TEDDY: A FAMILY OF FOUNDATION MODELS FOR UNDERSTANDING SINGLE CELL BIOLOGY

Alexis Chevalier [1] **Soumya Ghosh** [2] **Urvi Awasthi** [1] **James Watkins** [2] **Julia Bieniewska** [1] **Nichita Mitrea** [3] **Olga Kotova** [2] **Kirill Shkura** [3] **Andrew Noble** [2] **Michael J. Steinbaugh** [2] **Vijay Sadashivaiah** [2] **George Dasoulas** [2] **Julien Delile** [1] **Christoph Meier** [1] **Leonid Zhukov** [1] **Iya Khalil** [2] **Srayanta Mukherjee** [1] **Judith Mueller** [2]

Abstract

Understanding the biological mechanisms of disease is crucial for medicine, and in particular, for drug discovery. AI-powered analysis of genome-scale biological data holds great potential in this regard. The increasing availability of single-cell RNA sequencing data has enabled the development of large foundation models for disease biology. However, existing foundation models only modestly improve over task-specific models in downstream applications. Here, we explored two avenues for improving single-cell foundation models. First, we scaled the pre-training data to a diverse collection of 116 million cells, which is larger than those used by previous models. Second, we leveraged the availability of large-scale biological annotations as a form of supervision during pre-training. We trained the TEDDY family of models comprising six transformer-based state-of-the-art single-cell foundation models with 70 million, 160 million, and 400 million parameters. We vetted our models on several downstream evaluation tasks, including identifying the underlying disease state of held-out donors not seen during training, distinguishing between diseased and healthy cells for disease conditions and donors not seen during training, and probing the learned representations for known biology. Our models showed substantial improvement over existing works, and scaling experiments showed that performance improved predictably with both data volume and parameter count.

[*]Equal contribution [1]BCG AI Science Institute, Boston, USA [2]Merck & Co., Inc., Cambridge, MA, USA [3]MSD (UK) Limited, London, UK. Correspondence to: Soumya Ghosh <soumya.ghosh@merck.com>, Srayanta Mukherjee <mukherjee.srayanta@bcg.com>, Judith Mueller <judith.mueller@merck.com>.

Proceedings of the Workshop on Generative AI for Biology at the 42nd International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025.

1. Introduction

The complexity of cell biology and the mechanisms of disease pathogenesis are driven by an intricate regulatory network of genes (Chatterjee and Ahituv, 2017; Theodoris et al., 2015; 2021). A better resolution of this complex interactome network would enhance our ability to design drugs that target the causal mechanism of the disease rather than interventions that aim to modulate the downstream effects (Ding et al., 2022). However, accurate inference of gene regulatory networks is challenging. The possible space for genetic interactions is vast (Bunne et al., 2024), the networks to be inferred are highly context-dependent, different cell types and tissue types exhibit different regulatory networks and exhibit significant variations across donors (Chen and Dahl, 2024). Moreover, the data required to study gene regulatory networks for a specific disease is usually limited and highly specialized, often plagued by experimental artifacts (Hicks et al., 2018).

However, a confluence of recent technological progress promises to make this challenging problem more tractable. The advent of accurate single-cell sequencing technologies that remove the artifacts of bulk cell data, better reflect natural variability, and provide signals at higher resolutions. This, along with the increasing availability of atlas-scale scRNA-seq datasets that span an extensive range of diseases, cell types, tissue types, and donors provide an unprecedented opportunity for studying disease mechanisms at scale. Parallel and complementary progress (Vaswani et al., 2017) in artificial intelligence has produced tools, so-called foundation models, capable of absorbing and effectively learning from massive amounts of data (Devlin, 2018; Radford et al., 2019; Brown et al., 2020; Meta, 2024; Achiam et al., 2023). These approaches use self-supervised learning to learn from large volumes of unlabeled data, for instance, data scraped from all of the internet, and then adapt to specific tasks from modest amounts of task-specific, typically labeled, data (Devlin, 2018; Brown et al., 2020) and have revolutionized the ability of algorithms to understand natural language and images. Whether the same learning paradigm can be used to enhance our understanding of the

language of genes, specifically disease biology, has been a subject of recent research (Theodoris et al., 2023; Cui et al., 2024; Schaar et al., 2024; Chen et al., 2024). This interest is fueled by the fact that seemingly many of the basic principles that drive success in natural language carry over to the gene space: gene embeddings, i.e., gene representations in the vector space, can capture rich context-dependent relations (Wagner et al., 2016) that can be learned at scale from single-cell measurements in an unsupervised fashion.

Despite recent progress, open questions remain about the usefulness and the potential of foundation models for scRNA-seq. Theodoris et al. showed that downstream performance correlates with pre-training dataset size, but others (Liu et al., 2023; Boiarsky et al., 2023) found that untrained transformer models and traditional machine learning methods are competitive and can sometimes surpass foundation models pre-trained on large single cell corpora. Yet others (Kedzierska et al., 2023) found the zero-shot embeddings extracted from various foundation models to not substantially improve over less resource-intensive approaches. Moreover, methods for extracting gene regulatory information from foundation models remain under explored, with existing attempts (Cui et al., 2024) relying on simple clustering methods over gene embedding tables.

In this work, we took a systematic approach to the development of single-cell foundation models, seeking to advance the state-of-the-art in understanding disease biology. We trained a new family of foundation models adapted to disease biology, the TEDDY family of models[1]. These models incorporate architectural choices inspired by existing works SCGPT (Cui et al., 2024), GENEFORMER (Theodoris et al., 2023), and NICHEFORMER (Schaar et al., 2024), train on to-date the largest corpus of single-cell data, and use biological ontologies to supervise gene and cell representation learning. TEDDY is trained on data sourced from CELLXGENE (CZI Single-Cell Biology Program et al., 2023). The pre-training dataset contains 116 million (116M) cells from mouse, human, spatial, and dissociated scRNA-seq data. The TEDDY family ranges from 10M to 400M parameters, allowing us to examine the scaling behavior with respect to the number of parameters and the amount of pre-training data. To evaluate performance on the downstream task of disease understanding, we benchmark TEDDY against existing models on two disease classification tasks. We find that the larger TEDDY models improve on existing models on one task substantially and are within noise of the best-performing competitor in the other.

We believe TEDDY is an important step towards designing foundation models that understand disease biology. Nonetheless, significant innovations such as principled approaches for incorporating existing biological knowledge, continuing to scale the amount of pre-training data, and incorporating other complementary modalities will all likely be required to design foundation models that fully represent the complexity of single-cell biology.

2. Related work

The success of transformer-based foundation models in modeling natural language (Devlin, 2018; Achiam et al., 2023) and vision (Radford et al., 2021) has inspired a growing body of research in modeling single-cell transcriptomic data similarly. These models, analogously to natural language, treat cells as sentences and genes expressed in cells as words. They primarily differ in how they represent gene expression data, the training data quality and volume, and the self-supervised objectives they use for learning. We highlight these design choices below.

Owing to the lack of natural ordering among genes expressed in a cell, BERT (Devlin, 2018) style models that employ bi-directional self-attention and are trained via masked language modeling (MLM) are popular in the literature. scBERT (Yang et al., 2022), an early example of this category, was trained on a corpus of one million (1M) cells from PanglaoDB (Franzén et al., 2019) and primarily evaluated on the downstream tasks of cell-type annotation and discovery.

Subsequent models, including the GENEFORMERv1 (Theodoris et al., 2023) and v2 (Chen et al., 2024), were trained on larger corpora of 30M and 95M cells. These works represent a cell as a ranked list of genes it expresses. This rank-value encoding scheme ranks genes expressed in a cell based on their (normalized) expression values from most to least expressed. The pre-training task then involves predicting the gene expressed at a particular position in the ranked list, having observed the other genes in the ranked list. The authors demonstrate that pre-training in this fashion endows the models with diverse downstream capabilities, including network and chromatin dynamic predictions, in-silico gene-network analysis, and batch integration. Another recent model, NICHEFORMER (Schaar et al., 2024), also makes use of rank-value encoding but trains on a corpus of 110M cells containing both dissociated (53M) and spatially resolved (57M) cells drawn from both humans and mice to decode spatially resolved cellular information.

Others, for instance, SCGPT(Cui et al., 2024), a 53M[2] parameter model trained on 33M cells, bin gene expression counts independently for each cell and train the model to predict the bin a gene's expression in a cell discretizes to given the other binned gene expressions and optional metadata about the cell. Yet others, scFoundation (Hao et al.,

[1] TEDDY: *Transformer for Enabling Drug DiscoverY*

[2] CZI blog

2024), AIDO.Cell (Ho et al., 2024) represent gene expressions as weighted combinations of embeddings, where the mapping from the expression level of a gene to weights in the weighted combination are learned along with the embeddings during pre-training. They are trained on a corpus of 50M cells on a pre-training task of predicting observed gene expression levels from corrupted versions. Decoder-only architectures (Bian et al., 2024) and approaches (Rosen et al., 2023) combining transcriptomic foundation models with foundation models for other modalities are also starting to emerge.

Beyond model development, there is also fledgling literature focused on evaluating the effectiveness of foundation models (Liu et al., 2023; Kedzierska et al., 2023; Boiarsky et al., 2023) on various downstream tasks and reporting mixed results with foundation models showing promise but not consistently outperforming task-specific traditional methods that do not leverage large cell atlases.

Motivated by the shortcomings of current approaches, we explore two avenues for improving over the state-of-the-art. First, we scale the pre-training dataset to 116M cells which is an order of magnitude larger than data scBert, scMulan, and scGPT (Yang et al., 2022; Bian et al., 2024) were pre-trained on and a few million cells larger than those used by Nicheformer (Schaar et al., 2024) and Geneformerv2 (Chen et al., 2024). Second, we note that in contrast with domains like text and natural images, where training datasets are largely scraped from the internet, experimental data in scRNA-seq atlases such as CELLxGENE come with rich meta-data annotations that can aid in learning better representations. We explicitly leverage these annotations while training our TEDDY family of models by augmenting the self-supervised pre-training task of masked language modeling with an additional supervised task of predicting available meta-data annotations from gene expressions.

3. The TEDDY family of foundation models

The TEDDY family of models contains two variants that differ in how they represent gene expressions and what pre-training tasks they use. TEDDY-G follows the GENEFORMER recipe. It uses rank-value encoding and trains on the self-supervised task of predicting the gene expressed at a particular position in the ranked list representing a cell. TEDDY-X follows the SCGPT approach of binning gene expressions and training the model to predict the binned expression level of a gene. Both TEDDY-G and TEDDY-X additionally use the supervised annotation loss detailed in Section 3.2. We train a series of performant models for both variants containing approximately 70M, 160M, and 400M parameters. We also train smaller, less effective 10M and 30M parameter models to carefully probe the scaling behavior of these models and find that the two approaches present different challenges for downstream applications.

3.1. Pre-training dataset

We derive our training corpus from CELLxGENE (CZI Single-Cell Biology Program et al., 2023), a collection of 1,399 single-cell RNA-seq datasets from open-source publications. At the time of download[3], CELLxGENE contained 160M cells, of which 70M were from primary datasets, from 24,000 donors, 122 different diseases, 413 tissue types, 860 cell types, and 23M diseased cells.

We filtered low quality cells containing fewer than 225 gene counts and more than 10% mitochondrial transcript abundance. Dying or highly stressed cells often exhibit high levels of mitochondrial gene expression and were removed following recommendations from previous work (Satija et al., 2023). We also excluded studies using 10x Genomics Chromium v1 chemistry in favor of newer datasets captured with v2 and v3 chemistry. After quality control, our corpus contained 116M cells. We held out a subset of the data for validation and removed all data sets used for downstream testing in Section 4 to avoid data contamination.

3.2. Supervision via biological annotations

Table 1. **Biological annotations.** Each ontology term in CELLxGENE is mapped to one of the forty-three categories below, and the corresponding special token is added to the vocabulary.

Special tokens	*Annotation labels*
`<disease>`	brain, cancer, cardiovascular, genetic, immune, infectious, kidney, respiratory, other, healthy
`<tissue_type>`	adipose, cardiovascular, central nervous, digestive, embryonic, endocrine, exocrine, eye, hematopoietic, hepatic, integumentary, musculature, renal, reproductive, respiratory, sensory, unknown
`<cell_type>`	ciliated, connective, contractile, embryonic, epithelial, hematopoietic, immune, neural, perivascular, precursor, secretory, skeletal, unknown
`<sex>`	male, female, unknown

Annotations such as disease, cell type, tissue type, sex, and other labels capturing cell-specific metadata are available in CELLxGENE data. Existing foundation models either do not use these annotations during pre-training or only use these labels as metadata tokens to provide additional context for gene modeling, i.e., predicting gene expression levels in a cell. In early experiments, we found that adding meta-data as context for gene modeling brought no measurable improvements in gene modeling abilities (data not shown). Instead, we leverage the available annotations as supervisory signals during pre-training. Our pre-training task involves biological annotation modeling, i.e., predicting biological annotations associated with a cell, in addition to gene modeling. This encourages models to learn embeddings that, in addition to being predictive of gene expression levels, align with high-level biological properties encoded in the annotations.

[3]June 11, 2024.

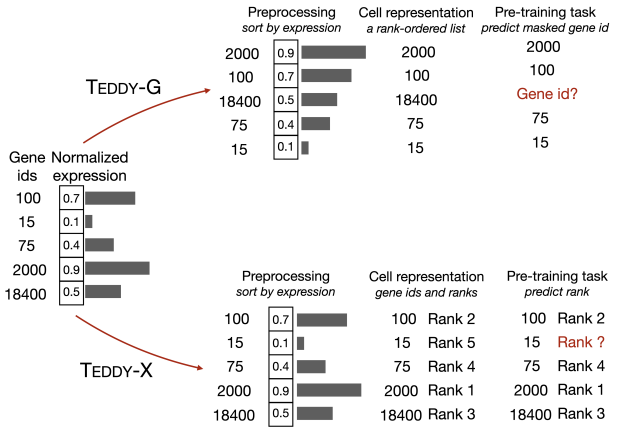

Figure 1. **An illustration of differences between TEDDY variants**. On the left we illustrate a cell with five non-zero expressed genes with non-zero median normalized expression values. TEDDY-G represents a cell as a list of genes ordered by their expression levels and the pre-training task involves predicting the index of masked out genes. TEDDY-X ranks expression values, scales them to the interval $[-1, +1]$, and then learns to predict a masked rank scaled to the interval $[-1, +1]$.

Focusing on disease, cell type, tissue type, and sex annotations, we use the CELLXGENE ontology tree to organize the 1399 labels that occur in these four categories into forty-three broad labels, which correspond to metadata labels for the data sets in the CELLXGENE corpus, shown in Table 1. We emphasize that the ontology terms we train with are coarse-grained annotations. They do not include fine-grained disease labels used in downstream model evaluation.

To incorporate the biological annotations, we add four special tokens `<disease>`, `<tissue_type>`, `<cell_type>`, `<sex>` to the model's vocabulary, one for each annotation category of interest. We prepend these special tokens to the sequence of genes expressed by a cell and train the model to predict the annotation labels from gene expression data.

3.3. Pre-training objective

We introduce notation to state the pre-training objective used by TEDDY variants formally. Let V denote the vocabulary size of the model, $\boldsymbol{x}_v \in \mathbb{R}^d$ denote a d-dimensional embedding, and $v \in [1, \ldots, V]$ denote the index of gene v in the vocabulary. Denote by $\boldsymbol{x}_{\texttt{<disease>}}, \boldsymbol{x}_{\texttt{<tissue>}}, \boldsymbol{x}_{\texttt{<cell>}}, \boldsymbol{x}_{\texttt{<sex>}}$ the embeddings associated with the special tokens. Let $\boldsymbol{t}_n = \{t_{nj}\}_{j=1}^{J_n}$ with $t_{nj} \in [1, \ldots, V]$ and $J_n < V$, represent the indices of the genes expressed by the cell $\boldsymbol{c}_n$, the self supervised objective employed by TEDDY-G is,

$$\ell_{\text{MLM}-\text{G}}(\theta) = \mathbb{E}_{\mathbf{c}_n \sim \mathcal{D}}\big[\mathbb{E}_{m \sim \mathcal{M}}[-\log \text{Cat}(t_{nm} | g_\theta(\mathbf{c}_{n\backslash m}))]\big],$$

where $\mathcal{D} = \{\mathbf{c}_n\}_{n=1}^{\text{N}}$ is a dataset containing N cells, θ are TEDDY-G parameters, Cat represents a V-way categorical distribution, $\mathcal{M}$ is the set of masked genes, and $\mathbf{c}_{n\backslash m} = \{\boldsymbol{x}_{nj}\}_{j \neq m} \bigcup \{\boldsymbol{x}_{\texttt{<disease>}}, \boldsymbol{x}_{\texttt{<tissue>}}, \boldsymbol{x}_{\texttt{<cell>}}, \boldsymbol{x}_{\texttt{<sex>}}\}$ represents the unmasked portion of the cell and $g_\theta(\mathbf{c}_{n\backslash m})$ is the softmax-transformed mean of the categorical distribution predicted by TEDDY-G. Throughout this work, we create $\mathcal{M}$ by masking out 15% of the genes uniformly at random. Let r_{nj} represent the rank of gene j in cell $\boldsymbol{c}_n$ scaled to the interval $[-1, 1]$. That is, the most expressed gene in cell $\boldsymbol{c}_n$ is mapped to 1 and the gene with the smallest non-zero expression is mapped to -1. The self-supervised objective employed by TEDDY-X is,

$$\ell_{\text{MLM}-\text{X}}(\theta) = \mathbb{E}_{\mathbf{c}_n \sim \mathcal{D}}\big[\mathbb{E}_{m \sim \mathcal{M}}[-\log \mathcal{N}(r_{nm} | f_\theta(\mathbf{c}_n), 1)]\big],$$

where $f_\theta(\mathbf{c}_n)$ is predicted by TEDDY-X and $\mathbf{c}_n = \mathbf{c}_{n\backslash m} \bigcup \{\boldsymbol{x}_{nm}\}$.

Let $\boldsymbol{y}_n^{\texttt{<disease>}}, \boldsymbol{y}_n^{\texttt{<cell>}}, \boldsymbol{y}_n^{\texttt{<tissue>}}, \boldsymbol{y}_n^{\texttt{<sex>}}$ denote one-hot vectors representing the disease, cell-type, tissue-type, and sex annotations associated with cell $\boldsymbol{c}_n$. The supervised component of the pre-trained objective is,

$$\begin{aligned}
&\ell_{\text{CLS}-\text{G}}(\theta) \\
&= \mathbb{E}_{\mathbf{c}_n \sim \mathcal{D}}\big[\mathbb{E}_{m \sim \mathcal{M}}[-\log \text{Cat}(\boldsymbol{y}_n^{\texttt{<disease>}} | g_\theta^{\text{dcls}}(\mathbf{c}_{n\backslash m})) \\
&- \log \text{Cat}(\boldsymbol{y}_n^{\texttt{<cell>}} | g_\theta^{\text{ccls}}(\mathbf{c}_{n\backslash m})) \\
&- \log \text{Cat}(\boldsymbol{y}_n^{\texttt{<tissue>}} | g_\theta^{\text{tcls}} \mathbf{c}_{n\backslash m})) \\
&- \log \text{Cat}(\boldsymbol{y}_n^{\texttt{<sex>}} | g_\theta^{\text{scls}}(\mathbf{c}_{n\backslash m}))]\big],
\end{aligned}$$

where $g_\theta^{\text{dcls}}(\mathbf{c}_{n\backslash m})), g_\theta^{\text{ccls}}(\mathbf{c}_{n\backslash m})), g_\theta^{\text{tcls}}(\mathbf{c}_{n\backslash m})), g_\theta^{\text{scls}}(\mathbf{c}_{n\backslash m}))$ represent TEDDY-G's softmax-transformed predictions. $\ell_{\text{CLS}-\text{X}}$ is analogously defined. Finally, the overall pre-training objective for TEDDY-G is,

$$\ell_{\text{pre}-\text{G}}(\theta) = \ell_{\text{MLM}-\text{G}}(\theta) + \ell_{\text{CLS}-\text{G}}(\theta). \tag{1}$$

The pre-training objective for TEDDY-X involves replacing $\ell_{\text{MLM}-\text{G}}$ with $\ell_{\text{MLM}-\text{X}}$ and $\ell_{\text{CLS}-\text{G}}$ with $\ell_{\text{CLS}-\text{X}}$ in Equation (1). In preliminary experiments, we found weighting the two components of the pre-training objective differently did not substantially affect pre-training or downstream performance. A more careful exploration is part of planned future work. Other recent work (Wang et al., 2025) cautions against using granular cell-type labels for supervision, demonstrating that models trained with such supervision distort underlying biological relationships among cells. They instead advocate for weaker supervision. The coarse cell-type labels we employ, along with the additional regularization provided by the masked language modeling loss, provide precisely such weak supervision.

3.4. Training details and scaling laws

We used a context length of 2048 genes for TEDDY models. We based this choice on results from pilot experiments on

a small subset of the data (see Appendix B.3. We trained all model variants for a single epoch, on 116M cells. We performed non-zero median normalization (Theodoris et al., 2021) of the expression data and use a vocabulary of $48,308$, which includes human and mouse protein-coding and micro-RNA genes as well as special tokens and tokens corresponding to biological annotation labels. We included micro-RNA genes (Theodoris et al., 2023) and mouse data (Schaar et al., 2024) based on similar choices made in previous work. We used a batch size of 256, a linear warmup with a linear decay learning rate scheduler with a warmup of 10000 steps, a maximum learning rate of 1e-4 decayed to zero at the end of training, and the AdamW optimizer. We also found that a relatively large weight decay of 0.1 was necessary to avoid loss spikes and divergences during training. See Appendix B for additional details and Figure 1 for a summary.

Figure 2 illustrates performance on held-out data for TEDDY-G models as a function of model size and training data volume. We observe that performance improves with both increasing model size and data volume but improves only modestly between the 160M and 400M parameter models and not in the straight power law regime.

4. A disease benchmark for transcriptomic foundation models

To assess the performance of the TEDDY foundation models, we carried out a benchmarking exercise. The extensive pre-training datasets used to train these models increase the risk of unintentionally contaminating the benchmark evaluation data with information that the model may have encountered during pre-training. Cognizant of these issues, we designed two downstream tasks that hold out data at different levels of granularity to assess the disease modeling capabilities of transcriptomic foundation models.

4.1. Held-out diseases task

In the first task, we evaluate the models' ability to generalize to unseen disease conditions. This involves assessing the models using data from donors with diseases that were not included in the pre-training phase.

We sourced five datasets from the CELLxGENE database, each containing cells from healthy donors and donors with a single disease condition. We sub-sampled these datasets to ensure that the proportion of diseased and healthy cells is equal. The disease conditions included are `Chronic Kidney Disease (CKD)`, `Alzheimer's Disease`, `Gastric Cancer`, `Rheumatoid Arthritis`, and `Pediatric Crohn's Disease`. To prevent data contamination, we excluded these five datasets, as well as any other datasets from CELLxGENE that contained these disease labels, from our pre-training corpus. Additional details about the data can be found in Appendix C.

The benchmark task involves solving five binary classification problems–one for each disease condition–to identify whether a cell is healthy or diseased. We performed three-fold cross validation for these tasks and ensure that there is no donor overlap across folds. The final reported accuracy is averaged over the each of the three held-out test sets in the folds.

4.2. Held-out donors task

We designed the second task to investigate whether foundation models generalize across the biological variability exhibited by different cell donors. To this end, we held out data from 82 donors, with thirteen different disease (or Normal) conditions — `COVID-19`, `Alzheimer's Disease`, `Acute Myeloid Leukemia`, `Blastoma`, `Luminal A Breast Carcinoma`, `Gingivitis`, `Luminal B Breast Carcinoma`, `Multiple Sclerosis`, `Myocardial Infarction`, `Normal`, `Periodontitis`, `Pilocytic Astrocytoma`, `Premalignant Hematological System Disease`, and `type-2 Diabetes Mellitus` from the pre-training corpus. We then evaluated the effectiveness of different methods at labeling whether a cell comes from a donor with one of these thirteen disease conditions or a healthy donor given the genes expressed by the cell. Unlike the coarse-grained disease annotations during pre-training (Table 1), these disease annotations are finer-grained, making this a non-trivial classification problem. We additionally collected 511 donors from the pre-training corpus and thirteen donors with Alzhiemer's disease which had been left out of the pre-training dataset for a total of 524 donors for training and validation. We finetuned on 70% of the 524 donors and held-out out 30% of the donors for validation. Finally, we tested on the 82 donors that were held out from the pretraining corpus. By construction our train, validation, and test splits had no overlapping donors. See Appendix C for additional details about the dataset.

5. Vetting the TEDDY family of foundation models

With the benchmark in hand, we performed careful experiments to vet the various modeling choices made by TEDDY models. We then proceed to compare TEDDY against existing state-of-the-art foundation models – NICHEFORMER (Schaar et al., 2024), SCGPT (Cui et al., 2024), six-layer GENEFORMER, and twelve-layer GENEFORMER 12L variants (Theodoris et al., 2023), and task-specific machine learning approaches.

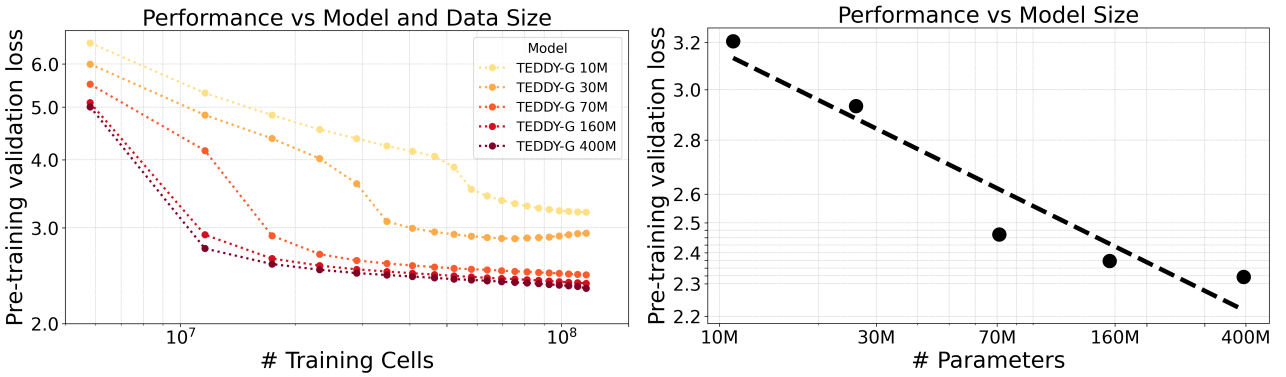

Figure 2. **Scaling behavior.** *Left:* Pre-training loss on held-out data for TEDDY-G as a function of pre-training data and model size. *Right:* Validation loss at the end of one epoch of training against the number of parameters in the model. The dashed line in black is the linear best-fit: $14.95 \times (\#\text{ parameters})^{-0.10}$.

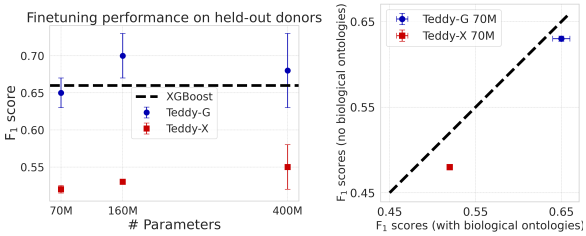

Figure 3. **Performance on held-out donors as a function of model size and biological annotaions** *Left:*F_1 scores improve for both TEDDY-G and TEDDY-X with increasing size. TEDDY-G outperforms TEDDY-X across model scale and improves on XGBoost trained exclusively for this task. *Right:* The x-axis plots the F_1 scores achieved by TEDDY-G 70M and TEDDY-X 70M when pre-trained with supervision from biological ontologies. The y-axis plots F_1 scores achieved by the same models without ontologies.

The held-out donors task exhibits class imbalance; we thus report both accuracy and weighted F_1 metrics. The results are aggregated over three random initializations of the foundation models. In the held-out diseases tasks, we evaluate on class-balanced datasets and hence only report accuracies. The results for this task are aggregated over the three cross-validation folds, each using a different random initialization.

5.1. Downstream performance improves with model scale and biological annotations

We begin with experiments comparing the impact of model size on our evaluations benchmark. Figure 3 plots finetuning performance of TEDDY-G and TEDDY-X across different model sizes on the held-out donors task. We observe that both model variants improve in performance with the number of parameters, but TEDDY-G performs substantially better across all parameter sizes. Moreover, while TEDDY-G 160M and 400M outperform a task-specific XGBoost classifier trained on log- transformed, filtered gene expression count data TEDDY-X does not.

We also evaluated the impact of adding supervision through biological ontologies as described in Section 3.2. To this end, we compared the performance of TEDDY-G 70M and TEDDY-X 70M pre-trained with and without explicit supervision. On the downstream held-out donors task, Figure 3 summarizes our results. We find that both TEDDY variants benefit from supervision; models pre-trained with biological ontologies result in better downstream performance. Moreover, we found that adding supervision led to more stable pre-training with fewer loss divergences and had no adverse impact on gene-modeling abilities of the model, as illustrated in A4.

Having observed TEDDY-G to generally outperform TEDDY-X on the held-out donors task, we only experimented with TEDDY-G on the remaining tasks.

5.2. TEDDY models outperform existing foundation models on held-out donors task

Table 2 summarizes the performance of different models on the held-out donors task. We observe that in comparison to competing foundation models, TEDDY-G generalizes better across donor variability. Upon finetuning, TEDDY-G 400M achieved an accuracy of 0.72, outperforming NICHEFORMER, the best-performing non-TEDDY model, by 8.0% (0.72 vs. 0.64) and GENEFORMER, the worst-performing model, by 45.8% (0.73 vs. 0.39). Similarly, TEDDY-G improves the F_1 score over NICHEFORMER by 17.6% (0.68 vs. 0.56) and a 67.6% improvement over GENEFORMER (0.68 vs. 0.22). These results highlight that TEDDY-G effectively generalizes across donors, an important use-case in practice.

Table 2. Performance of different foundation models on held-out donors task

	TEDDY-G 400M	NICHEFORMER	SCGPT	GENEFORMER	GENEFORMER12L
Held-out donors - Fine-tuning					
Accuracy	0.72±0.04	0.64±0.01	0.64±0.01	0.39±0.00	0.55±0.002
Weighted F1	0.68±0.06	0.56±0.01	0.54±0.01	0.22±0.02	0.45±0.02

Table 3. Logistic regression and TEDDY-G 400M on held-out diseases task.

	Logistic Regression	Linear probed TEDDY-G 400M	Finetuned TEDDY-G 400M
Chronic Kidney Disease	0.90±0.03	0.95±0.02	0.94±0.03
Alzheimer's Disease	0.64±0.08	0.86±0.07	0.75±0.18
Gastric Cancer	0.49±0.04	0.70±0.09	0.56±0.09
Rheumatoid Arthritis	0.47±0.03	0.93±0.10	0.56±0.06
Pediatric Crohn's Disease	0.63±0.20	0.72±0.16	0.74±0.18

5.3. TEDDY offers modest improvements over existing foundation models on held-out diseases task

The performance of TEDDY-G 400M and other foundational models on the held-out diseases dataset is summarized in Table A3 and Figure 4. We found that across the five binary classification sub-tasks TEDDY-G improves on all five when compared to GENEFORMER12L, on all but one task compared to the six layered GENEFORMERand SCGPT. TEDDY-G's performance was within noise of NICHEFORMER, a model trained on similar volume of data as TEDDY and an order of magnitude larger than SCGPTand GENEFORMER. This suggests that data volume is important for performing well on diseases not encountered during training. Scaling to larger volumes of data may lead to improved performance.

5.4. TEDDY improves over task-specific machine learning approaches

We also benchmarked TEDDY against a diverse set of task-specific machine learning methods—logistic regression, linear SVM, XGBoost, and lightGBM. We trained each of these methods to predict disease labels given log-transformed, filtered count data. We performed feature selection using the `highly_variable_genes` function from the SCANPY V1.10.1 library with default parameters. This method identifies genes exhibiting high variability across cells by calculating the normalized dispersion and selecting the top genes based on a predefined threshold. We performed the `highly_variable_genes` analysis for each cross-validation training fold and filtered the training and testing folds to only retain the identified genes.

In Table 3 we highlight comparisons of fine-tuned TEDDY-G 400M and linearly (with logistic regression) probed TEDDY-G 400M against logistic regression. We note that the TEDDY variants typically outperform logistic regression with linear-probing performing the best and often by a substantial margin. Additional results are detailed in Table A2. Comparing these with Figure 4 and Table A3 we found that TEDDY-G 400M and NICHEFORMER consistently improve over task-specific approaches.

We conducted an additional experiment to evaluate whether the representations learned by TEDDY-G are useful for disease classification. To do this, we replaced the hand-engineered features used by the task-specific methods with pre-trained embeddings from TEDDY-G 400M. We constructed cell-level embeddings by averaging over output gene embeddings produced by TEDDY-G 400M. We note that these embeddings are zero-shot, they were not fine-tuned on the held-out disease datasets. We then trained the task specific methods as before. The results are summarized in Figure A2, with numbers available in Table A2. Observe that across all methods and all diseases replacing hand-engineered features with zero-shot embeddings leads to substantially better performance. These results suggest that embeddings better capture disease-relevant information than hand-engineered features based on log-transformed transcript counts.

5.5. Zero-shot TEDDY embeddings produce coherent clusters

Next, to probe the zero-shot embeddings produced by TEDDY-G we looked at the embeddings produced by TEDDY-G 400M on data from (Thomas et al., 2024) that contains human gut single cell transcriptomes from donors with Ulcerative Colitis (UC), Crohn's Disease (CD), and those not afflicted by either disease. Importantly for our purposes, (Thomas et al., 2024) provided detailed cell-type annotations at varying levels of granularity, allowing us to benchmark our embeddings quantitatively. We used a version of the data reprocessed using Cell-Ranger v7.1.0. After filtering out cells with greater than 10% mitochondrial proportion and those with no author-provided labels, we ended up with 536,207 cells with cell-type annotations. We benchmarked TEDDY-G 400M against a recently proposed single-cell foundation model – Cell2Sentence (Rizvi et al., 2025). For this comparison, we selected the Cell2Sentence-410M model, which is of similar scale (410 million parameters) and fine-tuned for cell type prediction tasks.

In Figure A3, we plot two-dimensional embeddings from both models and present quantitative clustering metrics computed using the scib (Luecken et al., 2022) package in Table A6. We found TEDDY-G 400M to produce largely

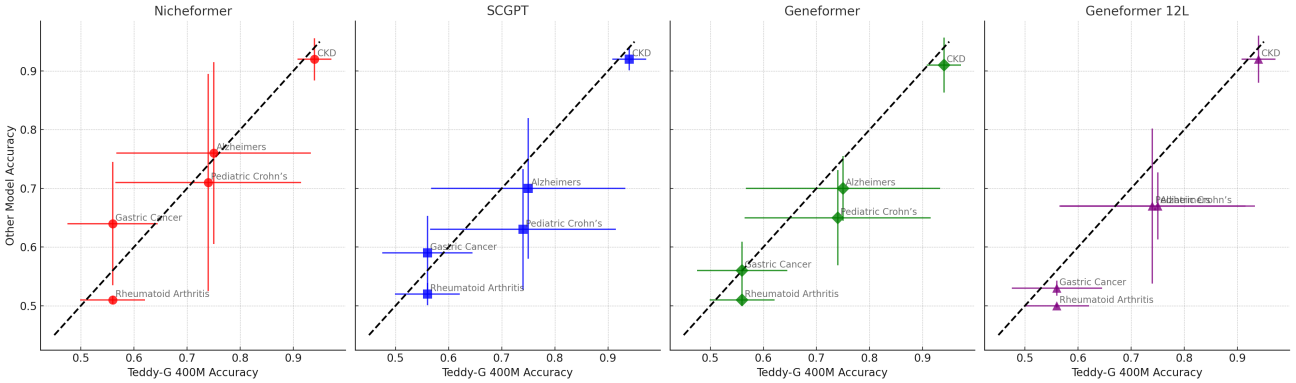

Figure 4. **Performance of existing foundation models compared with TEDDY-G** TEDDY-G consistently outperforms other foundation models. The x-axis plots fine-tuning accuracy achieved TEDDY-G 400M. The y-axis is the accuracy achieved by other foundation models. Points below the diagonal indicate TEDDY-G 400M achieves higher accuracy.

coherent clusters and outperform Cell2Sentence 410M at both the coarse (low) and intermediate cell-type annotation resolutions. In particular, we found TEDDY-G 400M embeddings capable of distinguishing subtle cell-state differences, for instance, between CD_4 and CD_8 cell-states that Cell2Sentence fails to separate. Additional details about this experiment are in Appendix D.1.

5.6. In-silico single-gene perturbations with TEDDY align with experimental evidence

To probe whether TEDDY-G encodes *causal* gene–network structure, beyond merely capturing co-expression statistics, we used the single-gene knockout benchmark introduced as an evaluation for GENEFORMER (Theodoris et al., 2023). We evaluated the models with an *in silico* single-gene perturbation that mimics a loss-of-function experiment for the cardiac transcription factor GATA4. Crucially, we did not update any model parameters for this experiment, so these results reflect *zero-shot* mechanistic knowledge acquired by the model during pre-training and underscores the potential of large pretrained models for *in-silico* target discovery.

First, we isolated two-hundred-thirty-six fetal-cardiomyocyte transcriptomes from the heart cell atlas (Knight-Schrijver et al., 2022) in which GATA4 is expressed, so that the knockout is biologically relevant. We then created a perturbed representation of these cells by setting the GATA4 expression to zero. We then embedded the before and after perturbation cells using TEDDY-G (and other competing models). For every gene g expressed in these cells we extracted its token-level representation before ($\mathbf{v}_g^{\text{before}}$) and after ($\mathbf{v}_g^{\text{after}}$) the perturbation and measured, $s_g = \cos\left(\mathbf{v}_g^{\text{before}}, \mathbf{v}_g^{\text{after}}\right)$, the cosine similarity between the two embeddings. This yielded one similarity score per gene per cell. We further partitioned genes into biologically annotated groups: *direct targets* of GATA4, *indirect targets*, and *housekeeping genes*, using labels provided in (Theodoris et al., 2023). For each cell, we averaged the similarity scores of the genes within each group to obtain average per-cell effects, $\bar{s}^{(i)}_{\text{direct}}$, $\bar{s}^{(i)}_{\text{indirect}}$, $\bar{s}^{(i)}_{\text{direct+indirect}}$ and $\bar{s}^{(i)}_{\text{housekeeping}}$, with i indexing the cell. We then subjected the sets of similarities, $\{\bar{s}^{(i)}_{\text{direct}}\}_{i=1}^{236}$, $\{\bar{s}^{(i)}_{\text{indirect}}\}_{i=1}^{236}$, $\{\bar{s}^{(i)}_{\text{direct+indirect}}\}_{i=1}^{236}$, $\{\bar{s}^{(i)}_{\text{housekeeping}}\}_{i=1}^{236}$ to paired one-sided Wilcoxon signed-rank tests whose alternative hypothesis matched the biological expectation (e.g. $\bar{s}_{\text{housekeeping}} > \bar{s}_{\text{direct}}$ for the *HK-vs-Direct* comparison, and $\bar{s}_{\text{housekeeping}} < \bar{s}_{\text{indirect}}$ for *Indirect-vs-Direct*). The four comparisons were (i) *housekeeping* vs. *direct*, (ii) *housekeeping* vs. *indirect*, (iii) *housekeeping* vs. *direct+indirect*, and (iv) *indirect* vs. *direct* and the resulting p-values are in Table 4. Across all four paired comparisons, the

Model	$p_{\text{HK}-\text{Dir}}$	$p_{\text{HK}-\text{Indir}}$	$p_{\text{HK}-\text{Dir+Indir}}$	$p_{\text{Indir}-\text{Dir}}$
TEDDY-G 400M	9.45×10^{-25}	6.47×10^{-22}	1.66×10^{-23}	1.27×10^{-20}
Geneformer 12-layer	1.43×10^{-5}	7.72×10^{-6}	9.26×10^{-7}	0.96
Geneformer (base)	5.01×10^{-6}	4.59×10^{-6}	2.22×10^{-6}	0.98
Nicheformer	0.72	2.62×10^{-6}	9.21×10^{-4}	1.00
scGPT	0.75	0.20	0.32	0.99

Table 4. Paired Wilcoxon signed-rank p-values for the GATA4 knockout assay ($n = 236$ cells; four tests per model).

TEDDY-G 400M model yields small p-values, showing that its embedding space separates gene sets exactly as GATA4 biology predicts: housekeeping genes change the least, direct targets change the most, indirect targets lie in between, and the indirect–vs.–direct contrast is strongly negative. Both GENEFORMER variants reproduce a similar qualitative ordering, but fail to distinguish between direct and indirect effects. NICHEFORMERcaptures only downstream effects: it distinguishes indirect targets from housekeeping genes yet fails to separate direct targets, suggesting it encodes transcriptomic consequences rather than primary transcription factor–target relationships. SCGPT shows no significant housekeeping–target

separation at all, indicating that its latent space is largely insensitive to this perturbation.

Taken together, these results underscore two points: (i) increasing model scale and using biologically informed objectives markedly enhance zero-shot mechanistic knowledge, and (ii) TEDDY-G can go beyond co-expression in the case of GATA4, and arrange genes in a manner that mirrors their causal proximity to the perturbed transcription factor.

6. Acknowledgements

The authors would like to thank Marinka Zitnik (Harvard Medical School, Harvard Data Science Initiative, Kemper institute, Broad Institute) for her thoughtful comments on the paper and to Michael Brochu (BCG) for overall support and direction. We also would like to thank Greg Hersch and many other colleagues that enabled the infrastructure for training TEDDY: (Merck Central IT) Ron Kim, Asheesh Chhabra, Manoj Vig, Amany Basily, Venki Balakrishnan, Mike Dickmann, Mohan Kumar, Rasti Matus, Kate Lipina, Amal Verghese, Karthik Palanis; (Merck Research Laboratories IT) Carol Rohl, Venkat Parakala, Antong Chen, Danny Bitton, Faisal Hoda, Atul Singal, Martin Radocha, Jan Lubojacky. And thank you to our partners at Databricks: Pedro Portilla, Srijit Nair, Tim Lortz, and Abraam Samuel.

Impact Statement

This paper presents work whose goal is to advance the use of machine learning in molecular biology. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Haiyang Bian, Yixin Chen, Xiaomin Dong, Chen Li, Minsheng Hao, Sijie Chen, Jinyi Hu, Maosong Sun, Lei Wei, and Xuegong Zhang. scmulan: a multitask generative pre-trained language model for single-cell analysis. In *International Conference on Research in Computational Molecular Biology*, pages 479–482. Springer, 2024.

Rebecca Boiarsky, Nalini Singh, Alejandro Buendia, Gad Getz, and David Sontag. A deep dive into single-cell rna sequencing foundation models. *bioRxiv*, pages 2023–10, 2023.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

Charlotte Bunne, Yusuf Roohani, Yanay Rosen, Ankit Gupta, Xikun Zhang, Marcel Roed, Theo Alexandrov, Mohammed AlQuraishi, Patricia Brennan, Daniel B Burkhardt, et al. How to build the virtual cell with artificial intelligence: Priorities and opportunities. *arXiv preprint arXiv:2409.11654*, 2024.

Sumantra Chatterjee and Nadav Ahituv. Gene regulatory elements, major drivers of human disease. *Annual Review of Genomics and Human Genetics*, 18(Volume 18, 2017): 45–63, 2017. ISSN 1545-293X.

Han Chen, Madhavan S. Venkatesh, Javier Gómez Ortega, Siddharth V. Mahesh, Tarak N. Nandi, Ravi K. Madduri, Karin Pelka, and Christina V. Theodoris. Quantized multitask learning for context-specific representations of gene network dynamics. *bioRxiv*, 2024.

Minhui Chen and Andy Dahl. A robust model for cell type-specific interindividual variation in single-cell rna sequencing data. *Nature Communications*, 15(1):5229, 2024.

Haotian Cui, Chloe Wang, Hassaan Maan, Kuan Pang, Fengning Luo, Nan Duan, and Bo Wang. scgpt: toward building a foundation model for single-cell multi-omics using generative ai. *Nature Methods*, pages 1–11, 2024.

CZI Single-Cell Biology Program, Shibla Abdulla, Brian Aevermann, Pedro Assis, Seve Badajoz, Sidney M. Bell, Emanuele Bezzi, Batuhan Cakir, Jim Chaffer, Signe Chambers, J. Michael Cherry, Tiffany Chi, Jennifer Chien, Leah Dorman, Pablo Garcia-Nieto, Nayib Gloria, Mim Hastie, Daniel Hegeman, Jason Hilton, Timmy Huang, Amanda Infeld, Ana-Maria Istrate, Ivana Jelic, Kuni Katsuya, Yang Joon Kim, Karen Liang, Mike Lin, Maximilian Lombardo, Bailey Marshall, Bruce Martin, Fran McDade, Colin Megill, Nikhil Patel, Alexander Predeus, Brian Raymor, Behnam Robatmili, Dave Rogers, Erica Rutherford, Dana Sadgat, Andrew Shin, Corinn Small, Trent Smith, Prathap Sridharan, Alexander Tarashansky, Norbert Tavares, Harley Thomas, Andrew Tolopko, Meghan Urisko, Joyce Yan, Garabet Yeretssian, Jennifer Zamanian, Arathi Mani, Jonah Cool, and Ambrose Carr. Cz cell×gene discover: A single-cell data platform for scalable exploration, analysis and modeling of aggregated data. *bioRxiv*, 2023. doi: 10.1101/2023.10.30.563174.

Jacob Devlin. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Jun Ding, Nadav Sharon, and Ziv Bar-Joseph. Temporal modelling using single-cell transcriptomics. *Nature Reviews Genetics*, 23(6):355–368, 2022.

Oscar Franzén, Li-Ming Gan, and Johan LM Björkegren. Panglaodb: a web server for exploration of mouse and human single-cell rna sequencing data. *Database*, 2019: baz046, 2019.

Minsheng Hao, Jing Gong, Xin Zeng, Chiming Liu, Yucheng Guo, Xingyi Cheng, Taifeng Wang, Jianzhu Ma, Xuegong Zhang, and Le Song. Large-scale foundation model on single-cell transcriptomics. *Nature Methods*, pages 1–11, 2024.

Stephanie C Hicks, F William Townes, Mingxiang Teng, and Rafael A Irizarry. Missing data and technical variability in single-cell rna-sequencing experiments. *Biostatistics*, 19(4):562–578, 2018.

Nicholas Ho, Caleb N. Ellington, Jinyu Hou, Sohan Addagudi, Shentong Mo, Tianhua Tao, Dian Li, Yonghao Zhuang, Hongyi Wang, Xingyi Cheng, Le Song, and Eric P. Xing. Scaling dense representations for single cell with transcriptome-scale context. *bioRxiv*, 2024.

Kasia Zofia Kedzierska, Lorin Crawford, Ava Pardis Amini, and Alex X Lu. Assessing the limits of zero-shot foundation models in single-cell biology. *bioRxiv*, pages 2023–10, 2023.

Vincent R Knight-Schrijver, Hongorzul Davaapil, Semih Bayraktar, Alexander DB Ross, Kazumasa Kanemaru, James Cranley, Monika Dabrowska, Minal Patel, Krzysztof Polanski, Xiaoling He, et al. A single-cell comparison of adult and fetal human epicardium defines the age-associated changes in epicardial activity. *Nature cardiovascular research*, 1(12):1215–1229, 2022.

Tianyu Liu, Kexing Li, Yuge Wang, Hongyu Li, and Hongyu Zhao. Evaluating the utilities of foundation models in single-cell data analysis. *bioRxiv*, pages 2023–09, 2023.

Malte D Luecken, Maren Büttner, Kridsadakorn Chaichoompu, Anna Danese, Marta Interlandi, Michaela F Müller, Daniel C Strobl, Luke Zappia, Martin Dugas, Maria Colomé-Tatché, et al. Benchmarking atlas-level data integration in single-cell genomics. *Nature methods*, 19(1):41–50, 2022.

Llama Team AI @ Meta. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

Syed Asad Rizvi, Daniel Levine, Aakash Patel, Shiyang Zhang, Eric Wang, Sizhuang He, David Zhang, Cerise Tang, Zhuoyang Lyu, Rayyan Darji, et al. Scaling large language models for next-generation single-cell analysis. *bioRxiv*, pages 2025–04, 2025.

Yanay Rosen, Yusuf Roohani, Ayush Agrawal, Leon Samotorcan, Tabula Sapiens Consortium, Stephen R Quake, and Jure Leskovec. Universal cell embeddings: A foundation model for cell biology. *bioRxiv*, pages 2023–11, 2023.

Rahul Satija, Satija Lab, and Collaborators. Seurat - guided clustering tutorial, October 2023. URL `https://satijalab.org/seurat/articles/pbmc3k_tutorial.html`.

Anna Christina Schaar, Alejandro Tejada-Lapuerta, Giovanni Palla, Robert Gutgesell, Lennard Halle, Mariia Minaeva, Larsen Vornholz, Leander Dony, Francesca Drummer, Mojtaba Bahrami, et al. Nicheformer: a foundation model for single-cell and spatial omics. *bioRxiv*, pages 2024–04, 2024.

Christina V Theodoris, Molong Li, Mark P White, Lei Liu, Daniel He, Katherine S Pollard, Benoit G Bruneau, and Deepak Srivastava. Human disease modeling reveals integrated transcriptional and epigenetic mechanisms of notch1 haploinsufficiency. *Cell*, 160(6):1072–1086, 2015.

Christina V Theodoris, Ping Zhou, Lei Liu, Yu Zhang, Tomohiro Nishino, Yu Huang, Aleksandra Kostina, Sanjeev S Ranade, Casey A Gifford, Vladimir Uspenskiy, et al. Network-based screen in ipsc-derived cells reveals therapeutic candidate for heart valve disease. *Science*, 371(6530), 2021.

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, et al. Transfer learning enables predictions in network biology. *Nature*, 618(7965):616–624, 2023.

Tom Thomas, Matthias Friedrich, Charlotte Rich-Griffin, Mathilde Pohin, Devika Agarwal, Julia Pakpoor, Carl Lee, Ruchi Tandon, Aniko Rendek, Dominik Aschenbrenner, et al. A longitudinal single-cell atlas of anti-tumour necrosis factor treatment in inflammatory bowel disease. *Nature Immunology*, pages 1–14, 2024.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Allon Wagner, Aviv Regev, and Nir Yosef. Revealing the vectors of cellular identity with single-cell genomics. *Nature biotechnology*, 34(11):1145–1160, 2016.

Hanchen Wang, Jure Leskovec, and Aviv Regev. Limitations of cell embedding metrics assessed using drifting islands. *Nature Biotechnology*, pages 1–4, 2025.

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, 2022.

Nick Youngblut, Chris Carpenter, Jaanak Prashar, Chiara Ricci-Tam, Rajesh Ilango, Noam Teyssier, Silvana Konermann, Patrick D. Hsu, Alex Dobin, Dave P. Burke, Hani Goodarzi, and Yusuf Roohani. scbasecamp: an ai agent-curated, uniformly processed, and continually expanding single cell data repository. *Arc Institute Technical Report*, 2025.

Jesse Zhang, Airol A Ubas, Richard de Borja, Valentine Svensson, Nicole Thomas, Neha Thakar, Ian Lai, Aidan Winters, Umair Khan, Matthew G. Jones, Vuong Tran, Joseph Pangallo, Efthymia Papalexi, Ajay Sapre, Hoai Nguyen, Oliver Sanderson, Maria Nigos, Olivia Kaplan, Sarah Schroeder, Bryan Hariadi, Simone Marrujo, Crina Curca Alec Salvino, Guillermo Gallareta Olivares, Ryan Koehler, Gary Geiss, Alexander Rosenberg, Charles Roco, Daniele Merico, Nima Alidoust, Hani Goodarzi, and Johnny Yu. Tahoe-100m: A giga-scale single-cell perturbation atlas for context-dependent gene function and cellular modeling. *bioRxiv*, 2025. doi: 10.1101/2025.02.20.639398.

Appendix

A. Discussion

We have introduced the TEDDY family foundation models and developed a benchmark for probing the disease biology learned by these models. The TEDDY family, trained on a comprehensive 116M dataset taken from the CELLxGENE corpus, consists of two variants, TEDDY-G trained with masked gene modeling and TEDDY-X trained with masked gene expression level modeling. Both come in 70M, 160M, and 400M parameter sizes, and our empirical validation found that larger models and data led to improved performance, though with diminishing returns. Deviating from previous work, the TEDDY models also used large-scale biological annotations as supervisory signals while pre-training. Our experiments showed that the inductive biases introduced by such supervision substantially improve downstream performance on the held-out donors task and do not hurt performance on the held-out diseases task.

The limited effectiveness of the current crop of transcriptomic foundation models, including TEDDY, on the held-out diseases task, could be attributed to either the failure of the models to adequately generalize to data not seen during pre-training or to the fact that the task itself is plagued by noise from experimental and data collection artifacts. After all, we do not have healthy or diseased annotations for individual cells; instead, we rely on whether the cells were derived from healthy or diseased donors. Carefully disentangling these factors is important for future work.

Other future directions include improvements in the data aspects, such as increasing the size of the pre-trained corpus, quantifying the diversity of the corpus, augmenting it with additional disease-specific data, and more thoroughly filtering it to remove data of poor quality, as well as complimentary modeling improvements – priors that carefully incorporate known biology improve model performance. Augmenting the training data with large scale interventional data (Zhang et al., 2025) and uniformly processed observational data (Youngblut et al., 2025) that reduce batch effects are of particular interest. Another direction includes evaluations on more complex biological tasks, like perturbation prediction and inference of gene regulatory networks (GRNs). Foundation models that can accurately predict expression changes after gene perturbations *in-silico* would be powerful tools for reconstructing disease mechanisms by identifying key regulatory pathways. By modeling how specific gene knockouts or activations impact downstream gene expression, these models could help biologists pinpoint the drivers of dysregulated pathways in diseases and prioritize targets for therapeutic intervention.

TEDDY models could help facilitate the reconstruction of the GRN in two different ways. First, we plan to introduce a new fine-tuning task in which the model predicts the probability that a regulatory edge exists between two genes, thus enabling the identification of novel interactions that are otherwise more difficult to extract from experimental data. Second, leveraging Perturb-seq fine-tuned models to use predicted expression profiles after gene knockouts as input for GRN inference algorithms offers a data-driven approach to constructing regulatory networks. This dual strategy could empower biologists to map regulatory interactions efficiently and on a large scale, thereby accelerating discoveries in complex regulatory networks and their roles in disease biology. Both these tasks directly apply to drug target discovery and a better understanding of disease biology. Finally, capturing multi-omics information via multi-modal integration would be critical, considering that cellular regulation is multi-layered. Hence, we plan to extend TEDDY beyond its current scRNA-seq modality. Notwithstanding, TEDDY establishes a new state-of-the-art for foundation models for scRNA-seq data.

B. Modeling Choices and Details

B.1. Architecture details and model sizes

The TEDDY models vary in the number of parameters from approximately 10 million parameters to approximately 400 million parameters. The exact parameter counts are detailed in Table A1. The 10M TEDDY models contain three transformer blocks (layers) and use a token embedding dimension of 128. The 30M models use six transformer blocks and a embedding dimension of 256. The 70M models use 12 transformer blocks and an embedding dimension of 512. The 160M models use 12 transformer blocks and an embedding dimension of 768 and finally the 400M models use 24 transformer blocks and 1024 dimensional embeddings.

B.2. Annotation labels

During pre-training, we balance the classes used for ontology classification. To achieve this, we compute the number of cells with each label over the entire train set, and during training, we randomly prompt the model with classification tokens with sampling probabilities designed to balance the probability of each label.

Table A1. Training data size and model size for TEDDY models and existing foundation models.

	Pretrain corpus size by number of cells	Model size by number of parameters
TEDDY-G 400M	116M	394.2M
TEDDY-X 400M	116M	414.2M
TEDDY-G 160M	116M	154.0M
TEDDY-X 160M	116M	164.8M
TEDDY-G 70M	116M	71.2M
TEDDY-X 70M	116M	72.0M
TEDDY-G 30M	116M	26.2M
TEDDY-X 30M	116M	26.1M
TEDDY-G 10M	116M	11.9M
TEDDY-X 10M	116M	11.8M
Nicheformer	110M	46.8M
SCGPT	33M	51.3M
Geneformer	30M	10.3M
Geneformer 12L	30M	39.6M
Cell2Sentence	57M	410M

In particular, we balance disease classification so that the cells with `<disease>` tokens have a 50% probability of being normal or diseased.

B.3. Other design explorations for the TEDDY family

To guide our model development efforts, we also explore the effect of different approaches to data pre-processing, the effect of sequence length, and the effect of training on multi-species data.

We implemented the median scaling method introduced in (Theodoris et al., 2023). We compute the median non-zero expression value of each gene after scaling each cell to 10,000 gene expression counts. We then scale each gene expression by this median value, so that each gene has a median expression value of 1 over the entire training dataset. We trained models with and without median-scaling on a subsampled version of our pre-training dataset comprising of four million (4M) cells. The model trained on unscaled data achieved a 60% improvement in training loss, clearly indicating the importance of median scaling. All our models are trained with median scaling.

We also compared the effect of pre-training with different sequence lengths.

We trained two TEDDY-G models (without biological annotations) on our 4M-cell dataset with sequence lengths of 2,048 and 1,500. After truncating the *test set* for both models down to 1,500-long sequences, the 2,048-long model achieved a test loss of 2.84, whereas the 1,500-long model achieved a loss of 2.88. The loss achieved by the 2,048-long model on the full 2,048-long sequences was 2.63. This shows that single-cell foundation models benefit from pre-training on longer sequences and understand shorter sequences better. Therefore, we train all our models with 2,048 tokens. We leave it as future work to study the importance of sequence length, especially in relation with recent work on long-context large language models.

Our 116M pre-training dataset contains both human and mouse genes. We tokenize ENSMUS mouse genes separately from human genes. Preliminary experiments on the 4M-cell data comparing models trained on both human and mouse data show that they achieve similar test loss to a model trained exclusively on human data. A more large scale exploration of using multi-species data is needed to better understand the benefit of jointly training on multi-species data and is interesting future work, especially with the advent of large multi-species datasets (Youngblut et al., 2025).

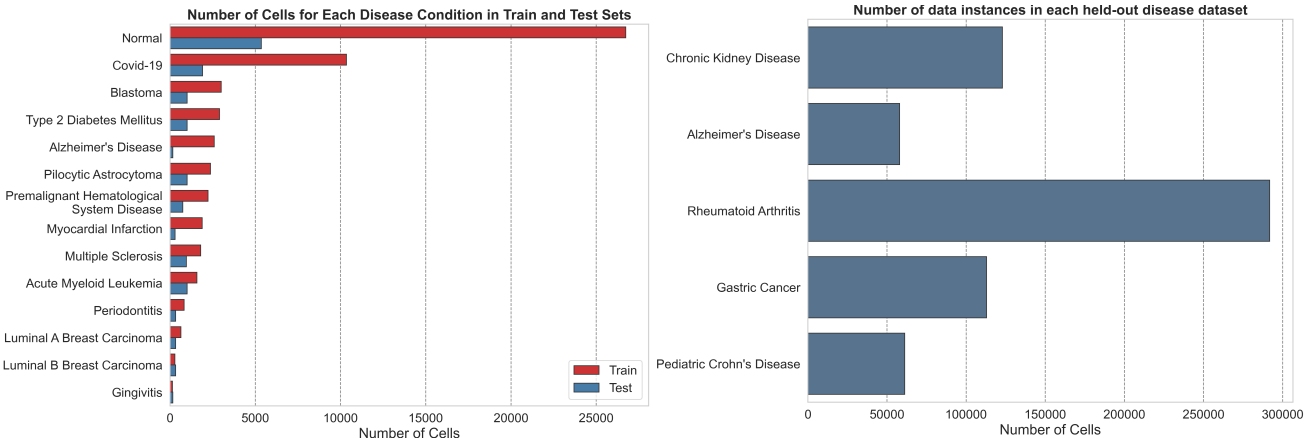

Figure A1. **Evaluation benchmark statistics.** *Left:* Distributions of disease conditions in train and test sets of the held-out donors dataset. *Right:* Number of instances (cells) in the different held-out disease datasets.

C. Evaluation benchmark details

C.1. Held-out donors dataset

The training set comprises of cells from 524 donors. The test set, contains data from 82 donors who were excluded from the pre-training data. The task is a 14-way classification problem, Figure A1 illustrates the distribution of cells across the fourteen labels (thirteen disease conditions and healthy). As can be seen the proportion of labels vary widely across disease conditions. To account for this, we report class-weighted F_1 scores in addition to accuracy.

Evaluation details We used a single train/test split, with 70% of the training set used for training and 30% for validation. We used AdamW to fine-tune the models. We ran a parameter sweep over six learning rates (1E-3, 1E-4, 2E-4, 2E-5, 5E-4, 5E-5) and using pytorch default values for other optimizer parameters. The best learning rate was selected based on accuracy on the validation set. The final model was trained on the full training set and tested on the held-out set and repeated with three random seeds per model.

C.2. Held-out diseases dataset

We use five datasets sourced from CellXGene for binary disease classification of cells as healthy or diseased in the held-out diseases task. These 5 datasets, and any other datasets containing disease labels present in these, were explicitly excluded from the pre-training corpus. The CellXGene datasets were preprocessed before inclusion, with minimal modifications applied to maintain their original state for reproducibility, except for filtering of non-primary datasets. The datasets chosen for evaluation were processed and required to meet several criteria: (1) inclusion of data from at least five donors to allow for robust cross-validation splits by donor; (2) an equal balance of healthy and diseased cells to address the class imbalance issue of the original data; and (3) a minimum of 10,000 cells per dataset, to provide sufficient training signal. We evaluated the performance of the model for the binary disease classification task using k-fold accuracy. Since these datasets are balanced between healthy and diseased cell types we just report accuracy as a measure of performance.

The dataset includes transcriptomic data from five different disease conditions, each with a balanced representation of diseased and normal cells. The Chronic Kidney Disease (CKD) dataset comprises 123,982 kidney cells from 25 donors, spanning 26 distinct kidney-related cell types, including epithelial cells of the proximal tubule, loop of Henle, and collecting ducts. The Alzheimer's Disease dataset consists of 57,010 cells from the prefrontal cortex, sourced from 11 donors, all classified as inhibitory interneurons. The Gastric Cancer dataset contains 122,922 stomach-derived cells from 9 donors, but all are labeled under an unknown classification. The Pediatric Crohn's Disease dataset features 61,244 cells from 16 donors, representing 31 diverse immune and epithelial cell types, such as enterocytes, B cells, and intestinal crypt stem cells, collected from the ileum. Lastly, the Rheumatoid Arthritis (RA) dataset is the largest, with 291,140 immune cells derived from 25 donors, predominantly consisting of T cells (CD4+ and CD8+), monocytes, dendritic cells, and natural killer cells, all obtained from blood samples. Together, these datasets offer a comprehensive resource for studying disease-related transcriptomic variations across multiple organ systems.

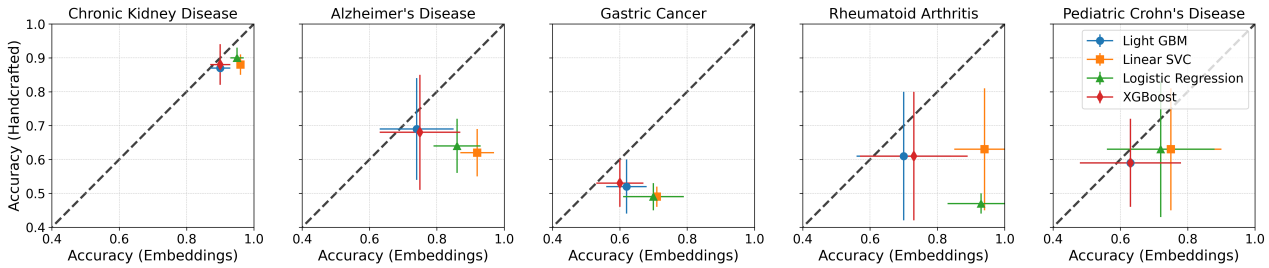

Figure A2. **Performance of task-specific machine learning methods using handcrafted features and TEDDY-G 400M embeddings** Accuracy increases with use of TEDDY-G embeddings. The x-axis plots accuracy achieved by different methods using TEDDY-G 400M embeddings as features. The y-axis is the accuracy achieved when using hand engineered features. Points below the diagonal indicate that embedding features achieve higher accuracy.

Evaluation details We performed 3-fold cross-validation, creating three train/test folds such that there is no overlap in donors across the train and test folds and the proportion of diseased and normal cells in each fold is 50%. Each training fold was further divided into train and validation subsplits for model selection. We used AdamW to fine-tune the mdoels. We ran a parameter sweep over six learning rates (1E-3, 1E-4, 2E-4, 2E-5, 5E-4, 5E-5) on each subsplit and selected the best learning rate per model and disease based on accuracy on the validation sub-split. Finally we retrained the model on the full training fold with the selected hyper-parameter and tested on the unseen test fold. Since the test folds are perfectly disease-balanced so we reported only classification accuracy.

D. Experimental details

In Table A2, Table A3, Table A4, Table A4, Table A5 we provide detailed numbers from our expriments described in Section 5 and Appendix C.

Table A2. Accuracy of task-specific methods on held-out diseases task.

	LightGBM	Linear SVC	Logistic Regression	XGBoost Classifier
Held-out diseases - Handcrafted features				
CKD	0.87±0.05	0.88±0.03	0.90±0.03	0.88±0.06
Alzheimers	0.69±0.15	0.62±0.07	0.64±0.08	0.68±0.17
Gastric Cancer	0.52±0.08	0.49±0.03	0.49±0.04	0.53±0.07
Rheumatoid Arthritis	0.61±0.19	0.63±0.18	0.47±0.03	0.61±0.19
Pediatric Crohn's	0.59±0.13	0.63±0.18	0.63±0.20	0.59±0.13
Held-out diseases - TEDDY-G embeddings as predictive features				
CKD	0.90±0.03	0.96±0.01	0.95±0.02	0.90±0.03
Alzheimers	0.74±0.11	0.92±0.05	0.86±0.07	0.75±0.12
Gastric Cancer	0.62±0.06	0.71±0.08	0.70±0.09	0.60±0.07
Rheumatoid Arthritis	0.70±0.14	0.94±0.09	0.93±0.10	0.73±0.16
Pediatric Crohn's	0.63±0.15	0.75±0.15	0.72±0.16	0.63±0.15

D.1. Zero-shot embeddings evaluation

We evaluated the clustering performance using four key metrics from scib (Luecken et al., 2022) package in Table A6: Normalized Mutual Information (NMI), which measures the shared information between clustering results and true cell labels (0 to 1, higher is better); Adjusted Rand Index (ARI), which assesses similarity while adjusting for chance (-1 to 1,

Table A3. Accuracy of different foundation models on held-out diseases task.

	TEDDY-G 400M	Nicheformer	SCGPT	Geneformer	Geneformer 12L
Held-out diseases - Fine-tuning					
CKD	0.94±0.03	0.92±0.04	0.92±0.02	0.91±0.05	0.92±0.04
Alzheimers	0.75±0.18	0.76±0.16	0.70±0.12	0.70±0.06	0.67±0.06
Gastric Cancer	0.56±0.09	0.64±0.11	0.59±0.06	0.56±0.05	0.53±0.01
Rheumatoid Arthritis	0.56±0.06	0.51±0.01	0.52±0.02	0.51±0.01	0.50±0.00
Pediatric Crohn's	0.74 ±0.18	0.71±0.19	0.63±0.10	0.65±0.08	0.67±0.13

Table A4. Adding the ontology classification training objective to the gene modeling objective during pre-training does not affect gene-modeling loss, both for TEDDY-G and TEDDY-X.

	Gene modeling loss
TEDDY-G	2.48
- ontology classification (TEDDY-G 70M)	2.48
TEDDY-X	0.123
- ontology classification (TEDDY-X 70M)	0.125

higher is better). NMI and ARI are computed based on Louvain clusters generated from the embedding space; Average Silhouette Width (ASW), which indicates cluster separation by computing within-cluster and between-cluster distances, and dividing this by the larger of the two values (0 to 1, higher is better); and Average Bio (AvgBIO), the arithmetic mean of ASW, NMI, and ARI. We used resolution optimized Leiden clustering using the `cluster_optimal_resolution` function with default settings provided by scib.

In addition, we generate UMAPs using the following parameters in Figure A3: we first perform Principal Component Analysis (PCA) on the dataset, setting the number of components to 50 with the "svd_solver" set to "arpack" and a random seed of 42. Next, we compute the neighborhood graph with 15 neighbors, also using the same random seed. Finally, we apply UMAP to visualize the data, with the same seed.

Table A5. Performance on held-out donors as a function of model size. Fine-tuning F_1 scores improve for both TEDDY-G and TEDDY-X with increasing size.

	TEDDY-G	TEDDY-X
70M parameters	0.65 ± 0.02	0.52 ± 0.00
160M parameters	0.70 ± 0.03	0.53 ± 0.00
400M parameters	0.68 ± 0.06	0.55 ± 0.03

Table A6. TEDDY-G 400M offers better zero-shot clustering performance.

	cell sub-clustering (low)				cell sub-clustering (intermediate)			
	NMI	**ARI**	**ASW**	**Average BIO**	**NMI**	**ARI**	**ASW**	**Average BIO**
Cell2Sentence 410M	0.7817	0.6385	0.5632	0.6611	0.7969	0.6768	0.5456	0.6731
Teddy-G 400M	0.8382	0.7154	0.6022	0.7186	0.8785	0.7929	0.5971	0.7562

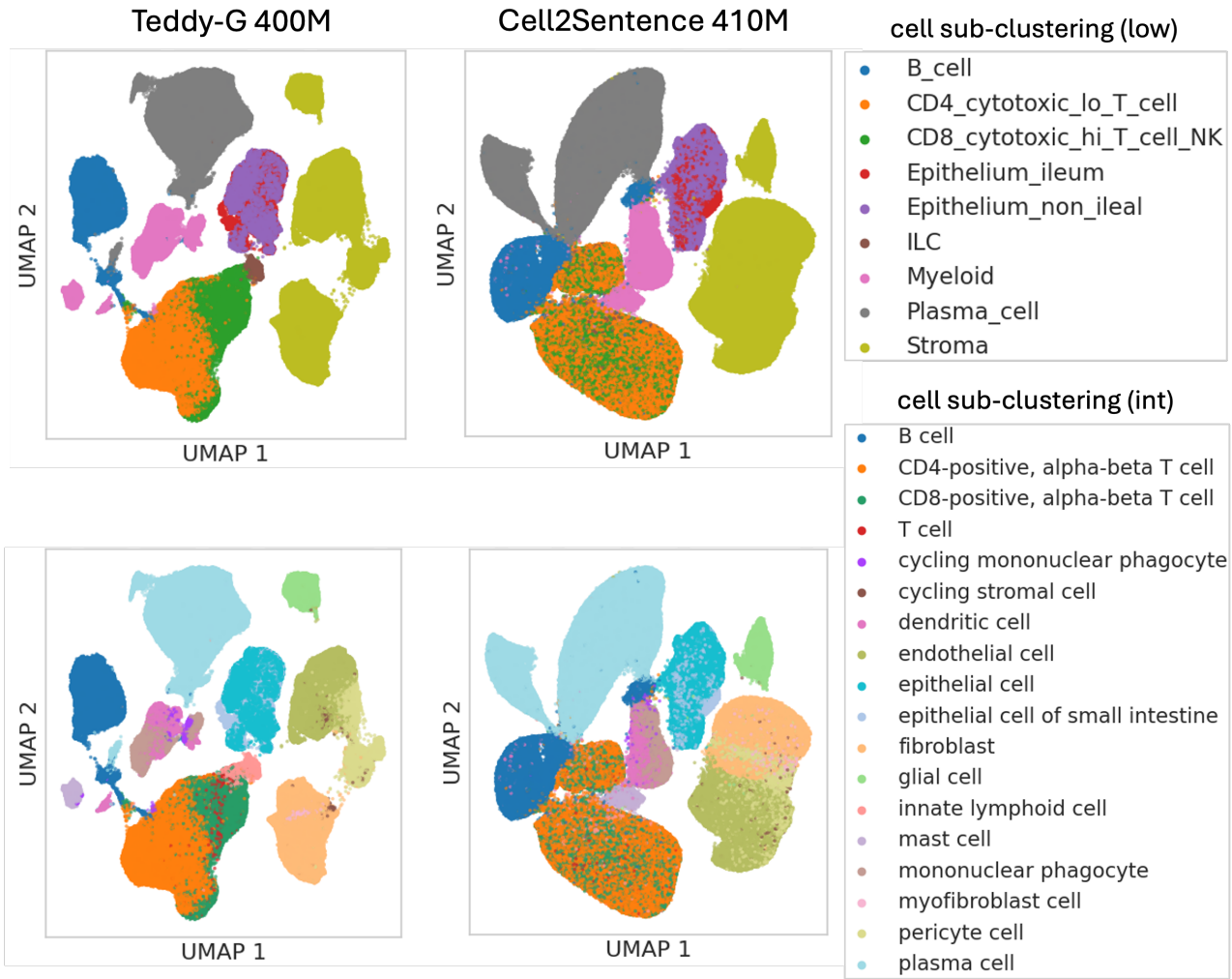

Figure A3. **Zero-shot embeddings.** Two dimensional umaps of zero-shot embeddings from TEDDY-G 400M (left) and Cell2Sentence 410M (right). The top row labels the umaps with coarse-grained cell labels while the bottom row uses finer-grained labels.