methods on the test GGI dataset provided in Du et al.[20], derived from shared Gene Ontology annotations: (i) sum of the GenePT embedding of two genes with an ℓ_2-regularized logistic classifier (LR), yielding an AUC of 0.80; (ii) sum of the Gene2vec/scGPT/Geneformer pretrained embeddings with an LR classifier (resulting in an AUC of 0.63–0.66); and (iii) sum of two random embeddings ($d = 1{,}536$, same dimension as GenePT) paired with an LR classifier, which served as a negative control (an AUC of 0.53). As shown in Fig. 2c, GenePT embeddings considerably enhance performance when compared with the other embedding models using the same downstream classifier. Even compared with a more complex downstream classifier (for instance, deep neural networks with a reported test set AUC of 0.88 in Du et al.[20]), GenePT still performs competitively in this task.

Next, we evaluated the ability to predict protein–protein interactions (PPIs) using GenePT gene embeddings, as depicted in Fig. 2d–f. We compared the ROC-AUC (refer to online Supplementary Information for the precision–recall curves) for three methods across three distinct PPI datasets: those derived from the literature (panel d), comprehensive assays (panel e) and biophysical contact annotations (panel f). For all three datasets, using the sum of the GenePT embeddings of two genes as input, combined with an ℓ_2-regularized logistic regression, results in better performance than all other models considered. These results suggest that GenePT's literature-based embedding captures information relevant to gene and protein interactions; a promising future direction is to combine GenePT embeddings with protein embeddings learned from 3D structures or protein language models.

Finally, we delved into cell-type specific activations among the GenePT-derived gene programmes within human immune tissue datasets through a 'zero-shot' approach. We first constructed a similarity graph based on cosine similarities between the GenePT embeddings by placing an edge between two genes if the cosine similarity is larger than 0.9 and applied Leiden clustering to the resulting graph at a resolution of 20. Randomly sampled 20 gene programmes comprising 10 or more genes are depicted in Fig. 2g. Here we show the average expression levels of these gene programmes, stratified by cell types. The observed selective activation of these programmes aligns with established biological knowledge where the identified gene sets are known to be functionally distinct and are differentially expressed across different cell types (as an example, gene set 8 comprising cathepsin families and gene set 24 comprising integrin (ITG) families). These findings underscore that GenePT-inferred gene programmes effectively capture biologically pertinent functional groups; additional results with different similarity thresholds can be found in Supplementary Information online.

Fig. 2 | GenePT gene embeddings encode underlying biology. **a**, 2D UMAP visualization of GenePT embeddings, coloured by different gene functionality groups. **b**, Confusion matrix of gene function prediction utilizing GenePT embeddings, combined with an ℓ_2-regularized logistic regression on a randomly held-out 30% test set. **c**, Prediction accuracy on a GGI benchmark dataset derived from GEO expression data[20]. **d**, Prediction accuracy for PPIs verified by high-quality binary literature datasets[49]. **e**, Prediction accuracy on the human binary protein interactions dataset[48]. **f**, Prediction accuracy on human heart tissue protein–protein functional interactions[50]. **g**, Cell-type specific activation among GenePT-embeddings-extracted gene programmes (a random subset of genes is shown for each programme) in a human immune tissue dataset[52]. The patterns of average gene expressions for identified gene programmes in different cell types are congruent with those previously identified in Cui et al.[3].



True \ Predicted	IG_V_gene	IG_V_pseudogene	TR_J_gene	Antisense	lincRNA	lncRNA	miRNA	Misc_RNA	Processed_transcript	Protein_coding	Pseudogene	rRNA	Sense_intronic	snRNA	snoRNA
IG_V_gene	80	0	0	0	0	0	0	0	0	20	0	0	0	0	0
IG_V_pseudogene	0	60	0	0	0	0	0	0	0	0	40	0	0	0	0
TR_J_gene	0	0	69	0	0	0	0	0	0	31	0	0	0	0	0
Antisense	0	0	0	82	3	5	2	0	1	6	0	0	0	0	1
lincRNA	0	0	0	9	39	24	4	0	1	18	4	0	0	1	1
lncRNA	0	0	0	28	6	55	0	0	0	7	2	0	0	0	0
miRNA	0	0	0	0	0	0	100	0	0	0	0	0	0	0	0
Misc_RNA	0	0	0	0	1	1	0	93	0	2	3	0	0	1	0
Processed_transcript	0	0	0	29	4	5	1	0	7	51	4	0	0	0	0
Protein_coding	0	0	0	0	0	0	0	0	0	99	0	0	0	0	0
Pseudogene	0	0	0	0	0	0	0	0	0	6	93	0	0	0	0
rRNA	0	0	0	0	0	0	0	0	0	0	0	100	0	0	0
Sense_intronic	0	0	0	0	0	39	9	0	0	30	9	0	13	0	0
snRNA	0	0	0	0	0	0	0	0	0	0	0	0	0	100	0
snoRNA	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100

Table 1 | Cross-validated AUC for GenePT predictions versus alternative embeddings for downstream tasks

Model	Fivefold Cross Validation AUC±s.d.			
	Dosage sensitivity	Bivalent versus non-methylated	Bivalent versus Lys4-methylated	TF range
Geneformer (fine-tuned)	0.91±0.02	**0.93±0.07**	0.88±0.09	0.74±0.08
Gene2vec+LR	0.91±0.03	0.66±0.07	0.91±0.04	**0.83±0.14**
Gene2vec+RF	0.86±0.05	0.63±0.14	0.89±0.04	0.66±0.15
BioLinkBert+LR	0.87±0.04	0.78±0.10	0.87±0.04	0.31±0.14
BioLinkBert+RF	0.87±0.02	0.80±0.06	0.85±0.07	0.54±0.23
Random Embed+LR	0.54±0.04	0.59±0.03	0.46±0.07	0.36±0.16
Random Embed+RF	0.49±0.04	0.60±0.08	0.42±0.12	0.54±0.18
GenePT (name only)+LR	0.85±0.05	0.85±0.01	0.89±0.05	0.61±0.25
GenePT (name only)+RF	0.89±0.02	0.90±0.02	0.91±0.04	0.58±0.22
GenePT+LR	0.89±0.03	0.91±0.06	0.94±0.03	0.73±0.25
GenePT+RF	0.92±0.02	0.92±0.06	**0.95±0.04**	0.64±0.07
GenePT (LLama-7B)+LR	**0.93±0.04**	0.88±0.07	0.93±0.05	0.67±0.25
GenePT (LLama-7B)+RF	0.92±0.02	0.89±0.07	0.93±0.03	0.63±0.32

Distinguishing (i) dosage-sensitive versus insensitive TFs; (ii) bivalent versus non-methylated gene; (iii) bivalent versus Lys4-only-methylated genes; and (iv) long-range versus short-range TFs. The performance for Geneformer is reproduced from Theodoris et al.[2] and is based on a fine-tuned sequence classification model. Here random embed denotes an embedding identical in size to GenePT with entries drawn from i.i.d. $\mathcal{N}(0,1)$. This serves as a 'negative control' to ensure that signals in GenePT are not merely due to a larger embedding dimension. We use RF and LR to denote random forest and logistic regression models with default parameters in `scikit-learn`, respectively. The best result for each task is bolded.

GenePT embeddings enable accurate predictions of chromatin dynamics and dosage sensitivity

In this section, we delve into specific biological tasks that predict the roles of genes in network dynamics with datasets curated from the literature by Theodoris et al.[2]: dosage-sensitive versus dosage-insensitive transcription factors (TFs), bivalent versus non-methylated genes, Lys4-only-methylated versus non-methylated genes, and long- versus short-range TFs. These tasks were used to demonstrate the utility of Geneformer. We assess the performance of GenePT and Gene2vec embeddings by fivefold cross-validated ROC-AUC with either an ℓ_2 penalized logistic regression (LR) or a random forest (RF) classifier using default parameters from `scikit-learn`[29]. By contrast, Geneformer results, as reported in Theodoris et al.[2], are based on a fine-tuned transformer model. We also reported some variants of the GenePT framework: BioLinkBert embedding of the gene summaries; or GPT-3.5 embedding of only the gene names (without context or descriptions); and random embeddings matching the GenePT dimension ($d = 1{,}536$).

Table 1 illustrates that GenePT embeddings consistently achieve competitive results, sometimes even surpassing Geneformer, although the latter benefits from a large pretraining dataset and a more intricate classification head. Interestingly, GPT-3.5 embeddings of only gene names also show high accuracies in some tasks. This might be due to two aspects: (1) gene nomenclature attempts to designate functionally related or homologous genes with similar symbols to enable grouping[30], and (2) the underlying language model and tokenizer for GPT-3.5 might grasp the biological significance of these gene symbols owing to extensive pretraining on scientific text[31]. Open-source embeddings also show competitive performance: LLama-7B[32] essentially matches the results of GPT-3.5-based embeddings, with BioLinkBert and Gene2vec following closely behind. As expected, random embeddings exhibit results similar to random guessing. The stark contrast in predictive performance between GenePT and random embeddings indicates that it is unlikely that the GenePT performance is simply due to a large embedding dimension ($d = 1{,}536$). In addition, since we used low-complexity, off-the-shelf ℓ_2-regularized logistic regression and random forests, and reported results based on fivefold cross-validation, it is unlikely that the performance is due to model overfitting. In summary, these results underscore the potential of our versatile GenePT approach, which compares favourably with state-of-the-art deep learning models specifically crafted for single-cell RNA sequencing data.

Finally, it is crucial to confirm that the promising results are not simply the result of information leakage, such as test set data being included in the original NCBI gene summaries used as input for GenePT. We expand on these concerns in detail in online Supplementary Information.

GenePT learns representations that reflect cell biology

In this section, we demonstrate that the 'zero-shot' GenePT embedding approaches capture the biology of single-cell datasets. We first evaluate the supervised and unsupervised learning results of using GenePT embeddings on six different single-cell RNA-seq datasets representing cells from circulatory systems (aorta and artery), bone tissues (bones, myeloid), the pancreas and immune cells collected from healthy individuals and patients with multiple sclerosis. In addition, we demonstrate that GenePT-w embeddings, which can be generated very efficiently using pretrained gene embeddings, can be further fine-tuned on a diverse panel of cell-level tasks using limited task-specific data to boost predictive accuracy.

We quantified the concordance between biological annotations (that is, cell types, cancer types and donor ages) and k-means clustering labels inferred from (i) pretrained Geneformer embeddings, (ii) pretrained scGPT embeddings, (iii) GenePT-w embeddings (as in Fig. 1b) and (iv) GenePT-s embeddings (as in Fig. 1c). We quantified the concordance using both adjusted mutual information (AMI) and adjusted Rand index (ARI) in Table 2. We see that latent representations via GenePT-w and scGPT broadly outperformed both the GenePT-s and Geneformer embeddings in terms of AMI and ARI metrics: across nine tasks, scGPT and GenePT each provide the most biological signal on five and four tasks, respectively. This demonstrates that GenePT cell embeddings capture biological variations comparable to two leading single-cell foundation models. An important caveat is that concordance with cell types and annotations is a limited measure of the utility of embedding, although it is widely used. We also included additional classification results for a cell-type annotation task via a nearest-neighbour approach on these datasets in online Supplementary Information. This analysis yielded similar findings that GenePT-w and scGPT are two of the best-performing methods in this setting, and both consistently

Table 2 | Assessing the association between different latent cell representations and biological annotations

		Geneformer			scGPT			GenePT-w			GenePT-s		
Dataset	**Annotation**	**ARI**	**AMI**	**ASW**	**ARI**	**AMI**	**ASW**	**ARI**	**AMI**	**ASW**	**ARI**	**AMI**	**ASW**
Aorta	Phenotype	0.10	0.12	−0.005	**0.12**	**0.12**	0.01	0.09	**0.12**	−0.04	**0.12**	0.11	**0.02**
	Cell type	0.21	0.31	−0.04	0.47	**0.64**	0.18	**0.54**	0.60	0.03	0.31	0.47	0.04
Artery	Cell type	0.39	0.59	0.10	0.36	0.59	0.15	**0.42**	**0.67**	**0.16**	0.36	0.56	0.06
Bones	Cell type	0.09	0.16	−0.01	0.12	0.21	−0.01	**0.21**	**0.29**	**0.02**	0.17	0.28	0.003
Myeloid	Cancer type	0.16	0.18	0.03	**0.27**	**0.29**	**0.08**	0.25	0.27	0.02	0.17	0.17	0.06
	Cell type	0.19	0.29	−0.02	**0.44**	**0.53**	**0.13**	0.21	0.28	0.001	0.30	0.41	0.03
Pancreas	Cell type	0.04	0.11	−0.09	0.21	0.41	0.05	**0.49**	**0.69**	**0.15**	0.30	0.50	0.10
Multiple sclerosis	Age	0.04	0.11	−0.1	0.04	0.11	−0.06	**0.07**	**0.13**	−0.07	0.06	0.12	**−0.03**
	Cell type	0.21	0.35	−0.05	**0.25**	**0.44**	**0.04**	0.17	0.32	−0.02	0.19	0.35	0.002

This analysis involves datasets representing cells from circulatory systems (aorta and artery), bone tissues (bones, myeloid), the pancreas and immune cells collected from healthy individuals and patients with multiple sclerosis. We utilized pretrained Geneformer and scGPT embeddings for this task. The ARI and AMI were computed to compare the labels derived from *k*-means clustering with the true annotations of the original samples (higher values indicate better alignment); the average silhouette width (ASW) was calculated using the true annotations of original samples to assess the cohesion and separation of the clusters. Best results for each task are bolded.

Table 3 | Test set performance on disease phenotype prediction tasks

		Classification metrics on the test set			
Dataset	**Embeddings**	**Accuracy**	**Precision**	**Recall**	**F1**
Cardiomyocyte (cell-level)	scGPT	0.67	0.63	0.62	0.61
	Geneformer	0.59	0.51	0.47	0.47
	GenePT-w	0.68	0.65	0.62	0.63
	Fine-tuned GenePT-w	0.78	0.76	0.77	0.76
Lupus (patient-level)	scGPT	0.91	0.83	0.86	0.84
	Geneformer	0.89	0.80	0.80	0.80
	GenePT-w	0.91	0.84	0.81	0.82
	Fine-tuned GenePT-w	0.96	0.98	0.89	0.93

We report test set accuracy, precision, recall and F1 for predicting phenotypic features on the cardiomyocyte (NF, non-failing heart; HCM, hearts with hypertrophic cardiomyopathy; DCM, hearts with dilated cardiomyopathy)[33] and lupus (disease or healthy) datasets[58]. The results are reported on the cell level for the cardiomyopathy dataset and patient level for the lupus dataset by applying an ℓ_2-regularized logistic regression on pretrained embeddings from scGPT, Geneformer, GenePT-w and fine-tuned GenePT-w.

outperform Geneformer and GenePT-s (except for one dataset) in terms of prediction accuracy. Interestingly, a simple ensembling of the nearest neighbours retrieved by different embeddings (GenePT-w, GenePT-s and scGPT) enhanced the predictive performance (see details in online Supplementary Information). This suggests that natural language embeddings, such as GenePT, could provide complementary insights to expression-derived foundation models such as scGPT in single-cell biology tasks.

In addition, we explored whether we could further improve the GenePT embeddings for specific downstream tasks by performing fine-tuning on the cardiomyocyte and lupus datasets. The cardiomyocyte dataset was first published in Chaffin et al.[33] with samples from non-failing hearts ($n = 9$) or hearts affected by hypertrophic ($n = 11$) or dilated ($n = 9$) cardiomyopathy. The lupus dataset consists of 120 lupus patients and 22 healthy controls, and the scientific question of interest is to predict clinical phenotypes on the individual patient level from scRNA-seq data[34].

For cardiomyopathy prediction, we first divided the datasets at the patient level (14 for training, 5 for validation and 5 for testing) and reported phenotypic prediction accuracies on the test sets, as the scientific question was to distinguish cardiomyocytes in non-failing hearts from those in hypertrophic or dilated cardiomyopathy samples.

We fine-tuned the GenePT-w embeddings on training patients to predict the three phenotypes: hypertrophic, dilated and non-failing. We then evaluated the performance of the fine-tuned embeddings for cells from the held-out patients.

For the lupus dataset, we divided the data into training and validation sets (72 lupus and 13 healthy participants) and a test set (48 lupus and 9 healthy participants). We fine-tuned GenePT-w on the training patients and reported patient-level classification by generating a pseudo-bulk expression for each patient. We also assessed a cell-type fine-tuning metric by examining cell-type prediction performance through fine-tuning on 1% of the training data, to evaluate the gain from fine-tuning GenePT-w embeddings under the setting where high-quality cell-type annotations are hard to acquire.

We report the phenotype classification results in Table 3: similar to the cell-type annotation results in the main text, pretrained GenePT-w embeddings perform competitively with both scGPT and Geneformer. Moreover, fine-tuning the GenePT data further increased the performance on both datasets by a considerable margin. For the cardiomyocyte tasks, we provided a more detailed visualization and class-wise prediction accuracy in Fig. 3a, where we see that fine-tuning primarily contributes to the improved distinction of NF (non-failing heart) and HCM (hearts with hypertrophic cardiomyopathy) samples. This aligns with the established clinical observation that HCM is a heterogeneous disease ranging from asymptomatic diagnoses to serious heart failure, and the causal gene mutations have also been of much interest[35].

In addition to the improved patient-level phenotype prediction in the lupus dataset, we also considered whether there was an improvement in cell-type annotation. In Fig. 3b, we plot the confusion matrices from predicting the cell types using different embeddings, coloured by annotated cell types. We note that fine-tuning GenePT-w on merely

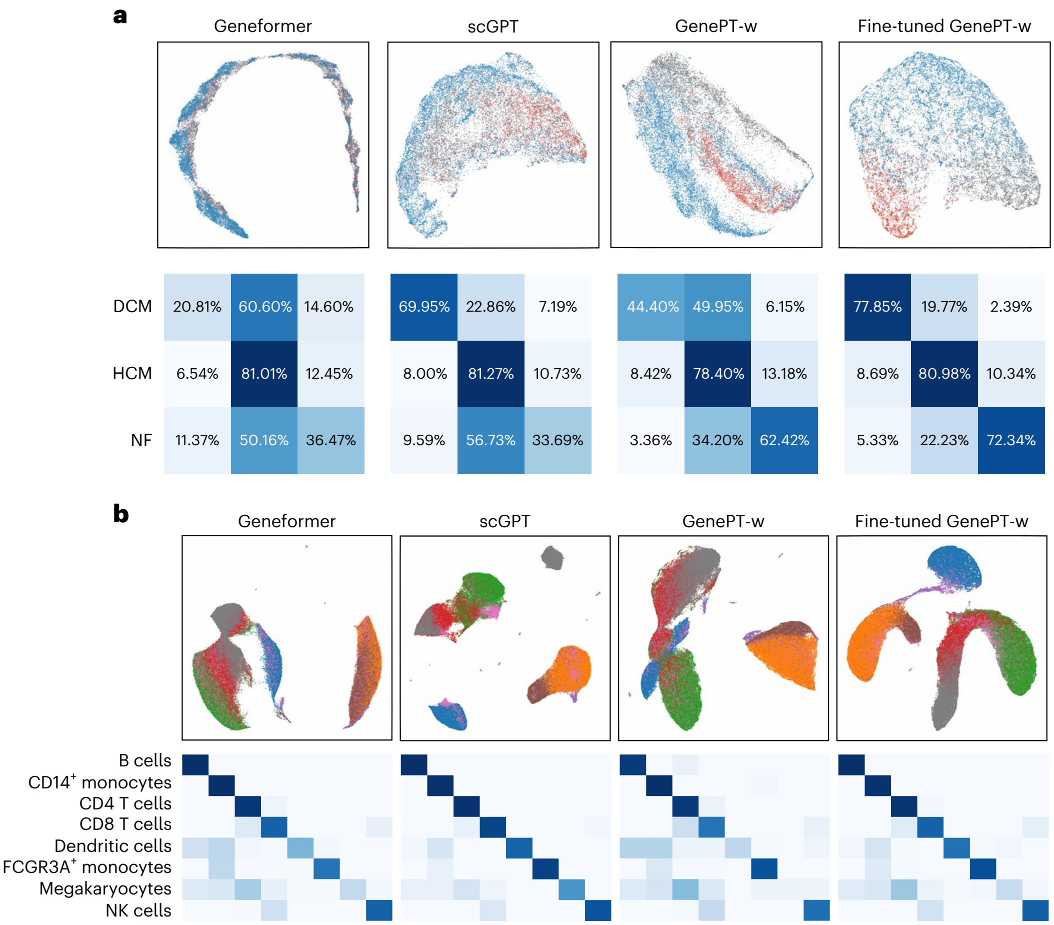


Fig. 3 | Fine-tuned GenePT embeddings improve downstream disease and cell type annotations. a, UMAP visualization for different embeddings on the test patients from the cardiomyopathy dataset, coloured by disease phenotypes, as provided in the original study[33]. **b**, UMAP visualization for different embeddings on the test patients from the lupus dataset, coloured by annotated cell types, as provided in the original study[58].

1% of the data improved the predictive performance (overall accuracy from 86% to 91%), and the most notable improvement was in correctly distinguishing dendritic cells from B cells and CD14$^+$ monocytes, a challenging task in lupus patients owing to altered immune system dynamics, especially in the dendritic cell compartment, subset frequency and localization[36].

GenePT embedding removes batch effect while preserving underlying biology

We next assess whether GenePT embeddings are robust to batch-dependent technical artefacts such as patient variability. We compared the performance of GenePT with pretrained Geneformer and scGPT using a 10% random sample from the cardiomyopathy dataset and a 20% random sample from the aorta dataset consisting of cells in healthy and dilated aortas[37].

Recall that the primary variation of interest is the difference between cardiomyocytes in non-failing hearts and those in hypertrophic or dilated cardiomyopathy samples. However, in practice, such effects can be confounded by non-biological variations such as technical artefacts. We performed the following analysis to quantify the patient-level batch effects: (i) we first project the data (either the original RNA-seq or one of the pretrained embeddings) into the top 50 principal components; (ii) we then applied k-means clustering with $k = 42$, which is the number of distinct patients; (iii) we compute ARI between the cell clusters and patient clusters. Higher ARI values indicate more patient-level batch effects. The original scRNA-seq data have a high ARI of 0.33, suggesting strong batch effects. Using the GenePT-s, Geneformer and scGPT, the ARI dropped to 0.07, 0.01 and 0.01 respectively, showing that these embeddings are robust to batch effects.

We also investigated whether these embeddings could preserve the underlying disease phenotype (that is, non-failing versus cardiomyopathy) of the patients from whom the cells were collected. To this end, we randomly split the cardiomyocytes into 80%/20% train/test sets and evaluated the predictive performance using the ℓ_2-regularized logistic regression on top of the following pretrained embeddings: (i) GenePT-s, (ii) scGPT and (iii) Geneformer. Overall, GenePT-s and scGPT achieve nearly identical performance on the held-out test set (88% accuracy, 88% precision and 88% recall for both embeddings for predicting disease label), whereas the performance for pretrained Geneformer trailed behind (71% accuracy, 72% precision and 71% recall).

We then conducted the same set of analyses on the aorta dataset. Recall that the cells were collected over 11 patients with four different phenotypes: healthy, ATAA with ascending only, ATAA with ascending with descending thoracic aortic aneurysm, and ATAA with ascending with root aneurysm. In Fig. 4, we show the UMAP embedding of the original data (top panel) and GenePT-s embeddings (bottom panel), coloured by patient phenotype (left panel), annotated cell types (middle panel) and patient identity (right panel). While the original data grouped primarily by the patient identity and displayed distinct clusters for even identical cell types, GenePT-s embeddings clustered primarily by cell types as well as disease phenotypes. In particular, GenePT-s embeddings were able to distinguish the cells collected from patients with ascending only aortic aneurysm from aortic aneurysm

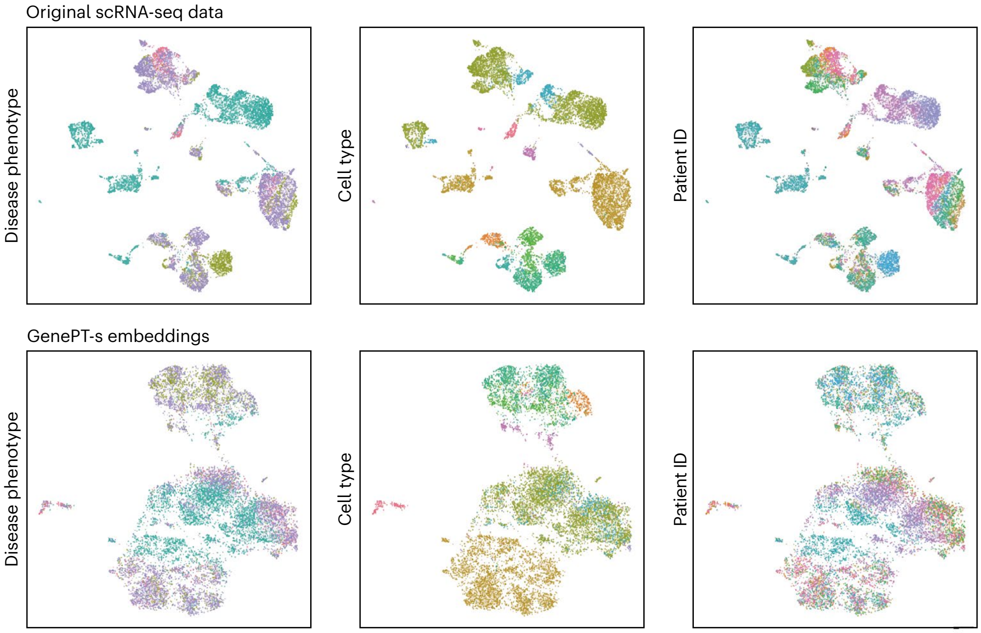


Fig. 4 | GenePT-s embeddings lead to better batch integration while preserving biological information for scRNA-seq data. Top: UMAP visualization of the subsampled aorta dataset, coloured by three different disease phenotypes (left), annotated cell types (middle) and patient identity (right), as provided in the original study[37]. Bottom: UMAP visualization of GenePT-s embeddings for the same set of cells, coloured by disease phenotype (left), annotated cell types (middle) and patient identity (right).

that includes the root (represented using green and purple points in the leftmost column in Fig. 4).

Clustering analysis confirmed the visual display: the ARI between patient identity and the estimated k-means clusters ($k = 11$) on the original scRNA-seq data is 0.24 versus 0.11, 0.10 and 0.18 when using Geneformer, GenePT-s and scGPT, respectively. We also evaluated the agreement between the phenotype labels (three ATAA subtypes and one control) and the clusters derived from embeddings and original scRNA-seq data. The resulting ARIs are 0.12, 0.11, 0.12 and 0.12 for Geneformer embeddings, GenePT-s embeddings, scGPT embeddings and scRNA-seq data, respectively. These findings suggest that GenePT-s, Geneformer and scGPT all exhibit some degree of robustness against batch effects while preserving information on the disease phenotype. This is corroborated by training a logistic regression model to predict the phenotypes: on the randomly held-out 20% test set, GenePT-s yields an accuracy of 73% (68% precision and 74% recall), similar to that of scGPT (75% accuracy, 75% precision and 75% recall) and moderately better than that of Geneformer (69% accuracy, 68% precision and 69% recall).

Discussion

With the advance of technologies to measure genetic and cellular functionalities at scale, embedding models have emerged as an attractive approach to make sense of the underlying biology. In this work, we introduced GenePT, a simple yet effective embedding model that leverages natural language models such as GPT-3.5 to represent genes and cells by utilizing their text summaries and ranked expression values, respectively. Across various contexts, including discerning gene functionality groups and predicting GGIs, this straightforward approach proves to be very effective even compared with state-of-the-art foundational models trained on large-scale single-cell transcriptomics data. Our work underscores the potential of complementing those specially crafted foundational models with a simple, natural language-guided representation, which could be substantially more resource and data efficient.

It is important to note the limitations in our work, primarily because the current GenePT framework only makes use of available gene summaries and descriptions. This may overlook the intricacies of lesser-known functionalities and new biological insights not documented in databases such as NCBI. Furthermore, unlike the embeddings trained on expression data, GenePT embeddings might not be optimal for specific tissues and cell types, as they are derived from pretrained LLMs. This may pose challenges in capturing the dynamic and context-dependent roles of genes and cells within those settings. Lastly, the effectiveness of the embeddings is constrained by the language models used, that is, GPT-3.5. While we demonstrated that fine-tuning is an effective strategy for the GenePT cell embeddings, exploring opportunities to fine-tune the underlying language model directly (for example, to generate LLama-7B-based context-specific gene embeddings) could further enhance the proposed approach.

Several promising pathways lie ahead for future research. First, extending the current GenePT approach to be more dynamic and context dependent could enhance its utility in real-world applications. In addition, investigating ways to integrate different embeddings across various modalities and models, as well as improving the simple representation of ordered gene names in GenePT-s to obtain more biologically meaningful representations, could further enhance the usability of embedding models in single-cell biology. In this vein, we recommend using GenePT-w as it more naturally integrates gene-expression information and outperforms GenePT-s across most of the tasks. Moreover, it is natural to investigate the performance of GenePT in additional downstream tasks, such as perturbation predictions and drug–gene interactions. Lastly, while this paper primarily focuses on gene and cell embeddings, it would be of great interest to explore whether the approach of leveraging the natural language descriptions with LLMs embedding could be applied to other biological domains and challenges, such as protein sequence modelling[38], genome-wide association studies and other computational biology tasks[23,39,40–44].

Methods

Data collection and transformation

To obtain embeddings for genes most pertinent to single-cell transcriptomics studies, we started by unifying the list of gene names provided in Geneformer[2] and scGPT[3]. The selection of these genes was informed by their expression levels across the pretraining datasets. In Geneformer cases, the genes were represented as Ensembl IDs rather than gene names, and we used the `mygene` package[45] for conversion, retaining in successful look-up of more than 90% of the Ensembl IDs. In addition, we incorporated genes detected in our downstream application datasets, totalling around 33,000 genes. For each gene, we extracted its information from the NCBI gene database's summary section after removing hyperlinks and date information. GPT-3.5 (`text-embedding-ada-002`) embeddings were obtained for the summaries for each gene (mean, 73 words; interquartile range, 25–116). Each embedding has a dimension of 1,536, which serves as a gene representation. Moreover, we mapped around 60,000 additional gene name aliases to the NCBI summary embedding using the HGNC database[46]. We conducted sensitivity analyses using four different sources of content input for gene summaries (gene names only, gene names with gene summaries, all summary card information and ChatGPT-generated gene summaries).

In addition to embedding the gene summaries using GPT-3.5, we conducted comparisons with alternative embedding methods, such as gene summary embeddings using the open-source biomedical language models such as BioLinkBert[47] and gene-expression-derived embeddings such as Gene2vec[20] and Geneformer[2].

To encode information at the cellular level, we developed two distinct approaches: GenePT-w (w for weighted) and GenePT-s (s for sentence). In both approaches, we first normalize and transform the scRNA-seq data as implemented in the `scanpy` package as follows: first, we row-normalize the count matrix so that each cell has 10,000 observed RNA transcripts, followed by a $\log(1+x)$ transformation of each matrix entry.

To construct GenePT-w embeddings, we first take a weighted average of the GenePT gene embeddings, where the weight is determined by the normalized expressions of each gene, and then normalize the embedding to have a unit ℓ_2 norm. This approach leverages the rich context of each gene embedding but is limited by the simplicity of the weighted average. As an alternative, inspired by the representation of cells using genes ordered by expression values in recent work[2,3], we represent cells using natural language sentences by creating a sequence of gene names, ordered by descending normalized expression levels (omitting genes with zero counts). We then pass this sentence representation for each cell to GPT-3.5 to obtain GenePT-s embeddings. While GenePT-s does not directly leverage gene-expression information, other than through gene ordering, empirical experiments with this approach could provide insight into the effectiveness of such gene ordering approaches.

Downstream gene-level and cell-level applications

Geneformer and scGPT demonstrate the biological knowledge encoded in the models using several downstream gene-level and cell-level tasks. In this paper, we evaluated the performance of GenePT on the same downstream applications wherever possible to compare GenePT with other embedding models for genes and cells, such as Geneformer, Gene2vec and scGPT.

Gene-level tasks.

- Gene functionality class prediction: this is a multi-class prediction challenge based on the 15 most common functional gene classes. Labels for these classes were curated as part of the Geneformer paper.
- Gene property prediction task: this involves four binary classification tasks using open-source data provided in Theodoris et al.[2]—distinguishing previously identified dosage-sensitive from dosage-insensitive TFs; differentiating between bivalent and non-methylated genes; differentiating between Lys4-only-methylated and non-methylated genes; and distinguishing long-range from short-range TFs.
- GGI prediction: we utilized a benchmark for GGI based on shared Gene Ontology annotations published by Du et al.[20]. The training and test datasets include over 200,000 pairs of examples in the tuple (gene 1, gene 2, label), where the binary label indicates whether a pair of genes is known to interact.
- PPI prediction: we assessed the ability to predict PPIs using GenePT embeddings with the following three datasets—(i) the human binary protein interactions (HuRI) dataset collected by Luck et al.[48] through screening with multiple PPI assays; (ii) comprehensive binary PPIs (Lit-BM) that are supported by at least two traceable pieces of evidence[49]; and (iii) tissue-specific protein–protein functional interaction networks derived by Greene et al.[50]. These PPI datasets contain tuples in the form of (protein 1, protein 2, binary label). The binary label indicates whether there is an observed interaction between the two proteins. We first converted the proteome identifiers for proteins into gene names using the UniProt conversion tool[51]. If multiple genes were returned, we randomly selected one. Since only positive interactions were reported in the HuRI and Lit-BM dataset, we constructed an equal amount of negative data by randomly sampling pairs of proteins examined in Luck et al.[48] that were not reported as interacting pairs. We explored the potential utility of creating context-dependent embeddings by providing context-dependent gene descriptions for PPI tasks in Supplementary Information.
- Unsupervised exploration of gene programmes: to examine the interaction between genes, we constructed a similarity network of GGIs using GenePT embeddings from a dataset of human immune tissues[52]. Our validation process follows that of Cui et al.[3] and consists of the following steps: (1) constructing gene networks based on the cosine similarities among the highly variable genes; (2) applying unsupervised Louvain clustering[53] to derive gene programmes; and (3) qualitatively comparing the trends of highlighted gene programmes with their cell-specific expression levels.

Cell-level tasks.

- Assessing association between embeddings and the underlying cell states: here we considered the following test datasets representing cells from circulatory systems (aorta, a random 20% subset of data originally published in Li et al.[37] comprising 11 cell types, and artery[54] with 10 cell types), bone tissues (bones[55] with 7 cell types; myeloid[56] containing 3 annotated cancer types and 11 cell types across 13,468 cells), the pancreas[52] (containing 11 annotated cell types across 4,218 cells) and immune cells collected from healthy individuals and patients with multiple sclerosis[57], totalling 18 annotated cell types and 12 donors across 3,430 cells. For each dataset and its associated metadata annotation, we applied *k*-means clustering on the pretrained GenePT, Geneformer or the scGPT embeddings to obtain clusters matching the classes in the metadata annotations. We select the number of clusters *k* to match the number of classes in the metadata annotation. We then computed the ARI and AMI to evaluate the concordance between derived cluster labels and the true metadata labels. A higher alignment, indicated by higher values of ARI or AMI, between the inferred and actual labels, suggests that the embedding captures more biological structure and signals. We also calculated the ASW using the true annotations of original samples to assess the cohesion and separation of the clusters.
- Disease phenotype prediction: we aimed to predict phenotypic features in cardiomyocytes (NF, non-failing heart; HCM,

hearts with hypertrophic cardiomyopathy; DCM, hearts with dilated cardiomyopathy)[33] and in lupus (disease or health) datasets[58]. The results were reported at the cell level for the cardiomyopathy dataset and at the patient level for the lupus dataset. We applied an ℓ_2-regularized logistic regression on pretrained embeddings from scGPT, Geneformer, GenePT-w and fine-tuned GenePT-w.

- Enhancing GenePT embeddings through fine-tuning: we investigated whether our pretrained embeddings could be fine-tuned on specific downstream tasks to enhance prediction performance, a technique previously successful in single-cell foundational models. To fine-tune the GenePT-w embeddings, we trained a two-layer MLP (multilayer perceptron) with leaky ReLU activations. Using disease phenotype and cell-type prediction as examples, we demonstrated that GenePT-w embeddings, efficiently generated from pretrained text-based gene embeddings, can also be fine-tuned to boost performance improvements in downstream tasks.
- Context awareness and batch integration: pretrained single-cell embedding models have shown to be robust against common batch-dependent technical artefacts while still encoding the underlying biological context. We assessed whether GenePT-s embeddings were impacted by common batch effects such as patient variability on two datasets used in Theodoris et al.[2]: the cardiomyocyte dataset originally published by Chaffin et al.[33] and the aforementioned aorta dataset.

Reporting summary

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

Data availability

All datasets used in the study have been previously published with pointers provided at https://github.com/yiqunchen/GenePT; embeddings generated in this work can be accessed at https://zenodo.org/records/10833191.

Code availability

The source code is available at https://github.com/yiqunchen/GenePT.

References

1. Yang, F. et al. scBERT as a large-scale pretrained deep language model for cell type annotation of single-cell RNA-seq data. *Nat. Mach. Intell.* **4**, 852–866 (2022).
2. Theodoris, C. V. et al. Transfer learning enables predictions in network biology. *Nature* **618**, 616–624 (2023).
3. Cui, H., Wang, C., Maan, H. & Wang, B. scGPT: towards building a foundation model for single-cell multi-omics using generative AI. *Nat. Methods* **21**, 1470–1480 (2024).
4. Vaswani, A. et al. Attention is all you need. In *Proc. 31st International Conference on Neural Information Processing Systems* Vol. 30 (eds Guyon, I. et al.) 6000–6010 (Curran Associates, 2017).
5. OpenAI. New and improved embedding model. https://openai.com/blog/new-and-improved-embedding-model (2023).
6. OpenAI. GPT-4 technical report. Preprint at https://arxiv.org/abs/2303.08774 (2023).
7. Chen, Q. et al. A comprehensive benchmark study on biomedical text generation and mining with ChatGPT. Preprint at *bioRxiv* https://doi.org/10.1101/2023.04.19.537463 (2023).
8. Biswas, S. S. Role of ChatGPT in public health. *Ann. Biomed. Eng.* **51**, 868–869 (2023).
9. Ayers, J. W. et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA Intern. Med.* **183**, 589–596 (2023).
10. Strong, E. et al. Chatbot vs medical student performance on Free-Response clinical reasoning examinations. *JAMA Intern. Med.* **183**, 1028–1030 (2023).
11. Bommasani, R. et al. On the opportunities and risks of foundation models. Preprint at https://arxiv.org/abs/2108.07258 (2021).
12. Connell, W., Khan, U. & Keiser, M. J. A single-cell gene expression language model. Preprint at https://arxiv.org/abs/2210.14330 (2022).
13. Chen, J. et al. Transformer for one stop interpretable cell type annotation. *Nat. Commun.* **14**, 223 (2023).
14. Hao, M. et al. Large scale foundation model on single-cell transcriptomics. *Nat. Methods* **21**, 1481–1491 (2024).
15. Lopez, R., Regier, J., Cole, M. B., Jordan, M. I. & Yosef, N. Deep generative modeling for single-cell transcriptomics. *Nat. Methods* **15**, 1053–1058 (2018).
16. Lotfollahi, M., Wolf, F. A. & Theis, F. J. scGen predicts single-cell perturbation responses. *Nat. Methods* **16**, 715–721 (2019).
17. Clough, E. & Barrett, T. The Gene Expression Omnibus database. *Methods Mol. Biol.* **1418**, 93–110 (2016).
18. Cellxgene Data Portal. https://cellxgene.cziscience.com/docs/08__Cite%20cellxgene%20in%20your%20publications (2023).
19. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proc. 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* Vol. 1 (eds Burstein, J. et al.) 4171–4186 (Association for Computational Linguistics, 2019).
20. Du, J. et al. Gene2vec: distributed representation of genes based on co-expression. *BMC Genom.* **20**, 82 (2019).
21. Duong, D., Ahmad, W. U., Eskin, E., Chang, K.-W. & Li, J. J. Word and sentence embedding tools to measure semantic similarity of Gene Ontology terms by their definitions. *J. Comput. Biol.* **26**, 38–52 (2019).
22. Chen, Q. et al. BioConceptVec: creating and evaluating literature-based biomedical concept embeddings on a large scale. *PLoS Comput. Biol.* **16**, 1007617 (2020).
23. Hou, W. & Ji, Z. Assessing GPT-4 for cell type annotation in single-cell RNA-seq analysis. *Nat. Methods* **21**, 1462–1465 (2024).
24. Wysocki, O. et al. Transformers and the representation of biomedical background knowledge. *Comput. Linguist.* **49**, 73–115 (2023).
25. Ye, R., Zhang, C., Wang, R., Xu, S. & Zhang, Y. Natural language is all a graph needs. In *Findings of the Association for Computational Linguistics: EACL 2024* (eds Graham, Y. & Purver, M.) 1955–1973 (Association for Computational Linguistics, 2024).
26. Sayers, E. W. et al. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **47**, 23–28 (2019).
27. Levine, D. et al. Cell2Sentence: teaching large language models the language of biology. In *Proc. 41st International Conference on Machine Learning (ICML 2024)* (PMLR, 2024)
28. Brown, G. R. et al. Gene: a gene-centered information resource at NCBI. *Nucleic Acids Res.* **43**, 36–42 (2015).
29. Pedregosa, F. et al. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
30. Bruford, E. A. et al. Guidelines for human gene nomenclature. *Nat. Genet.* **52**, 754–758 (2020).
31. Microsoft Research AI4Science & Microsoft Azure Quantum. The impact of large language models on scientific discovery: a preliminary study using GPT-4. Preprint at https://arxiv.org/abs/2311.07361 (2023).
32. Touvron, H. et al. LLama: open and efficient foundation language models. Preprint at https://arxiv.org/abs/2302.13971 (2023).
33. Chaffin, M. et al. Single-nucleus profiling of human dilated and hypertrophic cardiomyopathy. *Nature* **608**, 174–180 (2022).

34. He, B. et al. Cloudpred: predicting patient phenotypes from single-cell RNA-seq. In *Proc. Pacific Symposium on Biocomputing 2022* 337–348 (2021).
35. Marian, A. J. & Braunwald, E. Hypertrophic cardiomyopathy: genetics, pathogenesis, clinical manifestations, diagnosis, and therapy. *Circ. Res.* **121**, 749–770 (2017).
36. Son, M., Kim, S. J. & Diamond, B. SLE-associated risk factors affect DC function. *Immunol. Rev.* **269**, 100–117 (2016).
37. 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).
38. Rives, A. et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc. Natl Acad. Sci. USA* **118**, 2016239118 (2021).
39. Visscher, P. M., Brown, M. A., McCarthy, M. I. & Yang, J. Five years of GWAS discovery. *Am. J. Hum. Genet.* **90**, 7–24 (2012).
40. Lubiana, T. et al. Ten quick tips for harnessing the power of ChatGPT in computational biology. *PLoS Comput. Biol.* **19**, 1011319 (2023).
41. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.* **19**, 1–5 (2018).
42. Pliner, H. A., Shendure, J. & Trapnell, C. Supervised classification enables rapid annotation of cell atlases. *Nat. Methods* **16**, 983–986 (2019).
43. Pasquini, G., Arias, J. E. R., Schäfer, P. & Busskamp, V. Automated methods for cell type annotation on scRNA-seq data. *Comput. Struct. Biotechnol. J.* **19**, 961–969 (2021).
44. Traag, V. A., Waltman, L. & Van Eck, N. J. From Louvain to Leiden: guaranteeing well-connected communities. *Sci. Rep.* **9**, 5233 (2019).
45. Welcome to MyGene.py's documentation!—MyGene.py v3.1.0 documentation. https://docs.mygene.info/projects/mygene-py/en/latest/ (2023).
46. Seal, R. L. et al. Genenames.org: the HGNC resources in 2023. *Nucleic Acids Res.* **51**, 1003–1009 (2023).
47. Yasunaga, M., Leskovec, J. & Liang, P. LinkBERT: Pretraining language models with document links. In *Proc. 60th Annual Meeting of the Association for Computational Linguistics* Vol. 1 (eds Muresan, S. et al.) 8003–8016 (Association for Computational Linguistics, 2022).
48. Luck, K. et al. A reference map of the human binary protein interactome. *Nature* **580**, 402–408 (2020).
49. Rolland, T. et al. A proteome-scale map of the human interactome network. *Cell* **159**, 1212–1226 (2014).
50. Greene, C. S. et al. Understanding multicellular function and disease with human tissue-specific networks. *Nat. Genet.* **47**, 569–576 (2015).
51. UniProt Consortium. UniProt: the universal protein knowledgebase in 2023. *Nucleic Acids Res.* **51**, 523–531 (2023).
52. Luecken, M. D. et al. Benchmarking atlas-level data integration in single-cell genomics. *Nat. Methods* **19**, 41–50 (2022).
53. Blondel, V. D., Guillaume, J.-L., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. *J. Stat. Mech. Theory Exp.* **2008**, 10008 (2008).
54. Alsaigh, T., Evans, D., Frankel, D. & Torkamani, A. Decoding the transcriptome of calcified atherosclerotic plaque at single-cell resolution. *Commun. Biol.* **5**, 1084 (2022).
55. Chou, C.-H. et al. Synovial cell cross-talk with cartilage plays a major role in the pathogenesis of osteoarthritis. *Sci. Rep.* **10**, 10868 (2020).
56. Cheng, S. et al. A pan-cancer single-cell transcriptional atlas of tumor infiltrating myeloid cells. *Cell* **184**, 792–80923 (2021).
57. Schirmer, L. et al. Neuronal vulnerability and multilineage diversity in multiple sclerosis. *Nature* **573**, 75–82 (2019).
58. Subramaniam, M. *Implementing and Applying Multiplexed Single Cell RNA-sequencing to Reveal Context-specific Effects in Systemic Lupus Erythematosus*. PhD thesis, UCSF (2019).

Acknowledgements

Y.C. discloses support for the research described in this study from Stanford Data Science Postdoctoral Fellowship. J.Z. discloses support for the research described in this study from the Chan-Zuckerberg Initiative.

Author contributions

Y.C. led the analysis and J.Z. supervised the project. Both authors contributed to writing the paper and have approved the submitted version.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41551-024-01284-6.

Correspondence and requests for materials should be addressed to James Zou.

Peer review information *Nature Biomedical Engineering* thanks Jian Ma and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

nature portfolio

Corresponding author(s): James Zou

Last updated by author(s): Oct 8, 2024

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 (v3.2.2): https://pypi.org/project/mygene/3.2.2
requests (v2.28.1): https://pypi.org/project/requests/2.28.1
beautifulsoup4 (v4.11.1): https://pypi.org/project/beautifulsoup4/4.11.1
The code to collect raw data is available as an ipython notebook at https://github.com/yiqunchen/GenePT/blob/main/request_ncbi_text_for_genes.ipynb

Data analysis

anndata (v0.9.1): https://pypi.org/project/anndata/0.9.1
numpy (v1.24.4): https://pypi.org/project/numpy/1.24.4
pandas (v1.4.4): https://pypi.org/project/pandas/1.4.4
scipy (v1.10.0): https://pypi.org/project/scipy/1.10.0
umap-learn (v0.5.3): https://pypi.org/project/umap-learn/0.5.3
openai (v0.28.0): https://pypi.org/project/openai/0.28.0
matplotlib (v3.8.0): https://pypi.org/project/matplotlib/3.8.0
scikit-learn (v1.1.3): https://pypi.org/project/scikit-learn/1.1.3
scanpy (v1.9.3): https://pypi.org/project/scanpy/1.9.3
hnswlib (v0.7.0): https://pypi.org/project/hnswlib
Python (v3.10.4): https://www.python.org/downloads/release/python-3104
geneformer (v0.0.1): https://huggingface.co/ctheodoris/Geneformer
scgpt (v0.1.7): https://github.com/bowang-lab/scGPT

We provide analysis code for GenePT at the github repository https://github.com/yiqunchen/GenePT. Gene embeddings have also been deposited at https://zenodo.org/records/10030426. All single-cell and gene datasets used in GenePT have been previously published and we provide links at https://github.com/yiqunchen/GenePT/tree/main#datasets-used-in-genept.

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

Data

Policy information about availability of data

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

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

All datasets used in the study have been previously published, with pointers provided at https://github.com/yiqunchen/GenePT; embeddings generated in this work can be accessed at https://zenodo.org/records/10833191.

We make use of the following datasets in the manuscript, all of which are publicly available:

1. Gene-level tasks data can be accessed from https://huggingface.co/datasets/ctheodoris/Genecorpus-30M/tree/main/example_input_files/gene_classification

2. Gene-gene interaction network datasets are available at https://github.com/jingcheng-du/Gene2vec/tree/master/predictionData.

3. Protein-protein interaction datasets are available at https://github.com/CCSB-DFCI/HuRI_paper and https://proteinformatics.uni-leipzig.de/ProteinPrompt/download-data.

4. Pancreas, Myeloid, Multiple Sclerosis, Bones and Artery datasets are available at the following links:
- Multiple Sclerosis (M.S.) dataset: https://drive.google.com/drive/folders/1Qd42YNabzyr2pWt9xoY4cVMTAxsNBt4v?usp=sharing.
- Myeloid (Mye.) dataset: https://drive.google.com/drive/folders/1VbpApQufZq8efFGakW3y8QDDpY9MBoDS?usp=drive_link.
- Pancreas dataset: https://drive.google.com/drive/folders/1s9XjcSiPC-FYV3VeHrEa7SeZetrthQVV?usp=drive_link.
- Artery (GSE159677)
- Bone (GSE152805)

5. Cardiomyocyte dataset:
- Original data can be downloaded from https://singlecell.broadinstitute.org/single_cell/study/SCP1303/single-nuclei-profiling-of-human-dilated-and-hypertrophic-cardiomyopathy.
- We created a random 10% subset of the original dataset for our cell-level analysis, available at https://drive.google.com/drive/folders/1LgFvJqWNq9BqHbuxB2tYf62kXs9KqL4t?usp=share_link.

6. Aorta dataset:
- Original data have been deposited at https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE155468, and we used the additional cell-type annotation provided by the authors on https://github.com/LI-Yan-Ming/Circulation.-2020-142-1374-1388/blob/main/meta_addsubcluster_cellcycle.csv.
- We created a 20% random subset of the original dataset at https://drive.google.com/drive/folders/1LgFvJqWNq9BqHbuxB2tYf62kXs9KqL4t?usp=share_link.

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	Data used in this work have all been previously published, and we adhered to the reported sex from the original publications.
Reporting on race, ethnicity, or other socially relevant groupings	–
Population characteristics	Data used in this work have all been previously published, and we adhered to the reported population characteristics (such as age and disease status) from the original publications.
Recruitment	–
Ethics oversight	–

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

Field-specific reporting

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

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

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

Life sciences study design

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

Sample size	All data used for analysis in the manuscript are publicly available, and sample sizes are mentioned in Methods for each dataset.
Data exclusions	We did not perform additional exclusions, except for randomly selecting a subset of cells, on the original dataset.
Replication	For the accuracy evaluations, replicates were performed as part of cross-validation. We also included more than one cell type and gene type for cell-classification and gene-classification tasks, respectively.
Randomization	For gene-classification and cell-classification tasks, labelled genes and cells were randomly split into a training set (80% or 70%) and a test set (20% or 30%). Random splits were also done for five-fold cross-validation.
Blinding	Blinding was not relevant to the study, as our assessments are based on quantitative metrics such as accuracies, precision and recall.

Reporting for specific materials, systems and methods

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

Materials & experimental systems

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

Methods

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