GenePT: A Simple But Hard-to-Beat Foundation Model for Genes and Cells Built From ChatGPT

Yiqun T. Chen
Department of Biomedical Data Science
Stanford University
Stanford, CA 94305
yiqunc@stanford.edu

James Zou
Departments of Biomedical Data Science;
Electrical Engineering; and Computer Science
Stanford University
Stanford, CA 94305
jamesz@stanford.edu

Abstract

There has been significant recent progress in leveraging large-scale gene expression data to develop foundation models for single-cell transcriptomes such as Geneformer [1], scGPT [2], and scBERT [3]. These models infer gene functions and interrelations from the gene expression profiles of millions of cells, which requires extensive data curation and resource-intensive training. Here, we explore a much simpler alternative by leveraging ChatGPT embeddings of genes based on literature. Our proposal, GenePT, uses NCBI text descriptions of individual genes with GPT-3.5 to generate gene embeddings. From there, GenePT generates single-cell embeddings in two ways: (i) by averaging the gene embeddings, weighted by each gene's expression level; or (ii) by creating a sentence embedding for each cell, using gene names ordered by the expression level. Without the need for dataset curation and additional pretraining, GenePT is efficient and easy to use. On many downstream tasks used to evaluate recent single-cell foundation models — e.g., classifying gene properties and cell types — GenePT achieves comparable, and often better, performance than Geneformer and other methods. GenePT demonstrates that large language model embedding of literature is a simple and effective path for biological foundation models.

1 Introduction

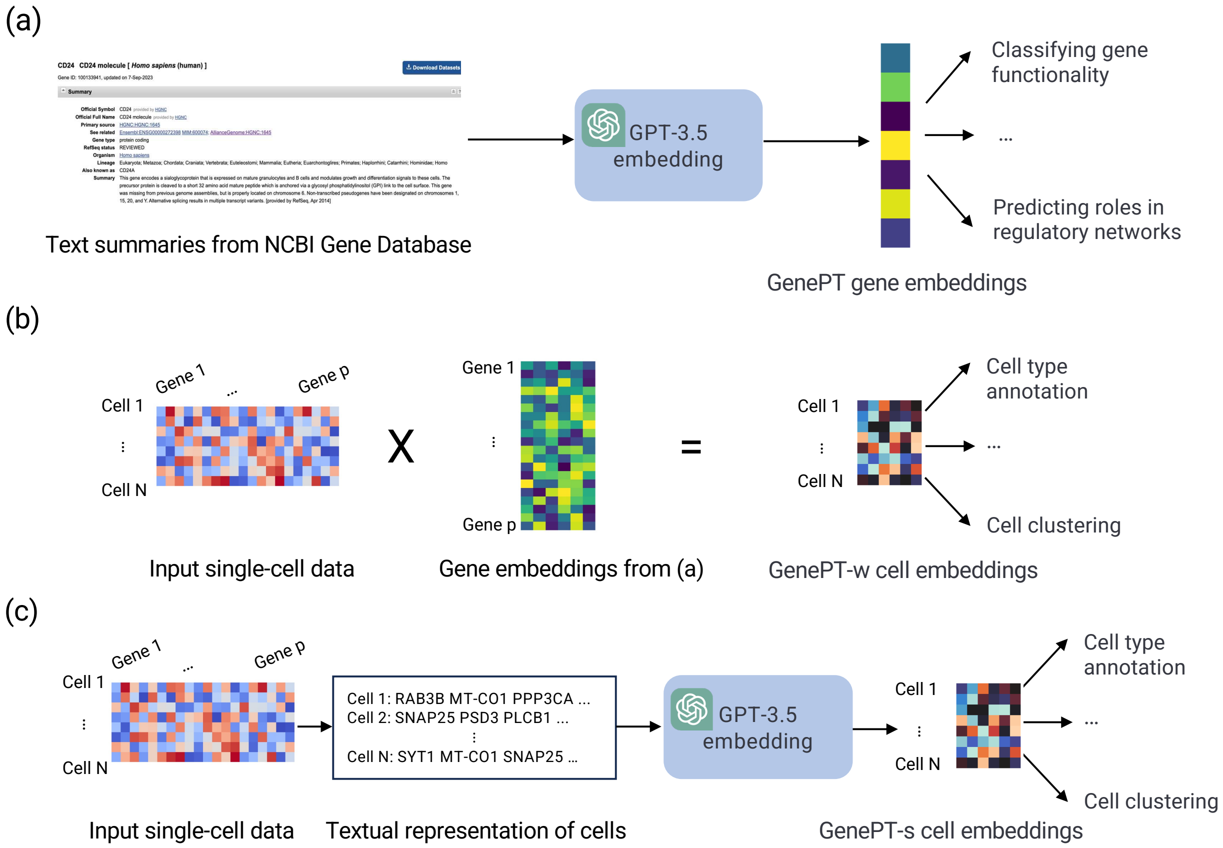
Recently, the field of single-cell biology has seen a surge in interests and efforts to develop foundation models, i.e., models designed to learn embeddings of genes and cells to facilitate various downstream analyses. Several methods, such as Geneformer [1] and scGPT [2], have been recently proposed 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 extensive single-cell gene expression datasets for pre-training the model in a self-supervised manner (e.g., 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, the computational power and time required to collect and process large-scale single-cell transcriptomics data used for pre-training (Step 2 above) can be prohibitive, particularly when researchers desire early signal detection and rapid iterations. Furthermore, the signals from extracted embeddings are heavily dependent on the gene expression data used in Step 2, which doesn't 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 explored an alternative, complementary approach and investigated the feasibility of encoding the biology of genes and cells using natural language.

The intuition for our approach is as follows: large-language models (LLMs) such as GPT-3.5 and GPT-4 have been trained on extensive text corpus [5], including biomedical literature, and have demonstrated remarkable ability in understanding, reasoning, and even generating biomedical text [6–9]. Consequently, we hypothesize that LLM-derived embeddings of gene summaries and functionalities — which often are curated from a broad spectrum of experiments and studies — might more directly capture the underlying biology.

Our contributions: We introduced GenePT — a method that represents genes and cells by utilizing OpenAI's ChatGPT text embedding API services [10]. We evaluated the generated embeddings on several biologically driven tasks and our findings reveal that our proposal exhibits performance comparable to, and sometime surpassing, specially designed models such as Geneformer across a diverse set of downstream tasks. GenePT offers several advantages to single-cell RNA-seq based foundation models: (i) it performs better on several biological tasks; (ii) it doesn't require expensive single-cell curation and additional pretraining; and (iii) it's very simple to use and to generate gene and cell embeddings. GenePT uses LLM-based embeddings which is an orthogonal source of information compared to the expression based representations; this suggests a promising new direction of combining these two ideas.



CD24 CD24 molecule [*Homo sapiens* (human)]

Gene ID: 100133941, updated on 7-Sep-2023

Summary

Official Symbol	CD24 provided by HGNC
Official Full Name	CD24 molecule provided by HGNC
Primary source	HGNC:HGNC:1645
See related	Ensembl:ENSG00000272398 MIM:600074; AllianceGenome:HGNC:1645
Gene type	protein coding
RefSeq status	REVIEWED
Organism	Homo sapiens
Lineage	Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; Homo
Also known as	CD24A
Summary	This gene encodes a sialoglycoprotein that is expressed on mature granulocytes and B cells and modulates growth and differentiation signals to these cells. The precursor protein is cleaved to a short 32 amino acid mature peptide which is anchored via a glycosyl phosphatidylinositol (GPI) link to the cell surface. This gene was missing from previous genome assemblies, but is properly located on chromosome 6. Non-transcribed pseudogenes have been designated on chromosomes 1, 15, 20, and Y. Alternative splicing results in multiple transcript variants. [provided by RefSeq, Apr 2014]

Figure 1: **An overview of the GenePT framework.** **(a)** For each gene, we source its corresponding functional summary from NCBI and use GPT-3.5 text embedding as its representation. **(b)** In GenePT-w cell embeddings framework, we use the average of the gene embeddings obtained from step (a), each weighted by its respective normalized expression level in that cell. **(c)** In GenePT-s cell embeddings framework, each cell from the input single-cell data is translated into natural language sentence based on ranked gene expressions, and the GPT-3.5 embedding of the entire sentence is used to represent the cell.

The remainder of the paper is structured as follows: Section 2 reviews pertinent literature on LLMs tailored for transcriptomics and on probing approaches in language modeling. Section 3 details our process for data collection, embedding, and analysis. Section 4 provides empirical findings that underscore the efficacy of GenePT embeddings for genes and cells. Section 5 discusses potential limitations and chart out avenues for future research.

2 Related Work

Foundation models for single-cell transcriptomics: Foundation models have shown unprecedented performance for a myriad of tasks including text classification, question answering, and text generation [11]. Efforts have naturally been made to adapt these models to tackle tasks in biology, especially in the field single-cell transcriptomics [1, 2, 12]. Examples for such efforts include cell type annotation, where a cell is labeled based on its biological identity [3, 13]); gene functional and regulatory network inference, where the functionality of individual genes and clustered gene groups are examined [2, 14]; and sample integration [13], which accounts for transcript abundance influenced primarily by technical replicate noise instead of underlying biology.

With the advent of large-scale, open-source expression datasets such as Gene Expression Omnibus [15] and the Human Cell Atlas [16], several models have been trained on such data. The aspiration behind these models is to craft a foundational model for single-cell transcriptomics, analogous to foundational models in natural language processing. These models are intended to display broad capabilities across an array of biological tasks rather than just a niche subset. For instance, Geneformer [1] employs extensive pretraining on the ranks of gene expression levels through masked token prediction across 30 million cells collected from a wide range of sources using a transformer architecture. It shows good performance in tasks ranging from understanding network dynamics to deciphering network hierarchy. Another noteworthy model is scGPT [2]: it hinges on generative pretraining (with gene expression prediction as the task) and used 33 million cells from the CELLxGENE collection for training [17]. Its capabilities are demonstrated through downstream evaluations in perturbation prediction, batch integration, and cell type annotation.

Using LLMs for cell biology: Beyond the large-scale models tailored for structured, non-linguistic data, recent initiatives have explored the direct manipulations of LLMs for biomedically-focused tasks. For example, Hou and Ji [18] employed ChatGPT for cell type annotation; Wysocki et al. [19] probed biomedical information on BioBERT and BioMegatron embeddings; and Ye et al. [20] utilized instruction fine-tuning to achieve competitive results on graph data task benchmarks with an LLM. While our paper is under preparation, Levine et al. [21] has 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 cell type annotation and generation of new cells conditional on cell types, with an emphasis on *generative* tasks.

Deciphering natural language embeddings: Understanding how large-scale unsupervised representations capture linguistic nuances is a central research question in Natural Language Processing (NLP). One avenue of exploration, sometimes referred to as "probes", trains supervised models to predict downstream properties from the language model embeddings [22–24]. These techniques have achieved impressive accuracy across NLP tasks, suggesting that the embeddings exhibit a substantial amount of understanding of input attributes.

The success of probing and foundational models in biology inspire our primary research questions (RQs):

RQ1: Do natural language embeddings of genes capture the intrinsic biological functionalities of a gene?

RQ2: Do natural language embeddings of cells capture the underlying biology of a cell?

In addressing these RQs, our study makes the following contribution to the literature: we show that natural language embeddings of gene functions — such as summaries readily available from sources like the NCBI gene database [25] — successfully encapsulate the underlying biological relationships and insights associated with genes, when assessed on biologically relevant tasks. Moreover, for single cells, language models embeddings of the gene names, ordered by expression levels, encode substantial biological signals that can be used, e.g., for cell type annotation.

3 Methods

3.1 Data Collection and Transformation:

To obtain embeddings for genes most pertinent to single-cell transcriptmotics studies, we began with unifying of the list of gene vocabularies utilized in Geneformer [1] and scGPT [2]. The selection of these genes was informed by their expression levels across the pretraining datasets, with detailed methodologies for gene vocabulary construction available in the cited works. In Geneformer cases, the genes were represented as Ensembl IDs rather than gene names, and we used the `mygene` package [26] for conversion, retaining in successful look up of more than 90% of the Ensembl

IDs. Additionally, we incorporated genes detected in our downstream application datasets, totaling around 33,000 genes. For each gene, we extracted its information from the NCBI gene database's summary section (see an examples in Appendix A), and removed hyperlinks and dates. Averaging 73 words (interquartile range: 25–116), the parsed gene summaries were then processed through GPT-3.5 (`text-embedding-ada-002` embedding model), resulting in embeddings of 1,536 dimensions, which served as gene representations (see Figure 1). On the rare occasion that an NCBI gene page was unavailable ($< 5\%$ of our final gene embeddings), we turned to GPT-3.5 with the prompt: *"Provide a brief overview of the functionality of gene gene name"*. In addition to embedding the gene summaries using GPT-3.5, we conducted comparisons with alternative embedding methods, such as (i) embedding the text using the open-source biomedical language model BioLinkBert [27]; and (ii) Gene2vec [28] derived from gene expression data.

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: firstly, 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 get the GenePT-w embedding, as its name suggests, we represent the cell under consideration by taking the average of the gene embeddings, weighted by the corresponding normalized expression levels, which can be efficiently implemented using matrix multiplication (see Figure 1(b)). Alternatively, instead of pooling the gene embeddings via an explicitly weighted average, we can represent cells in natural language by creating a sequence of gene names as an analog to sentences, ordered by the descending normalized expression level, with genes with zero counts omitted. We can then pass this sentence representation for each cell to GPT-3.5 to obtain GenePT-s embeddings (see Figure 1(c)).

3.2 Downstream gene-level and cell-level applications:

Geneformer and scGPT demonstrate the biological value offered by their foundation 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 Geneformer and other single-cell foundation models. In particular, for gene-level tasks, we primarily contrast our results with those from Geneformer and Gene2vec. This is because their results on the same datasets have been previously documented without the need to re-train or fine-tune. By constrast, we exclude comparisons to scGPT for gene-level tasks, as there isn't a readily available gene-level classification model built on scGPT, necessitating extensive training and fine-tuning. Regarding cell-level tasks, we leveraged *pre-trained* embeddings from all the foundational models.

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. [1]: Distinguishing previously identified dosage-sensitive from dosage-insensitive transcription factors. Differentiating between bivalent and non-methylated genes. Differentiating between Lys4-only-methylated and non-methylated genes. Distinguishing long-range from short-range transcription factors (TFs).
- Gene-Gene Interaction Datasets: We utilized a benchmark for gene-gene interaction based on shared gene ontology annotations published by Du et al. [28]. 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.
- Unsupervised Exploration of Gene Programs: To examine the interaction between genes, we constructed a similarity network of gene-gene interactions using GenePT embeddings from a dataset of human immune tissues [29]. Our validation process follows that of Cui et al. [2] 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 [30] to derive gene programs; and 3. qualitatively comparing the trends of highlighted gene programs with their cell-specific expression levels.

Cell-level tasks:

- Assessing Association Between Embeddings and Underlying Cell States: Here, we considered the following test datasets re-processed and used to demonstrate the use of scGPT and Geneformer — Myeloid [31] (containing 3 annotated cancer types and 11 cell types across 13,468 cells), Multiple Sclerosis [32] (containing 18 annotated cell types and 12 donors across 3,430 cells), hPancreas [29] (containing 11 annotated cell types across 4,218 cells), and Aorta (a random 20% susbet of data originally published in Li et al. [33] and comprise 11 cell types across 9,625 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 cluster k to match the number of classes in the metadata annotation. We then computed the Adjusted Rand Index (ARI) and Adjusted Mutual Information (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.

- Context awareness and batch integration: Pretrained single-cell foundation models have been demonstrated 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 effect such as patient variability on two datasets used in Theodoris et al. [1]: the cardiomyocyte dataset originally published by Chaffin et al. [34], and the Aorta dataset originally published in Li et al. [33].

4 Results

4.1 GenePT embeddings capture underlying gene functionality

In Figure 2(a), we display a 2D UMAP of the GenePT embeddings (using the `text-embedding-ada-002` model), for over 34,000 genes that belong to the top 15 most prevalent functional classes (see the detailed class breakdown in Table 3 in Appendix B). The UMAP reveals distinct clusters when colored by various gene functionality groups, implying that the language model embeddings are able to capture the inherent functions of the genes. We further divided the genes into a 70%/30% train/test split and evaluated the prediction accuracy of 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 commendable class-specific accuracies. with only minor misclassifications between closely related functional groups like lincRNA, lncRNA, and processed transcripts (see Figure 2(b)).

We further assessed the efficacy of GenePT embeddings in predicting gene-gene interactions (GGI) in Figure 2(c). We compared the ROC-AUC for three methods on the test GGI dataset provided in Du et al. [28]: (i) sum of the GenePT embedding of two genes with a random forest (RF) classifier (yielding an AUC of 0.84); (ii) sum of the Gene2Vec embeddings with an RF classifier (resulting in an AUC of 0.67); and (iii) sum of two random embeddings ($d = 1,536$, same as GenePT) with entries drawn from independent $\mathcal{N}(0,1)$ paired with an RF classifier, which served as a negative control (an AUC of 0.51). As shown in Figure 2(c), GenePT embeddings considerably enhance performance when compared to the Gene2Vec embeddings under the same downstream classifier. Even when leveraging a more intricate deep neural network, Du et al. [28] reported an AUC of 0.77, underscoring the competitive edge of GenePT in this task.

Next, we delved into cell-type specific activations among the GenePT-derived gene programs within human immune tissue datasets through a "zero-shot" approach. We first constructed 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 programs comprising 10 or more genes are depicted in Figure 2(d). Here, we display the average expression levels of these gene programs, stratified by cell types. The observed selective activation of these programs aligns with established biological knowledge where the identified gene sets are known to be functionally distinct and are differentially expressed across different cell types (e.g., Gene set 8 comprising of IFI families and gene set 24 comprising of CDC families). These findings underscore that GenePT-inferred gene programs effectively capture biologically pertinent functional groups.

4.2 GenePT embeddings enable accurate predictions in 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. [1]: dosage-sensitive versus dosage-insensitive 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 five-fold cross-validated ROC-AUC with either an ℓ_2 penalized logistic regression (LR) or a Random Forest (RF) classifier using default parameters from `sklearn` [35]. By contrast, Geneformer results, as reported in Theodoris et al. [1], 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, despite the fact that the latter benefits from a substantial pre-training dataset and a more intricate expressive classification head. Intriguingly, GPT-3.5 embeddings of only gene names also show high accuracies in some tasks, suggesting that the underlying language model and tokenizer for GPT-3.5 might grasp the biological significance

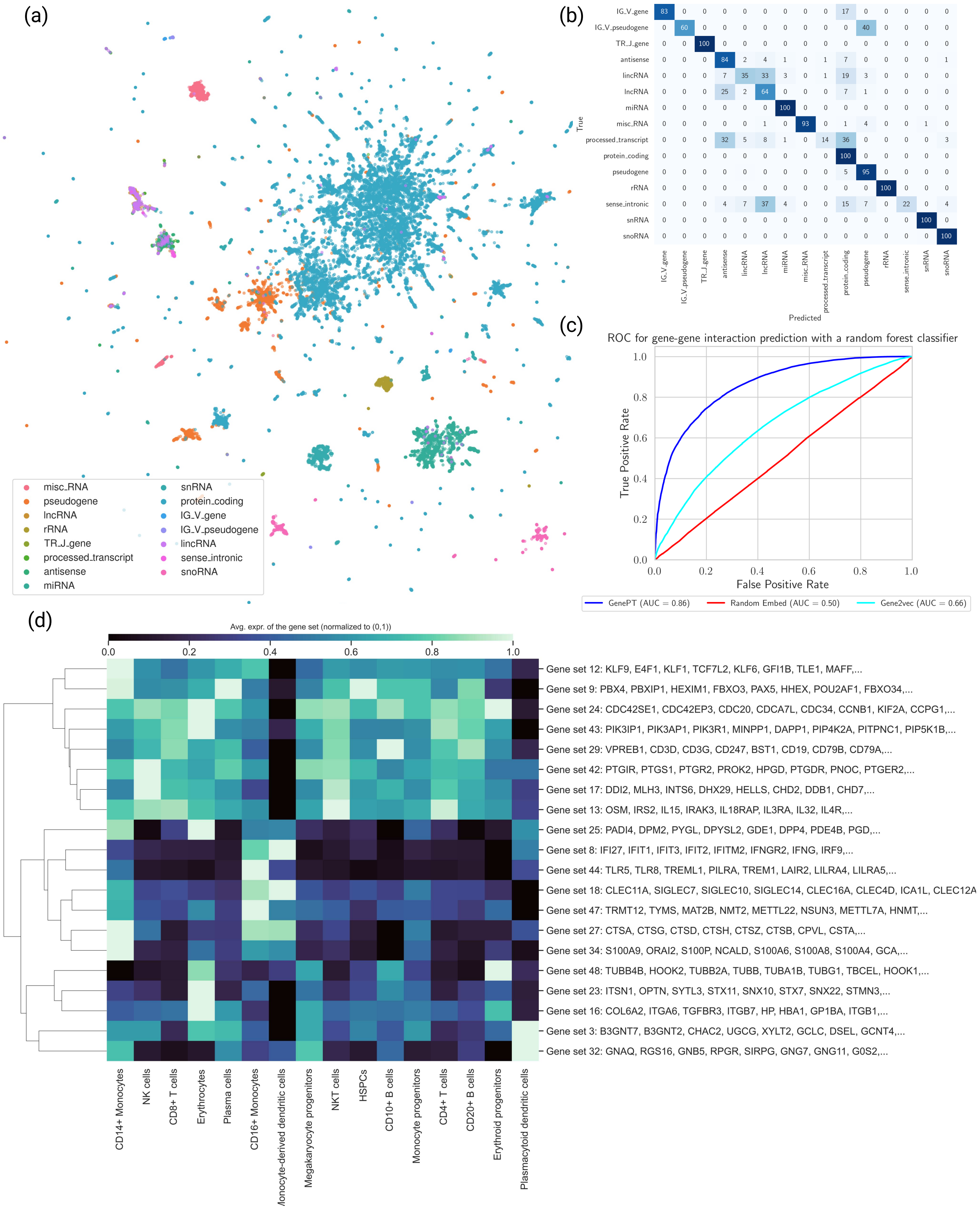


Figure 2: **GenePT embeddings encode underlying biology.** **(a)** 2D UMAP visualization of GenePT embeddings, colored by different gene functionality groups. **(b)** Confusion matrix of gene function prediction utilizing GenePT embeddings, combined with an ℓ_2-regularized logistic regression on a randomly held-out 30% test set. **(c)** Prediction accuracy on a gene-gene interaction benchmark dataset derived from GEO expression data [28] **(d)** Cell-type specific activation among GenePT-embeddings-extracted gene programs (a random subset of genes is displayed for each program) in a human immune tissue dataset [29]. The patterns of average gene expressions for identified gene programs in different cells types are congruent with those previously identified in Cui et al. [2].

of these gene symbols. Open-source embeddings like BioLinkBert and Gene2vec have slightly less competitive performance; and, as expected, random embeddings exhibit results akin to noise or blind guessing. The stark contrast between GenePT and random embeddings indicates that it's unlikely that the GenePT performance is simply due to a large embedding dimension ($d = 1536$) or the use of particularly expressive models (in our case, a simple LR or RF). In summary, these results underscore the potential of our versatile GenePT approach, which compares favorably with state-of-the-art deep learning models specifically crafted for single-cell RNA sequencing data.

Table 1: Cross-validated AUC for GenePT predictions versus alternative embeddings for downstream task of distinguishing (i) dosage-sensitive vs. insensitive transcription factors; (ii) bivalent versus non-methylated gene; (iii) bivalent versus Lys4-only methylated genes; and (iv) long-range versus short-range transcription factors (TFs). The performance for Geneformer is taken from Theodoris et al. [1] 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 sklearn, respectively.

	5-fold CV AUC ± SD			
Model	Dosage sensitivity	Bivalent vs non-methylated	Bivalent vs 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

4.3 GenePT learns representations that reflect known biology on a cell level

In this section, we focus on determining the capacity of our cell embedding approaches, as depicted in Figure 1(b)–(c), in capturing the biology underpinning selected single-cell datasets. We sought to evaluate whether the GenePT embeddings are congruent with metadata annotations across the four datasets — hPancreas, Myeloid, Multiple Sclerosis, and Aorta (detailed annotations are in Section 3; see additional pre-processing steps for the first three datasets in Cui et al. [2]).

Table 2: Assessing the association between different latent sample representations and biological annotations. Pretrained Geneformer and scGPT embeddings are used in this task. Adjusted Rand index (ARI) and adjusted mutual information (AMI) were computed between k-means clustering derived labels and true annotations of original samples.

		Geneformer		scGPT		GenePT-w		GenePT-s	
Dataset	Annotation	ARI	AMI	ARI	AMI	ARI	AMI	ARI	AMI
Myeloid	Cancer type	0.18	0.20	**0.27**	**0.29**	0.07	0.09	0.17	0.17
	Cell type	0.19	0.29	**0.44**	**0.52**	0.05	0.07	0.32	0.40
Multiple Sclerosis	Cell type	0.21	0.35	**0.29**	**0.49**	0.06	0.13	0.19	0.35
	Age	0.04	0.11	0.04	0.11	0.02	0.06	**0.06**	**0.12**
hPancreas	Cell type	0.09	0.20	0.20	0.39	0.09	0.23	**0.44**	**0.50**
Aorta	Cell type	0.20	0.33	N/A	N/A	0.15	0.24	**0.39**	**0.53**

We quantified the concordance between biological annotations (i.e., cell types, cancer types, donor ages) and k-means clustering labels inferred from: (i) pretrained Geneformer embeddings; (ii) pretrained scGPT embeddings; (iii) GenePT-w embeddings (as in Figure 1(b)); and (iv) GenePT-s embeddings (as in Figure 1(c)). We quantified the concordance using both AMI and ARI in Table 2. We see that latent representations via GenePT-s broadly outperformed both the GenePT-w and Geneformer embeddings in terms of AMI and ARI metrics and stays competitive with the scGPT embeddings (metrics for the scGPT embeddings in the `Aorta` dataset are not available because scGPT returned a code execution error): across six tasks, scGPT and GenePT-s each provide the most biological signal on three task subsets. 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, though it is widely used. Interestingly, we observe that the conceptually straightforward GenePT-w approach does not perform optimally on these datasets, indicating that cell representation may necessitate a more nuanced pooling strategy than the current expression-weighted approach illustrated in Figure 1(b). We also included additional results for a cell type annotation task via a nearest neighbor approach on these datasets in Appendix C. Similar to the findings in Table 2, GenePT-s and GenePT-w consistently outperform Geneformer in term of prediction accuracy and produce results comparable to pretrained scGPT embeddings. Interestingly, a simple ensembling of the nearest neighbors retrieved by different embeddings (GenePT-w, GenePT-s, and scGPT) enhanced the predictive performance. This suggests that natural language embeddings, such as GenePT-s, could provide complementary insights to existing expression-derived foundation models like scGPT in single-cell biology tasks.

4.4 GenePT embedding removes batch effect while preserving underlying biology

In this section, we 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 a cardiomyocyte dataset by Chaffin et al. [34] and a 20% random sample from the Aorta dataset consisting of cells in healthy and dilated aortas [33], both of which were used to demonstrate the utility of Geneformer.

In the cardiomyocyte dataset, the scientific question was to distinguish cardiomyocytes in non-failing hearts from those in hypertrophic or dilated cardiomyopathy samples. Notably, the original data exhibited significant patient batch effects (see Figure 3(b) in the Appendix). 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 adjusted Rand index (ARI) between the cell clusters and patient clusters. Higher ARI values indicate more patient-level batch effects. The original scRNA-seq data has 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.

In addition to reducing batch effects, we also investigated whether these embeddings could preserve the underlying disease phenotype (i.e., non-failing versus cardiomyopathy) of the patients from whom the cells were collected. To this end, we randomly split the cardiomyocytes into a 80%/20% train/test sets and evaluated the predictive performance using the ℓ_2-regularized logistic regression on top of the following pre-trained 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).

Next, we conducted a similar study with the Aorta dataset, collected over 11 patients (eight patients with Ascending thoracic aortic aneurysm (ATAA) and three control subjects; the eight ATAA patients are further divided into three different phenotypes: ascending only, ascending with descending thoracic aortic aneurysm, and ascending with root aneurysm). We demonstrate the use of GenePT on a random 20% sample of the original Aorta dataset. In Figure 4 (see Appendix D), we display the original data (top panel) and GenePT-s embeddings (bottom panel) using UMAP, colored by patient phenotype (left panel), annotated cell types (middle panel), and patient identity (right panel). While the original data was highly influenced by patient batch effect (see Figure 4(c)) and displayed distinct clusters for identical cell types (e.g., T cells and Mono/Maph/Dend cells in Figure 4(b)), GenePT-s embeddings clustered primarily by cell types (Figure 4(e)) as well as disease phenotype (Figure 4(d)). In particular, GenePT-s embeddings were able to distinguish the phenotype of ascending only aortic aneurysm (green points in Figure 4(d)), a different phenotype than aortic aneurysm that includes the root (purple points in Figure 4(d)).

We repeated the clustering analysis above on the Aorta dataset to get a more quantitative measure of patient-level batch effects: The Adjusted Rand Index (ARI) between patient labels and the estimated k-means clusters ($k = 11$) on the original scRNA-seq data is 0.24 versus 0.11 and 0.10 when using Geneformer and GenePT-s, 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, and 0.12 for Geneformer embeddings, GenePT-s embeddings, and scRNA-seq data, respectively. These findings suggest that both GenePT-s and Geneformer embeddings exhibit robustness against batch effects while preserving information of the disease phenotype. This is further corroborated by training a logistic regression model to predict the phenotype. GenePT-s and Geneformer yield accuracies of 73% (68% precision, 74% recall) and 69% (68% precision, 69% recall), respectively. scGPT gave errors on the Aorta dataset and could not be included in this analysis.

5 Discussion

With the advance of technologies to measure genetic and cellular functionalities, enhancing our understanding of the underlying biology through latent embedding representations has attracted much interest. In this work, we introduced GenePT, a simple yet effective approach that leverages GPT-3.5 to represent genes and cells by utilizing their text summaries and ranked expression values, respectively. Remarkably, across various contexts, including discerning gene functionality groups and predicting gene-gene interactions, this straightforward approach has proven to be quite effective even compared to state-of-the-art foundational models trained on large-scale single-cell transcriptomic 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 time-efficient. Combining LLM embeddings with expression-derived embeddings is an interesting direction of future work.

Limitations: It is important to note the limitations in our work, primarily due to the fact that the current GenePT framework only makes use of available gene summaries and descriptions. This may may overlook the intricacies of lesser-known functionalities not documented in databases like NCBI. Furthermore, unlike the embeddings trained on expression data, GenePT embeddings might not be optimal for specific tissues and cell types. This might pose challenges in capturing the dynamic and context-dependent roles of genes and cells within those settings. Lastly, the effectiveness of the embeddings is inherently constrained by the language models employed, i.e., GPT-3.5. Fine-tuning the language models could further enhance understanding of the domain-specific language prevalent in genomics.

Future work: Several promising pathways lie ahead for future research. First, extending the current GenePT approach to be more dynamic and context-dependent — such as via fine-tuning on GPT-3.5 or other open-source LLMs — could enhance its utility in real-world applications. Moreover, it's 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 applicable to other biological domains and challenges, such as protein sequence modeling [36] and Genome-Wide Association Studies [37].

Data availability

All datasets used in the study have been previously published with pointers provided at `https://github.com/yiqunchen/GenePT`.

Code availability

GenePT is available at `https://github.com/yiqunchen/GenePT`.

Acknowledgments

YC is supported by a Stanford Data Science Postdoctoral Fellowship. JZ is supported by the National Science Foundation (CCF 1763191 and CAREER 1942926), the US National Institutes of Health (P30AG059307 and U01MH098953) and grants from the Silicon Valley Foundation and the Chan-Zuckerberg Initiative.

References

[1] Christina V Theodoris, Ling Xiao, Anant Chopra, Mark D Chaffin, Zeina R Al Sayed, Matthew C Hill, Helene Mantineo, Elizabeth M Brydon, Zexian Zeng, X Shirley Liu, and Patrick T Ellinor. Transfer learning enables predictions in network biology. *Nature*, 618(7965):616–624, June 2023.

[2] Haotian Cui, Chloe Wang, Hassaan Maan, and Bo Wang. scGPT: Towards building a foundation model for Single-Cell multi-omics using generative AI. May 2023.

[3] Fan Yang, Wenchuan Wang, Fang Wang, Yuan Fang, Duyu Tang, Junzhou Huang, Hui Lu, and Jianhua Yao. scBERT as a large-scale pretrained deep language model for cell type annotation of single-cell RNA-seq data. *Nature Machine Intelligence*, 4(10):852–866, September 2022.

[4] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the 31st International Conference on Neural*

Information Processing Systems, NIPS'17, pages 6000–6010, Red Hook, NY, USA, December 2017. Curran Associates Inc.

[5] OpenAI. GPT-4 technical report. March 2023.

[6] Qijie Chen, Haotong Sun, Haoyang Liu, Yinghui Jiang, Ting Ran, Xurui Jin, Xianglu Xiao, Zhimin Lin, Zhangming Niu, and Hongming Chen. A comprehensive benchmark study on biomedical text generation and mining with ChatGPT. April 2023.

[7] Som S Biswas. Role of chat GPT in public health. *Annals of biomedical engineering*, 51(5):868–869, May 2023.

[8] John W Ayers, Adam Poliak, Mark Dredze, Eric C Leas, Zechariah Zhu, Jessica B Kelley, Dennis J Faix, Aaron M Goodman, Christopher A Longhurst, Michael Hogarth, and Davey M Smith. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA internal medicine*, 183(6):589–596, June 2023.

[9] Eric Strong, Alicia DiGiammarino, Yingjie Weng, Andre Kumar, Poonam Hosamani, Jason Hom, and Jonathan H Chen. Chatbot vs medical student performance on Free-Response clinical reasoning examinations. *JAMA Internal Medicine*, 183(9):1028–1030, September 2023.

[10] OpenAI. New and improved embedding model. `https://openai.com/blog/new-and-improved-embedding-model`, March 2023. Accessed: 2023-10-4.

[11] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W Thomas, Florian Tramèr, Rose E Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. On the opportunities and risks of foundation models. August 2021.

[12] William Connell, Umair Khan, and Michael J Keiser. A single-cell gene expression language model. October 2022.

[13] Romain Lopez, Jeffrey Regier, Michael B Cole, Michael I Jordan, and Nir Yosef. Deep generative modeling for single-cell transcriptomics. *Nature Methods*, 15(12):1053–1058, December 2018.

[14] Mohammad Lotfollahi, F Alexander Wolf, and Fabian J Theis. scgen predicts single-cell perturbation responses. *Nature Methods*, 16(8):715–721, August 2019.

[15] Emily Clough and Tanya Barrett. The gene expression omnibus database. *Methods in molecular biology*, 1418: 93–110, 2016.

[16] Aviv Regev, Sarah A Teichmann, Eric S Lander, Ido Amit, Christophe Benoist, Ewan Birney, Bernd Bodenmiller, Peter Campbell, Piero Carninci, Menna Clatworthy, Hans Clevers, Bart Deplancke, Ian Dunham, James Eberwine, Roland Eils, Wolfgang Enard, Andrew Farmer, Lars Fugger, Berthold Göttgens, Nir Hacohen, Muzlifah Haniffa, Martin Hemberg, Seung Kim, Paul Klenerman, Arnold Kriegstein, Ed Lein, Sten Linnarsson, Emma Lundberg, Joakim Lundeberg, Partha Majumder, John C Marioni, Miriam Merad, Musa Mhlanga, Martijn Nawijn, Mihai Netea, Garry Nolan, Dana Pe'er, Anthony Phillipakis, Chris P Ponting, Stephen Quake, Wolf Reik, Orit Rozenblatt-Rosen, Joshua Sanes, Rahul Satija, Ton N Schumacher, Alex Shalek, Ehud Shapiro, Padmanee Sharma, Jay W Shin, Oliver Stegle, Michael Stratton, Michael J T Stubbington, Fabian J Theis, Matthias Uhlen, Alexander van Oudenaarden, Allon Wagner, Fiona Watt, Jonathan Weissman, Barbara Wold, Ramnik Xavier, Nir Yosef, and Human Cell Atlas Meeting Participants. The human cell atlas. *eLife*, 6, December 2017.

[17] Cellxgene data portal. `https://cellxgene.cziscience.com/docs/08__Cite%20cellxgene%20in%20your%20publications`. Accessed: 2023-10-4.

[18] Wenpin Hou and Zhicheng Ji. Reference-free and cost-effective automated cell type annotation with GPT-4 in single-cell RNA-seq analysis. *bioRxiv : the preprint server for biology*, April 2023.

[19] Oskar Wysocki, Zili Zhou, Paul O'Regan, Deborah Ferreira, Magdalena Wysocka, Dónal Landers, and André Freitas. Transformers and the representation of biomedical background knowledge. *Computational linguistics (Association for Computational Linguistics)*, 49(1):73–115, March 2023.

[20] Ruosong Ye, Caiqi Zhang, Runhui Wang, Shuyuan Xu, and Yongfeng Zhang. Natural language is all a graph needs. August 2023.

[21] Daniel Levine, Syed Asad Rizvi, Sacha Lévy, Nazreen Pallikkavaliyaveetil, Ruiming Wu, Zihe Zheng, Antonio Oliveira Fonseca, Xingyu Chen, Sina Ghadermarzi, Rahul M Dhodapkar, and David van Dijk. Cell2Sentence: Teaching large language models the language of biology. September 2023.

[22] John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. September 2019.

[23] Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. What you can cram into a single vector: Probing sentence embeddings for linguistic properties. *arXiv preprint arXiv:1805.01070*, 2018.

[24] Yonatan Belinkov, Nadir Durrani, Fahim Dalvi, Hassan Sajjad, and James Glass. What do neural machine translation models learn about morphology? *arXiv preprint arXiv:1704.03471*, 2017.

[25] Garth R Brown, Vichet Hem, Kenneth S Katz, Michael Ovetsky, Craig Wallin, Olga Ermolaeva, Igor Tolstoy, Tatiana Tatusova, Kim D Pruitt, Donna R Maglott, and Terence D Murphy. Gene: a gene-centered information resource at NCBI. *Nucleic acids research*, 43(Database issue):D36–42, January 2015.

[26] Welcome to MyGene.py's documentation! — MyGene.py v3.1.0 documentation. `https://docs.mygene.info/projects/mygene-py/en/latest/`. Accessed: 2023-10-4.

[27] Michihiro Yasunaga, Jure Leskovec, and Percy Liang. Linkbert: Pretraining language models with document links. *arXiv preprint arXiv:2203.15827*, 2022.

[28] Jingcheng Du, Peilin Jia, Yulin Dai, Cui Tao, Zhongming Zhao, and Degui Zhi. Gene2vec: distributed representation of genes based on co-expression. *BMC genomics*, 20(Suppl 1):82, February 2019.

[29] Malte D Luecken, M Büttner, K Chaichoompu, A Danese, M Interlandi, M F Mueller, D C Strobl, L Zappia, M Dugas, M Colomé-Tatché, and Fabian J Theis. Benchmarking atlas-level data integration in single-cell genomics. *Nature Methods*, 19(1):41–50, January 2022.

[30] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, 2008.

[31] Sijin Cheng, Ziyi Li, Ranran Gao, Baocai Xing, Yunong Gao, Yu Yang, Shishang Qin, Lei Zhang, Hanqiang Ouyang, Peng Du, Liang Jiang, Bin Zhang, Yue Yang, Xiliang Wang, Xianwen Ren, Jin-Xin Bei, Xueda Hu, Zhaode Bu, Jiafu Ji, and Zemin Zhang. A pan-cancer single-cell transcriptional atlas of tumor infiltrating myeloid cells. *Cell*, 184(3):792–809.e23, February 2021.

[32] Lucas Schirmer, Dmitry Velmeshev, Staffan Holmqvist, Max Kaufmann, Sebastian Werneburg, Diane Jung, Stephanie Vistnes, John H Stockley, Adam Young, Maike Steindel, Brian Tung, Nitasha Goyal, Aparna Bhaduri, Simone Mayer, Jan Broder Engler, Omer A Bayraktar, Robin J M Franklin, Maximilian Haeussler, Richard Reynolds, Dorothy P Schafer, Manuel A Friese, Lawrence R Shiow, Arnold R Kriegstein, and David H Rowitch. Neuronal vulnerability and multilineage diversity in Multiple Sclerosis. *Nature*, 573(7772):75–82, September 2019.

[33] Yanming Li, Pingping Ren, Ashley Dawson, Hernan G Vasquez, Waleed Ageedi, Chen Zhang, Wei Luo, Rui Chen, Yumei Li, Sangbae Kim, et al. Single-cell transcriptome analysis reveals dynamic cell populations and differential gene expression patterns in control and aneurysmal human aortic tissue. *Circulation*, 142(14):1374–1388, 2020.

[34] Mark Chaffin, Irinna Papangeli, Bridget Simonson, Amer-Denis Akkad, Matthew C Hill, Alessandro Arduini, Stephen J Fleming, Michelle Melanson, Sikander Hayat, Maria Kost-Alimova, et al. Single-nucleus profiling of human dilated and hypertrophic cardiomyopathy. *Nature*, 608(7921):174–180, 2022.

[35] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[36] Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C Lawrence Zitnick, Jerry Ma, et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, 118(15):e2016239118, 2021.

[37] Peter M Visscher, Matthew A Brown, Mark I McCarthy, and Jian Yang. Five years of gwas discovery. *The American Journal of Human Genetics*, 90(1):7–24, 2012.

A Example gene summary inputs to GenePT

In this section, we present additional details on the example gene summary inputs to GenePT. As an example, here is the paragraph we passed into GPT-3.5 embedding for gene CD24:

> CD24 Official Full Name CD24 molecule Primary source HGNC:HGNC:1645 See related Ensembl:ENSG00000272398 MIM:600074; AllianceGenome:HGNC:1645 **Gene type protein coding** RefSeq status REVIEWED Also known as CD24A **Summary This gene encodes a sialoglycoprotein that is expressed on mature granulocytes and B cells and modulates growth and differentiation signals to these cells. The precursor protein is cleaved to a short 32 amino acid mature peptide which is anchored via a glycosyl phosphatidylinositol (GPI) link to the cell surface. This gene was missing from previous genome assemblies, but is properly located on chromosome 6. Non-transcribed pseudogenes have been designated on chromosomes 1, 15, 20, and Y. Alternative splicing results in multiple transcript variants. Expression Biased expression in thyroid (RPKM 586.8), esophagus (RPKM 431.3) and 12 other tissues** See more Orthologs mouse all.

We see that generally there are generally two types of interest provided for a given gene: (i) various names, symbols, Refseq status, and orthologs; and (ii) summaries of gene types, functions, and notable expressions levels (highlighted in bold in the paragraph above). While our initial work did not quantitatively explore the sensitivity of the GenePT approach to various text cleaning and pre-processing methods, it is a critical direction for our future research.

B Additional results for the gene level functionality and property predictions

In Figure 2(b), we display the results from a prediction task that leverages GenePT embedding to predict the 15 most prevalent gene functional class, as detailed in Table 3 below. For baseline comparison, we also applied LR classification using Gene2vec embeddings on the subset of the genes that had a corresponding Gene2vec embeddings ($21,000$ genes). This yielded the five-fold cross-validated accuracy of 0.86 (SD: 0.03). On the same subset of $21,000$ genes, GenePT achieved a considerably higher average accuracy of 0.95 (SD:0.05).

Type	Count	Percentage (%)
Protein coding	20,184	71.6
pseudogene	3,725	13.2
miRNA	2,061	7.3
snRNA	1,793	6.4
misc RNA	1,043	3.7
lncRNA	724	2.6
snoRNA	692	2.5
antisense	592	2.1
rRNA	512	1.8
lincRNA	499	1.8
processed transcript	248	0.9
IG V gene	97	0.3
sense intronic	84	0.3
IG V pseudogene	79	0.3
TR J gene	76	0.3

Table 3: Gene functional classes used in the prediction task of Figure 2(b).

C Cell type annotation results

In this section, we also consider the cell type annotation task, where the primary aim is to predict annotated cell type labels based on the input cell representation. This annotation step is critical in single-cell analysis, as accurately distinguishing various cell populations within sequenced tissues can significantly enrich downstream biological insights. Mirroring the experimental design in the scGPT paper [2], we evaluated different embeddings' efficacy for cell-type annotation using with a 10-nearest-neighbor classifier on three datasets hPancreas, Myeloid, and Multiple Sclerosis with the same train/test split. We report the test set classification accuracy by applying a 10-nearest neighbor classifier on various pretrained embeddings in Table 4 and note that GenePT embeddings held the ground against pretrained scGPT embeddings and outperformed the pretrained Geneformer embeddings. Furthermore, we explored an ensemble approach that aggregates the 10 nearest neighbors from GenePT-w, GenePT-s, and scGPT, resulting in 30 predictions for each cell. This method demonstrated enhanced performance across various datasets and metrics. This indicates that literature-based natural language embeddings, such as GenePT-s, and expression-profile-derived embeddings like scGPT, provide complementary insights in single-cell biology tasks.

Table 4: Classification performance on cell-type annotation benchmarks used in scGPT, which includes the Myeloid, Multiple Sclerosis, and hPancreas datasets, as well the Aorta dataset (subsampled from data originally published in Li et al. [33]. Reported metrics include accuracy, precision, recall, and F1 (Macro-weighted), and are based on applying 10-nearest neighbor classifiers (using cosine similarity as the distance metric) on pretrained emebddings from scGPT, Geneformer, GenePT-w and GenePT-s. We also report the performance of ensembling the nearest neighbors retrieved by scGPT (or Geneformer in the case where we were not able to obtain scGPT embeddings for a dataset), GenePT-w, and GenePT-s.

		Classification metrics on the test set			
Dataset	Embeddings	Accuracy	Precision	Recall	F1
Myeloid	scGPT	0.53	0.33	0.29	0.30
	Geneformer	0.44	0.26	0.18	0.20
	GenePT-w	0.48	0.31	0.26	0.27
	GenePT-s	0.52	0.33	0.27	0.29
	Ensemble scGPT + GenePT-w + GenePT-s	**0.57**	**0.37**	**0.32**	**0.34**
Multiple Sclerosis	scGPT	**0.76**	**0.67**	**0.62**	**0.61**
	Geneformer	0.44	0.47	0.36	0.34
	GenePT-w	0.34	0.43	0.27	0.22
	GenePT-s	0.49	0.50	0.41	0.40
	Ensemble scGPT + GenePT-w + GenePT-s	0.71	0.66	0.57	0.55
hPancreas	scGPT	0.77	0.61	0.56	0.55
	Geneformer	0.50	0.25	0.34	0.27
	GenePT-w	**0.94**	0.72	0.63	0.65
	GenePT-s	0.89	0.65	0.53	0.56
	Ensemble scGPT + GenePT-w + GenePT-s	**0.94**	**0.80**	**0.67**	**0.70**
Aorta	Geneformer	0.86	0.70	0.60	0.62
	GenePT-w	0.87	**0.91**	**0.68**	**0.72**
	GenePT-s	0.86	0.70	0.60	0.62
	Ensemble Geneformer + GenePT-w + GenePT-s	**0.88**	0.84	0.64	0.68

D Additional visualization on the batch effect and underlying diseases biology via GenePT in the cardiomyocytes and Aorta data

In Figure 3, we visualize the original single-cell data and GenePT embeddings, colored by disease type (top row) and patient id (bottom row). We see that while the original data was capturing the underlying disease biology (top left; NF: non-failing heart; HCM: hearts with hypertrophic cardiomyopathy; DCM: hearts with dilated cardiomyopathy), it also was highly affected by patient batch effect (panel (b) in Figure 3; different color indicates individual patients). On the other hand, cell embeddings generated by GenePT-s clustered primarily by disease phenotype rather than patients.

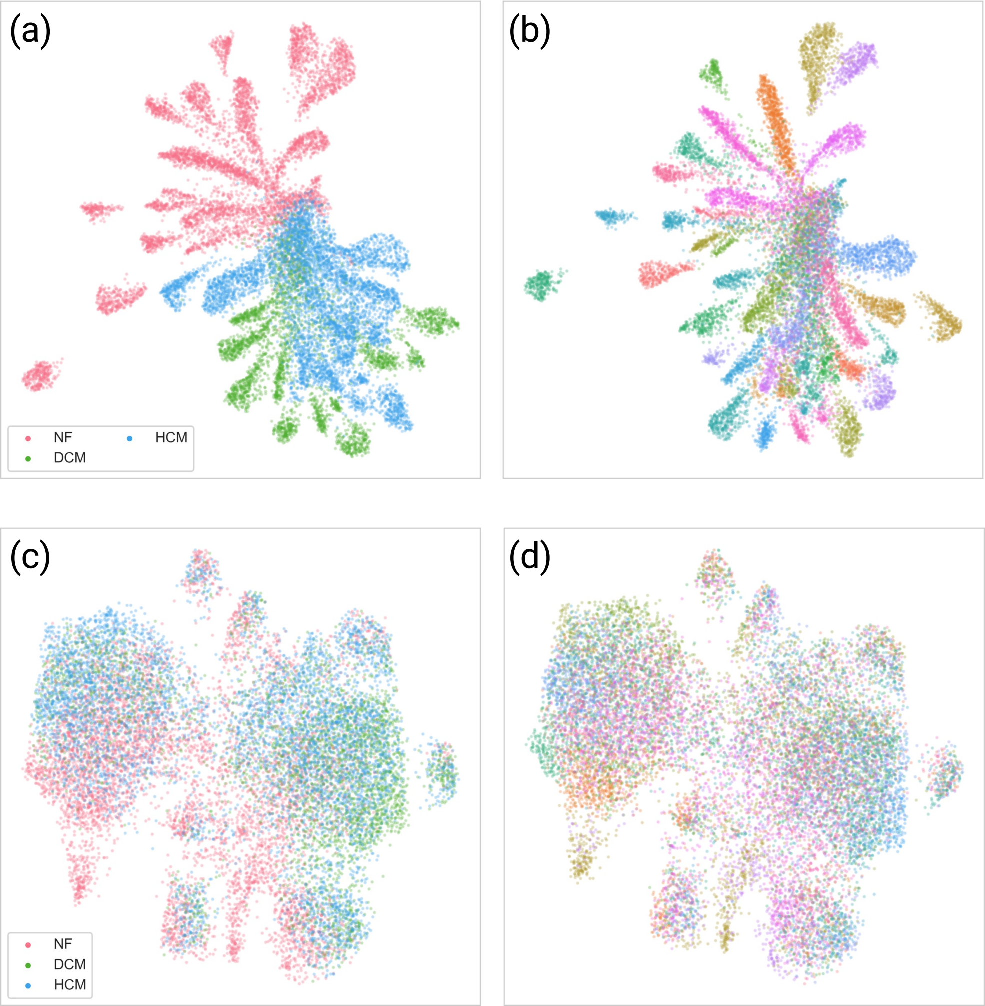


Figure 3: **(a)** UMAP visualization of single-cell data, colored by disease phenotype where NF, HCM, and DCM stand for non-failing heart, hearts with hypertrophic cardiomyopathy, and hearts with dilated cardiomyopathy, respectively. **(b)** Same as **(a)**, but colored by patient id. **(c)** UMAP visualization of GenePT-s embeddings of the same set of cells as **(a)**, colored by disease phenotype. **(d)** Same as **(c)**, but colored by patient id.

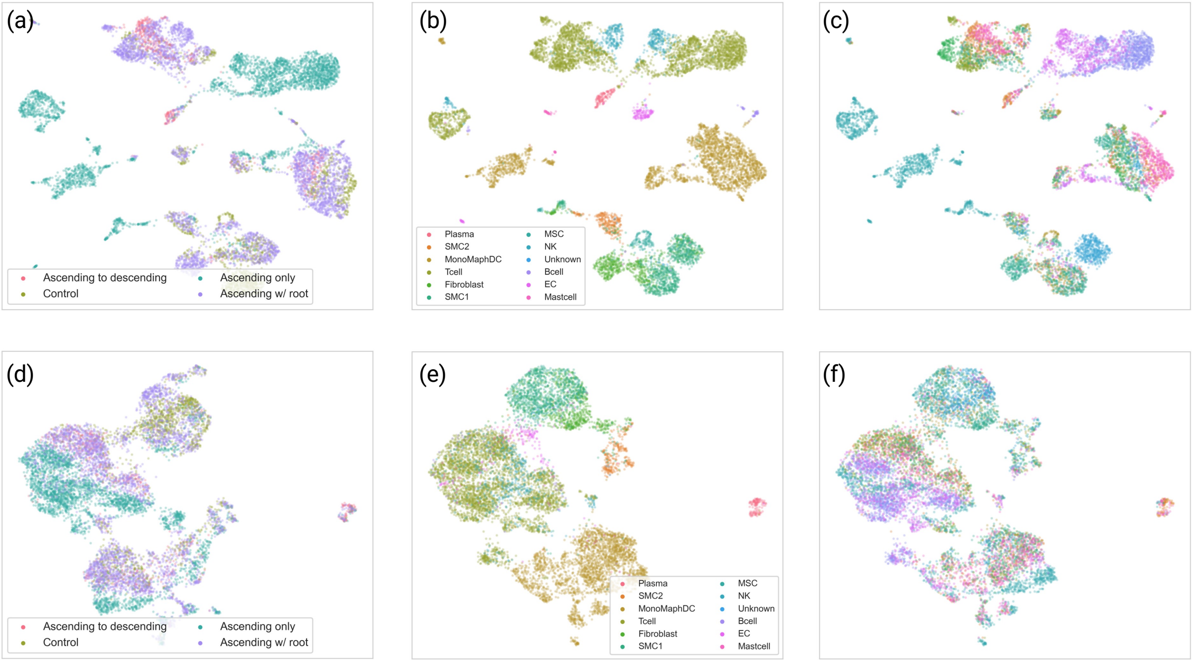


Figure 4: **(a)** UMAP visualization of the subsampled Aorta dataset, colored by disease phenotype (three different disease phenotypes: ascending only, ascending with descending thoracic aortic aneurysm, and ascending with root aneurysm; one control phenotype comprising patients with healthy hearts after transplant) provided in the original study [33]. **(b)** Same as **(a)**, but colored by cell types annotated by the original study [33]. **(c)** Same as **(a)**, but colored by patient id. **(d)** UMAP visualization of GenePT-s embeddings of the same set of cells as **(a)**, colored by disease phenotype. **(e)** Same as **(d)**, but colored by cell types. **(f)** Same as **(d)**, but colored by patient id.