Evaluating the role of pre-training dataset size and diversity on single-cell foundation model performance

Alan DenAdel[1], Madeline Hughes[2], Akshaya Thoutam[2], Anay Gupta[3], Andrew W. Navia[4], Nicolo Fusi[2], Srivatsan Raghavan[4-7,§], Peter S. Winter[4,§], Ava P. Amini[2,§,*], and Lorin Crawford[2,§,*]

[1]Center for Computational Molecular Biology, Brown University
[2]Microsoft Research
[3]School of Computer Science, Georgia Institute of Technology
[4]Broad Institute of MIT and Harvard
[5]Department of Medical Oncology, Dana-Farber Cancer Institute
[6]Harvard Medical School
[7]Department of Medicine, Brigham and Women's Hospital
[§]Authors Contributed Equally
[*]Corresponding Email(s): ava.amini@microsoft.com; lcrawford@microsoft.com

Abstract

The success of transformer-based foundation models on natural language and images has motivated their use in single-cell biology. Single-cell foundation models have been trained on increasingly larger transcriptomic datasets, scaling from initial studies with 1 million cells to newer atlases with over 100 million cells. This study investigates the role of pre-training dataset size and diversity on the performance of single-cell foundation models on both zero-shot and fine-tuned tasks. Using a large corpus of 22.2 million cells, we pre-train a total of 375 models which we evaluate by conducting 3,750 experiments. Our results show that current methods tend to plateau in performance with pre-training datasets that are only a fraction of the size.

Introduction

The development of single-cell transcriptomic assays has empowered the dissection of biological heterogeneity at multiple scales. Machine learning models have been effectively applied to a range of tasks using single-cell data as input, including batch integration, imputation, perturbation prediction, and reference mapping. In parallel, large general-purpose neural network models, known as foundation models, have revolutionized the fields of natural language processing (NLP) and computer vision [1], and they have also led to groundbreaking progress in protein modeling[2–4]. Foundation models typically follow a transfer learning paradigm where the model is trained on a generic task, often predicting masked tokens, and then fine-tuned for specific tasks of interest [1]. Notably, recent NLP foundation models have shown strong performance in zero-shot and few-shot scenarios, where no task-specific fine-tuning is performed [5].

In single-cell biology, foundation models hold the potential to enable complex analyses without requiring large amounts of new data for each new scientific question. Recently, several foundation models for single-cell transcriptomics have been developed, including scBERT [6], Geneformer [7], scGPT [8], and

scFoundation [9]. Although these models differ in their tokenization and training schemes, most rely on the transformer architecture with self-attention [10]. Typically trained on tens of millions of single-cell transcriptomes, these models are applied to diverse tasks such as batch integration, cell type annotation, gene regulatory network inference, and perturbation prediction. The ambition of these single-cell foundation models (scFMs) is to achieve state-of-the-art performance across these tasks and beyond. While foundation models in NLP and computer vision exhibit remarkable zero- and few-shot capabilities [5, 11], the extent to which scFMs can demonstrate analogous levels of performance remains uncertain [12, 13].

Pre-training transformer-based models typically requires expensive, specialized hardware such as graphics processing units (GPUs), tensor processing units (TPUs), or neural processing units (NPUs), as well as days to weeks of time for computation. To date, many scFMs are closed-source or only provide source code for fine-tuning and evaluation, rather than releasing the source code needed to pre-train a model from scratch. Keeping these computational and practical challenges in mind, more data is not always better. For a fixed training budget, curating a high-quality subset of the data and making additional training passes on that subset may yield better model performance [14]. Although there is some evidence suggesting that current large language models may be under-trained [15, 16], this phenomenon has not yet been evaluated in current scFMs. Nevertheless, developers of scFMs continue to scale up both model and training dataset size. Curating datasets with increased quality and diversity, rather than focusing solely on size, could plausibly improve the performance of scFMs.

In this work, we investigate the role of pre-training dataset size and diversity on the performance of scFMs (Fig. 1). We pre-train a total of 375 models which we evaluate by conducting 3750 experiments across a comprehensive suite of factors including both in zero-shot scenarios and when fine-tuned for specific tasks. Our results indicate that current scFMs do not benefit from increasing pre-training dataset size or from naive approaches to increasing pre-training dataset diversity.

Results

We evaluated the performance of three model architectures of increasing complexity: (1) a variational autoencoder (scVI [17]), (2) a masked autoencoder (SSL [18]), and (3) a transformer (Geneformer [7]). Detailed descriptions of the specific parameters and training procedures for each model architecture are described in the Methods section.

We pre-trained each model architecture on datasets of varying size and diversity (Fig. 1A). These pre-training datasets were sampled from the 22.2 million cell scTab corpus of single-cell transcriptomes [19]. To generate these datasets, we implemented three distinct downsampling schemes: (1) random downsampling, which selects cells uniformly at random; (2) cell type re-weighting, which attempts to include equal proportions of each cell type in the dataset; and (3) geometric sketching, which samples cells uniformly across transcriptional space without relying on metadata [20]. Random downsampling conserves diversity (i.e., all subsets have diversity that is approximately equal to the full dataset), while both cell type re-weighting and geometric sketching increase diversity relative to the full corpus. Pre-training datasets were generated using each of these three downsampling schemes at sizes of 1%, 10%, 25%, 50%, and 75% of the total training corpus. Each pre-training corpus was split into train, validation, and test sets in proportions of 80%, 10%, and 10%, respectively. To ensure robust evaluation, we generated five independent splits for each combination of downsampling scheme and dataset size, resulting in 80 pre-training datasets. All three model architectures were pre-trained on these datasets, yielding a total of 240 pre-trained models for this initial analysis.

To confirm that each downsampling scheme had the intended effect on dataset diversity, we computed three diversity metrics for each pre-training dataset: (1) the Shannon index [21], (2) the Gini-Simpson index [22], and (3) the Vendi Score [23]. The Shannon index and the Gini-Simpson index evaluate diversity based on metadata (in this case, the cell type labels), while the Vendi Score directly evaluates the transcriptional diversity. Both cell type re-weighting and geometric sketching increased diversity

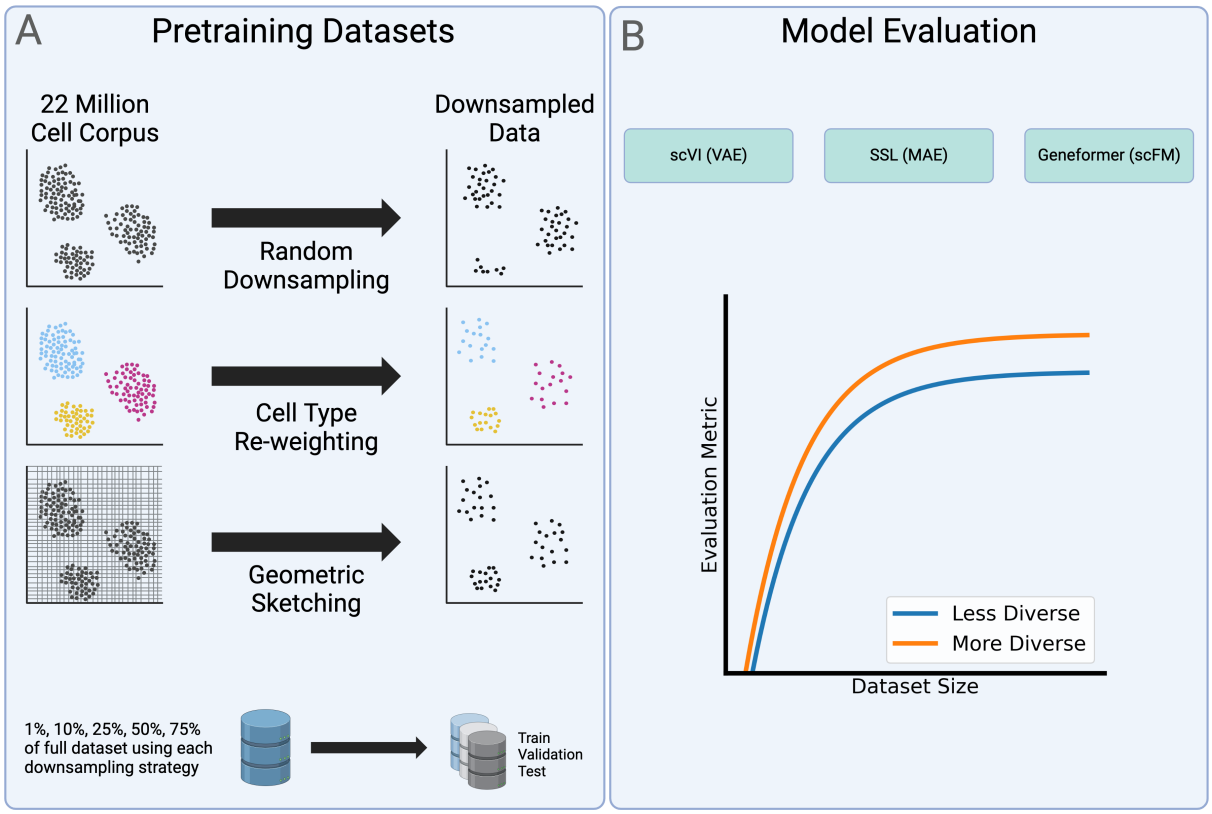


Figure 1. Strategy to assess the effects of pre-training dataset size and diversity on scFM performance. **(A)** Schematic of the downsampling approaches, sizes of downsampled pre-training datasets, and data splitting strategy. **(B)** An example of what evaluation performance might *a priori* be expected to look like as a function of pre-training dataset size and diversity.

compared to random downsampling (Supplemental Fig. 1). Interestingly, geometric sketching resulted in higher Vendi Scores compared to cell type re-weighting, whereas cell type re-weighting resulted in higher Shannon and Gini-Simpson index values. These differences reflect the underlying mechanisms of the two downsampling methods: cell type re-weighting enhances diversity based on cell type metadata, while geometric sketching directly targets transcriptional data.

To test whether increased dataset size and/or diversity improved performance (Fig. 1B), we evaluated all 240 pre-trained models in the two primary regimes in which a scFM might be utilized: (1) zero-shot and (2) fine-tuned. In the zero-shot regime, the model's internal representation of data, referred to as an embedding, is used directly without additional training. In contrast, the fine-tuning regime involves re-training all or part of a model to address a specific task. Fine-tuning typically tackles a supervised learning problem where both labels and the task are well-defined, while the zero-shot approach is suited for tasks where no ground truth labels are available.

A common task in many single-cell RNA-sequencing (scRNA-seq) workflows is labeling cell types based on their transcriptional profiles. To evaluate initial model performance, we used all pre-trained models to classify cells from a clonal hematopoiesis dataset [24] in both the zero-shot scenario and after fine-tuning the models to classify cell types present in the dataset. As simple zero-shot baselines for comparison, we included embeddings derived from highly variable genes (HVGs) and from the top 50 principal components. For a fine-tuned baseline, we included a regularized logistic classifier using HVGs as input features. Model performances, as assessed by the micro F1 score, are shown in Fig. 2. Interestingly, in the zero-shot case, Geneformer saw no increase in performance as a function of pre-training dataset size across all downsampling schemes (Fig. 2A). Similarly, scVI and SSL saw minimal improvement from increasing pre-training dataset size beyond 10% of the entire corpus, except in the case of geometric sketching. When the models had been fine-tuned on the clonal hematopoiesis dataset, this pattern was amplified (Fig. 2B). Neither SSL nor Geneformer showed any performance increase from the pre-training dataset size, while scVI exhibited a slight improvement in performance going from 1% of the full corpus to 10% of the full corpus, but no improvement beyond that (except for geometric sketching, which demonstrated reduced performance overall). Model performance was comprehensively assessed using accuracy, precision, recall, micro F1 score, and macro F1 score in both the zero-shot scenario (Supplemental Fig. 2) and the fine-tuning scenario (Supplemental Fig. 3). Across all metrics, the phenomenon of model performance plateauing at a small fraction of the total training data was consistently observed. Furthermore, increasing the diversity of pre-training datasets did not improve model performance in this classification task.

To determine whether the performance plateau observed on the clonal hematopoiesis dataset generalized to other datasets, we evaluated cell type classification in both the zero-shot and fine-tuned scenarios across several distinct datasets representing diverse biological contexts: (1) clonal hematopoiesis (as shown in Fig. 2) [24], (2) intestine-on-a-chip [25], (3) placental infection (which we refer to as "placenta") [26], and (4) periodontitis [27]. Importantly, these datasets are not present in the scTab subset of CELLxGENE and thus represent independent test sets.

For each evaluation dataset, we sought to identify the pre-training corpus fraction where performance gains became marginal, which we refer to as the "learning saturation point" (Fig. 3A). For each combination of model, evaluation dataset, and downsampling scheme, we identified the maximum Micro F1 score across all pre-training dataset sizes and defined the "saturation threshold" as 95% of this value. We then identified the smallest percentage of the training data that produced a model surpassing this threshold and labeled this percentage the "learning saturation point".

Across all model architectures, downsampling strategies, and evaluation datasets, zero-shot classification performance typically saturated at a small fraction of the total pre-training corpus (Fig. 3B and Supplemental Figs. 4-6), and fine-tuned performance typically saturated at 1% of the training data and never required more than 10% to reach saturation (Fig. 3C and Supplemental Figs. 7-9). For scVI, zero-shot learning saturation largely occurred at 1% and never exceeded 10%; in contrast, for SSL, sat-

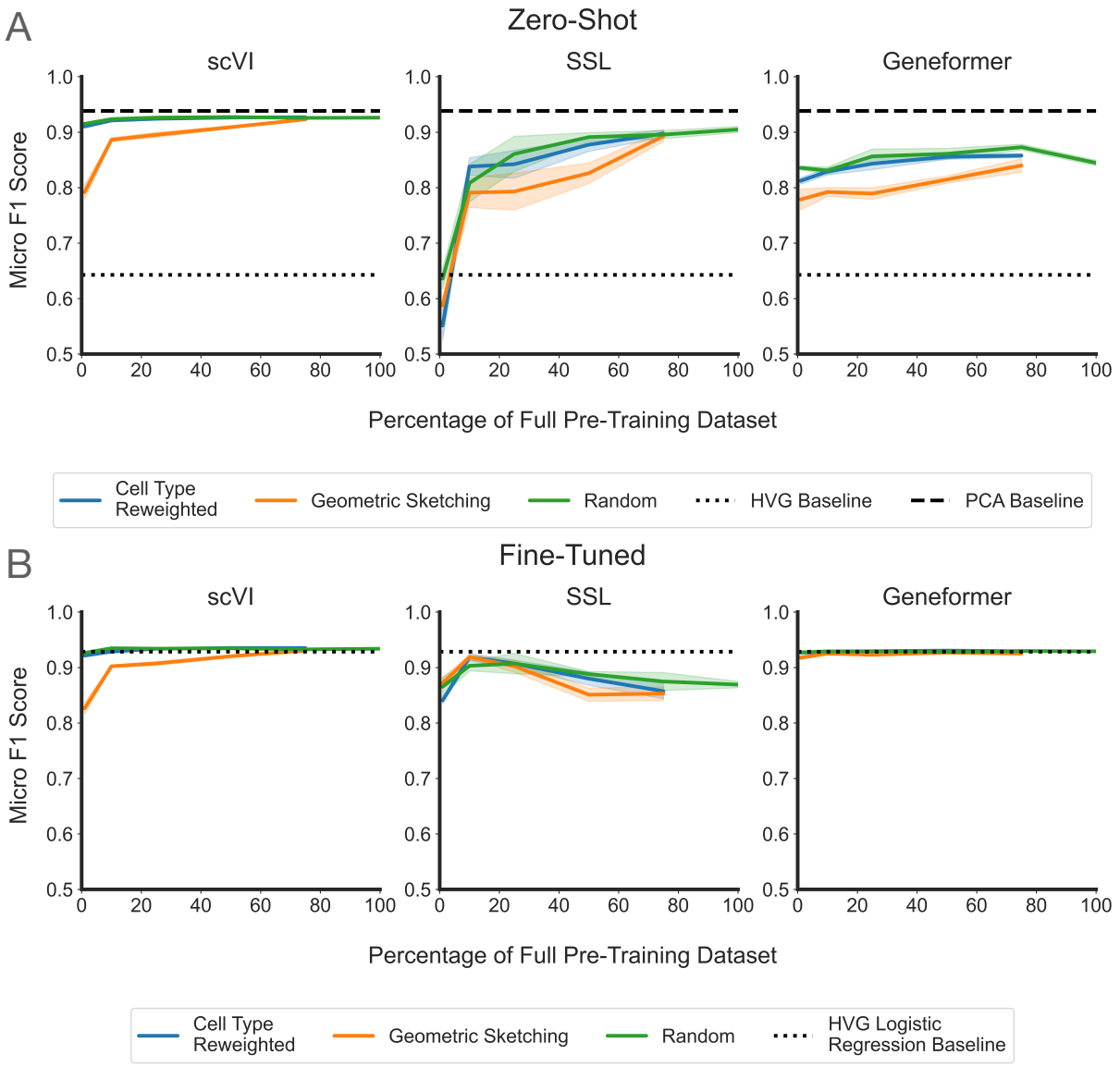


Figure 2. Zero-shot and fine-tuned performance on classifying cells from a clonal hematopoiesis dataset plateaus at a small fraction of the total data available for pre-training. **(A)** Line plots showing zero-shot classification performance for each model's embeddings, as evaluated by the Micro F1 Score. For each model, the different colors correspond to the downsampling strategy used to generate the data used for pre-training. The dotted line shows the performance of using the highly variable genes as an embedding; the dashed line shows the performance of using principal component projections as an embedding. **(B)** Line plots showing classification performance for each model after fine-tuning, as evaluated by the Micro F1 Score. For each model, the different colors correspond to the downsampling strategy used to generate the data used for pre-training. The dotted line shows the performance of training a regularized logistic regression classifier using the highly variable genes as input features.

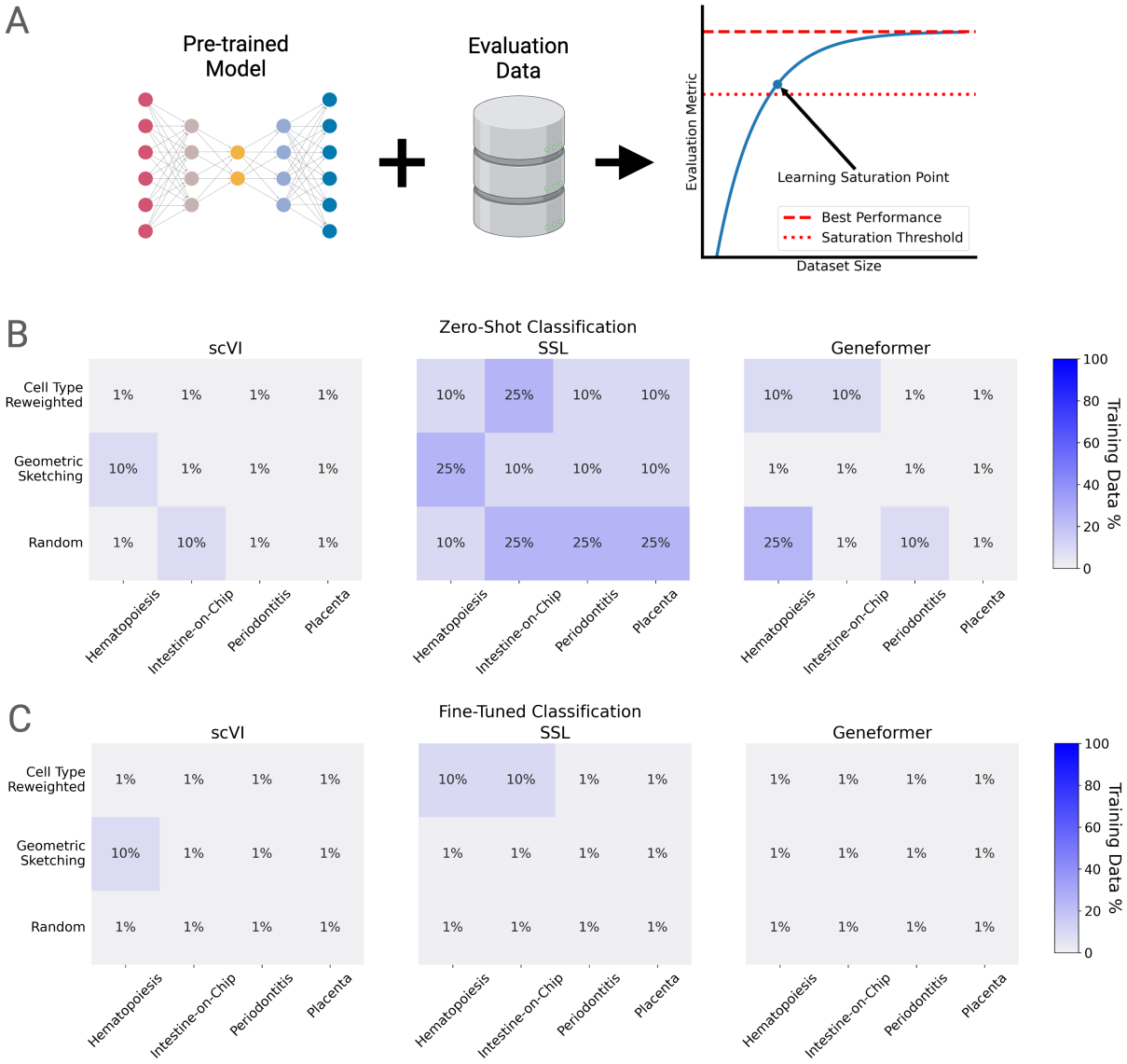


Figure 3. Performance saturation as a function of pre-training dataset size across multiple evaluation datasets. (A) Schematic of analysis to find the learning saturation point. For each family of models (i.e., a downsampling strategy paired with a model) a "saturation threshold" of 95% of the maximum performance was computed, and the minimum pre-training dataset size that produced a model surpassing that threshold was identified. This dataset size was denoted the "learning saturation point" and is considered the point at which model performance saturated as a function of pre-training dataset size. **(B)** Heatmap visualizing the learning saturation point for the clonal hematopoiesis, intestine-on-chip, periodontitis, and placental infection datasets for each of scVI, SSL, and Geneformer, across each downsampling strategy and when evaluated in the zero-shot regime. Each sub-panel corresponds to the model architecture, the x-axis corresponds to the dataset evaluated, and the y-axis corresponds to the downsampling strategy used to pre-train each model. **(C)** Heatmap visualizing the learning saturation point for the clonal hematopoiesis, intestine-on-chip, periodontitis, and placental infection datasets for each of scVI, SSL, and Geneformer, across each downsampling strategy and when evaluated in the fine-tuning regime. Each sub-panel corresponds to the model architecture, the x-axis corresponds to the dataset evaluated, and the y-axis corresponds to the downsampling strategy used to pre-train each model.

uration occurred at 10% or 25% in the zero-shot case. Geneformer overwhelmingly saturated at 1% or 10%, though it reached 25% in one case. Across all four evaluation datasets, more diverse pre-training datasets did not demonstrate improved performance relative to the randomly downsampled datasets. These results suggest that, for cell type classification, the learning saturation point for scFMs can occur as early as 1% of the pre-training corpus – amounting to approximately 200,000 cells. This trend holds for both transformer-based and non-transformer model architectures, as well as for both random and diversity-based downsampling schemes.

We also assessed the performance of the model embeddings on the task of batch integration (Supplemental Figs. 10, 11). Since obtaining ground truth scRNA-seq data that has been batch effect corrected is not possible, batch integration is inherently a zero-shot task. For this task, model embeddings were evaluated in terms of how well cells of the same type—but from different batches—were mixed in the embedding space. We utilized two datasets that contained cells from multiple batches: (1) the periodontitis dataset [27], previously used in the classification evaluation, and (2) a lung dataset [28] for this evaluation.

We adopted the batch integration evaluation methodology from Kedzierska et al. [12] and Luecken et al. [29]. This involved generating cluster assignments based on Louvain clusters obtained from the embedding space and comparing them to the ground truth cell type labels. The quality of these cluster assignments was assessed using normalized mutual information (NMI) and adjusted Rand index (ARI). To evaluate the degree of batch mixing in the embedding space, we calculated the average silhouette width (ASW) with respect to both cell type labels and batch labels. The average BIO (AvgBIO) score was then used to summarize overall performance across these metrics, combining ASW, NMI, and ARI. Results for the periodontitis dataset are presented in Supplemental Fig. 10 and for the lung dataset in Supplemental Fig. 11.

We repeated the learning saturation point analysis for the AvgBIO batch integration performance curves across all models, evaluation datasets, and downsampling schemes on this batch integration task, and observed that, consistent with the classification task analysis, model performance generally saturated at a small fraction of the total pre-training corpus (Supplemental Fig. 12). Specifically, saturation of batch integration performance occurred at 1% for scVI, 10% or 25% for SSL, and between 1% and 25% for Geneformer.

Since increasing the diversity of the pre-training datasets did not improve models' performance on downstream tasks, we reasoned that a small fraction (at maximum, 25%) of the scTab corpus was sufficient to represent cells with a normal transcriptional phenotype. We thus hypothesized that supplementing the pre-training corpus with cells that had been systematically perturbed could enhance the contribution of pre-training on downstream performance.

Here, we performed two sets of spike-in experiments using genome-wide Perturb-seq data from Replogle et al. [30], which contains over 2.5 million systematically perturbed cells. For both sets of experiments, we started with the uniformly downsampled scTab pre-training datasets. For each of these datasets, we generated two new datasets by "spiking in" cells from the Replogle Perturb-seq experiments. This involved replacing a portion of the scTab cells with an equal number of cells from the Replogle Perturb-seq data. The spike-in experiments were conducted at two proportions of spike-in perturbation data: 10% and 50%. The Replogle dataset had enough cells to replace 10% of the scTab cells in all of the uniformly downsampled datasets. However, there were only enough cells to replace 50% of the scTab cells in the 1%, 10%, and 25% downsampled datasets. These experiments allowed us to assess whether incorporating perturbed cells could increase the value of pre-training for downstream tasks.

To evaluate the effect of adding in perturbation data, we analyzed the learning saturation points of the Micro F1 score performance curves for the cell type classification task, evaluating all models in both the zero-shot (Supplemental Fig. 13) and fine-tuning (Supplemental Fig. 14) cases. Consistent with the initial analysis of the classification task, model performance generally saturated at a small fraction of the total pre-training corpus. In the fine-tuned case, all model architectures typically saturated at 1%,

although Geneformer saturated at 25% for the 10% spike-in in one instance. For scVI, saturation in the zero-shot case typically occurred at 1% and never exceeded 10%. In contrast, SSL saturation occurred between 10% and 25% in the zero-shot case. Finally, Geneformer typically saturated at 1% or 10% in the zero-shot case.

We also repeated the learning saturation point analysis on the AvgBIO performance curves for the batch integration task, evaluating all models pre-trained on the spike-in datasets (Supplemental Fig. 15). Similar to the classification tasks, across all combinations of model, evaluation dataset, and downsampling schemes, model performance generally saturated at a small fraction of the total pre-training corpus, even with the inclusion of spike-in data. When pre-trained on the spike-in datasets, scVI consistently saturated at 1%, SSL saturated between 10% and 25%, and Geneformer typically saturated at 1% or 10%. From these evaluations in both cell type classification and batch integration, it is clear that naively spiking-in perturbation data does not facilitate an improvement in model performance.

Discussion

In this study, we investigated the role of pre-training dataset size and diversity on the performance of general-purpose single-cell foundation models (scFMs). We evaluated three representative general-use models for single-cell RNA sequencing data on downstream tasks in both the zero-shot case and after fine-tuning the models for specific tasks. Our work addresses a critical consideration in training large-scale models: the size and diversity of the pre-training corpus. While neural scaling laws observed in other domains suggest that increasing dataset size leads to better model performance, our findings demonstrate that, past a learning saturation point, simply increasing pre-training datasets, even with greater diversity, does not necessarily improve performance on downstream tasks. These results suggest that further scaling up of pre-training datasets from tens of millions of cells to hundreds of millions or even billions of cells without appropriately modeling the non-sequential nature of single-cell data may not yield tangible returns.

Importantly, our analysis used the scTab dataset as-is, without additional quality control or data curation beyond downsampling. Recent work has highlighted that some reference atlases contain substantial numbers of low-quality cells and contaminants [31]. Our findings underscore the importance of prioritizing data quality and content over sheer dataset size. Developers of scFMs and large databases should consider data quality and content rather than attempting to gather the largest datasets possible, which we have shown is unlikely to meaningfully improve performance.

Our results highlight the need for a more nuanced approach, balancing dataset size and diversity with careful attention to model architectures and model benchmarking. The experiments showing that spiking in perturbation data does not necessarily improve performance on cell type classification or batch integration additionally emphasize the importance of using pre-training data that is aligned with the downstream tasks for which a model will be applied. While conceptualizing a cell as "a bag of RNA" [32] is a useful starting point for many models, the concept of a cell as "a sentence of genes" [6–9] may be less effective. Well-designed embeddings have long been used to great success in genomics [17, 33], and simple embeddings can capture a large amount of the information present in single-cell data [12, 34]. Future work should focus on understanding the interplay between feature representation, model architecture, and pre-training dataset size to optimize performance on downstream tasks.

Methods

Pre-training data and downsampling schemes

The main pre-training dataset used in this study is scTab [19] which is a subset of the CELLxGENE [35] census (version 2023-05-15). In this particular dataset, there are a total of 22.2 million cells from

164 unique cell types, 5,052 unique human donors, and 56 human tissues. For each of the downsampling strategies described below (uniform random downsampling, cluster adaptive downsampling, and geometric sketching downsampling), cells in scTab were downsampled in percentages of 1%, 10%, 25%, 50%, and 75% with five replicates for each percentage. Each of these downsampled datasets was then divided into training, validation, and test sets in proportions of 80%, 10%, and 10% splits, respectively. We also downsampled the entire scTab dataset (denoted as the 100% dataset) into 80% training, 10% validation, and 10% test sets with five replicates of this split.

Uniform downsampling. Uniform downsampling is the simplest scheme used. The uniformly downsampled datasets sampled each cell uniformly at random without consideration for any additional metadata or the expression counts matrix.

Cell type re-weighted downsampling. The goal of the cell type re-weighted scheme is to downsample such that each cell type is present in equal proportions in the resulting dataset. Let $p = (p_1, \ldots, p_K)$ be the proportion of each cell type in the scTab dataset where K is the total number of cell types. For many cell types present in the scTab dataset, there are not enough cells to satisfy this goal even when downsampling to 1%. Therefore, to generate a cell type re-weighted dataset, we first choose all cells that belong to a cell type that is not common enough to satisfy the equal proportion condition. Then, we randomly downsample among the remaining cell types with equal proportions. Let N be the total number of cells and D be the downsampling proportion. The algorithm is as follows:

1. For any k-th cell type satisfying $p_k ND < ND/K$ (i.e., it is impossible to satisfy the re-weighting condition, even by selecting all of the cells of this cell type), include all of the cells of this cell type.
2. Let N' be the number of unselected cells after the previous step and K' be the number of remaining cell types. Some of these cell types that previously satisfied the re-weighting condition may no longer do so with N' and K'. Repeat step 1 with N' in place of N and K' in place of K.
3. Let M be the number of cells already selected. When all remaining cell types satisfy the re-weighting condition, sample $(DN - M)/K'$ cells from each cell type (i.e., re-weight these cell types to have equal proportions).

The resulting downsampled dataset should have proportions $p^* = (p_1^*, \ldots, p_K^*)$ such that $p_1^* \approx p_2^* \approx \cdots \approx p_K^* \approx 1/K$.

Geometric sketching. Geometric sketching is a method that downsamples based on the intrinsic diversity of a dataset without reference to any metadata [20]. Geometric sketching samples uniformly over transcriptional space without regard to density (i.e., a cell type that has a large number of cells with very similar gene expression will be downweighted in the resulting "sketch"). Let N be the total number of cells, D be the downsampled proportion, and C be the number of covering hypercubes. The geometric sketching algorithm is as follows:

1. Generate a "plaid covering" of the dataset in transcriptomic space. A plaid covering is a covering of the dataset consisting of equally sized hypercubes. Each hypercube in the covering will contain a different number of cells.
2. For each covering hypercube, sample DN/C cells.

Due to the fact that each hypercube is of equal size, higher-density regions of transcriptomic space (i.e., regions with many similar cells) will be downweighted in the resulting "sketch" of the dataset.

Spike-in Experiments with Perturbation Data

In order to assess the extent to which including only normal transcriptomic phenotypes impacts downstream model performance, we performed two sets of "spike-in" experiments using the Replogle et al. [30] Perturb-seq study (CRISPR-based screens with single-cell RNA-sequencing readouts) archived in the scPerturb database [36]. The Replogle dataset contains 2,547,877 cells and 10,117 genes. For both sets of experiments, we began with the downsampled datasets generated using the uniform downsampling scheme. Then, for each of these datasets, we generated two new datasets that had cells from the Replogle Perturb-seq experiments "spiked in". This was done by effectively removing some of the scTab cells and replacing them with an equal number of cells from the Replogle study. We spiked in these cells at rates of 10% and 50%. The Replogle dataset had enough cells to replace 10% of the scTab cells in all of the uniformly downsampled datasets; however, there were only enough cells to replace 50% of the scTab cells in the 1%, 10%, and 25% downsampled datasets.

Evaluation Datasets

Below is a description of each dataset we used for evaluation in this study. These were all downloaded from the CELLxGENE census and can be found online at `https://cellxgene.cziscience.com/`.

Clonal hematopoiesis. The clonal hematopoiesis data results from profiling peripheral blood gene expression from 17 patients with clonal hematopoiesis and 7 controls [24]. For the zero-shot evaluations, the data were divided into 50/50 train/test splits. For the fine-tuning evaluations, the data were divided into 80/10/10 train/test/validation splits. There are 66,985 total cells and 8 labeled cell types in this dataset.

Placenta. The placenta data results from profiling placental responses to three pathogens [26]. For the zero-shot evaluations, the data were divided into 50/50 train/test splits. For the fine-tuning evaluations, the data were divided into 80/10/10 train/test/validation splits. In this dataset, there are 20 donors (counting both mother and fetus), 158,978 total cells, and 7 labeled cell types.

Intestine-on-chip. The intestine-on-chip data results from profiling iPSC-derived small intestine-on-chips that were exposed to either: (1) different media conditions or (2) interferon-beta or interferon-gamma [25]. For the zero-shot evaluations, the data were divided into 50/50 train/test splits and for the fine-tuning evaluations, the data were divided into 80/10/10 train/test/validation splits. There are 22,280 total cells and 8 labeled cell types.

Periodontitis. The periodontitis data results from profiling the periodontium of 34 individuals [27]. For the zero-shot evaluations, the data were divided into 50/50 train/test splits and for the fine-tuning evaluations, the data were divided into 80/10/10 train/test/validation splits. There are 105,918 total cells and 27 labeled cell types.

Pre-training of machine learning models

In this study, we implemented three machine learning methods for single-cell transcriptomics: scVI [17], SSL [18], and Geneformer [7]. For each model class, we standardized the amount of compute (in terms of training steps) that was allocated to pre-training so that simply increasing the pre-training dataset size did not increase the amount of training performed. To determine the number of training epochs allocated to each model, we split the entire scTab corpus into an 80/10/10 split of train, test, and validation datasets, respectively. We then trained each model class on the training split until the validation loss had converged. Next, we plotted the validation loss as a function of the number of epochs and selected the

number of epochs that resulted in convergence of the loss function. Finally, this value was normalized based on the fraction of training data used in order to determine the number of epochs needed for each training split. For example, if the number of epochs taken to converge on the full corpus was 40 and we want to find the number of epochs for 25% of the corpus, then we would use $40/0.25 = 160$ epochs for this split of the data.

Single-cell variational inference (scVI). The default parameters were used to designate the architecture of the pre-trained scVI models. In particular, the number of nodes per hidden layer was 128, the latent space had 10 dimensions, the encoder and decoder both have a single hidden layer, the dropout rate was 10%, the gene likelihood was zero-inflated negative binomial, the dispersion parameter of the negative binomial distribution was constant per gene, and the latent distribution was normal. For training, the batch size was 128, early stopping with respect to the validation loss was turned on, and the number of epochs $40 \times (100/P)$ where P is the training proportion. For example, for a 25 percent split the number of epochs is $40 \times (100/25) = 160$ epochs.

Self-supervised learning (SSL). The parameters specified in the SSL repository on Github were used to specify the model architecture and pre-training parameters (except for the number of epochs, which was determined as described above). In particular, the hidden layer sizes were (512, 512, 256, 256) for the encoder (reversed for the decoder) and the latent space dimension was 64. For training, the batch size was 16,384, the masking rate was 0.5, the dropout rate was 0.1, weight decay was set to 0.01, and the learning rate was 0.001.

Geneformer. The parameters specified in the Geneformer repository on Hugging Face were used to specify the Geneformer architecture and pre-training parameters (except for the number of epochs, which was determined as described above). In particular, the model type was BERT (bidirectional encoder representations from transformers), the maximum input size was 2048, the number of layers was 6, the number of attention heads was 4, the number of embeddings dimensions was 256, the intermediate size was 512 (i.e., twice the number of embedding dimensions), and the activation function was ReLU. For training, the batch size was 18 per GPU (across 8 GPUs total), the maximum learning rate was 0.001, the learning schedule was linear, the number of warmup steps was 10,000, the attention dropout was 0.02, the hidden layer dropout was 0.02, the optimizer was AdamW, the layer norm epsilon was 1×10^{-12}, and the initializer range was 0.02, and the weight decay parameter was 0.001.

Note on pre-processing datasets for each model. The software for scVI takes in raw gene expression counts. As a result, we did not perform any pre-processing on the pre-training or evaluation data before applying scVI. Similarly, the Geneformer software requires raw gene expression counts before it tokenizes the data; thus, we do not perform any pre-processing before this step in the model. For SSL, gene expression profiles were normalized to 10,000 counts per cell and then logarithmized with a pseudocount of 1 (i.e., processed with the function $f(\mathbf{x}) = \log(\mathbf{x} + 1)$).

Description of fine-tuning

Below is a description of how we fine-tune pre-trained versions of scVI, Geneformer, and SSL.

Single-cell variational inference (scVI). The variational autoencoder implemented in the scVI library does not have a fine-tuning mechanism built-in. To get around this fact, we implemented a simple multi-layer perceptron (MLP) that utilized the latent space of an scVI model as input. We utilized an MLP that closely matched the architecture of the scVI decoder, but whose output predicted the cell type label rather than reconstructing gene expression. The model architecture was an input layer of 10

dimensions (i.e., the latent space of the pre-trained scVI models), a hidden layer of 128 neurons, a ReLU function after the hidden layer, and an output linear layer the size of the number of cell type labels. This architecture was chosen to be simple and was not optimized. For training, we used a dropout rate of 0.2, the Adam optimizer with learning rate 1×10^{-4}, cross-entropy loss, early stopping based on the validation loss, and 500 maximum epochs.

Self-supervised learning (SSL). To fine-tune SSL, we used the built-in cell type classification model from the SSL library. We used the default model architecture where the hidden units had sizes of 512, 512, 256, 256, and 64. For training, we used a learning rate of 9×10^{-4}, a dropout rate of 0.1, a weight decay of 0.05, a batch size of 16384, a minimum of 30 epochs, used early stopping based on validation loss, and the same number of maximum epochs used during pre-training. The learning rate and weight decay were non-default values and were taken from the SSL training scripts.

Geneformer. To fine-tune Geneformer, we used the built-in cell type classification model from the Geneformer library. We froze the first 2 layers of the pre-trained model. For training, we used a single epoch (following the example set by Theodoris et al. [7]), with 1000 warm-up steps, a linear learning rate schedule, a batch size of 18 per device, and 8 GPUs. We used a learning rate 0.000210031 and a weight decay 0.233414 as obtained by using Geneformer's hyperparameter optimizer.

Classification tasks and evaluation metrics

We used two simple embeddings as baselines for zero-shot cell type classification: (1) highly variable genes and (2) the cell loadings of the first 50 principal components computed on the dataset. As a baseline for fine-tuned cell type classification, we trained a regularized logistic classifier on the highly variable genes determined on the training dataset. In both cases, for a classification task with cell type labels $\mathbf{y} = (y_1, \ldots, y_N)$ and predicted cell types $\hat{\mathbf{y}} = (\hat{y}_1, \ldots, \hat{y}_N)$, we use the following metrics: accuracy, precision, recall, macro F1 score, and micro F1 score. **Accuracy** measures the fraction of cells that are correctly predicted

$$\text{accuracy}(\mathbf{y}, \hat{\mathbf{y}}) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}(\hat{y}_i = y_i) \tag{1}$$

where $\mathbb{1}(\bullet)$ is an indicator function. For a classification problem, **precision** and **recall** are defined as

$$\text{precision} = \frac{TP}{TP + FP}, \qquad \text{recall} = \frac{TP}{TP + FN} \tag{2}$$

where TP denotes the number of true positives, FP is the number of false positives, and FN is the number of false negatives. Lastly, the **F1 score** is the harmonic mean of the precision and recall such that

$$\text{F1} = \frac{\text{TP}}{\text{TP} + (\text{FP} + \text{FN})/2} \tag{3}$$

where, in addition to previous notation, TN is the number of true negatives. In this study, we assessed performance in datasets with multiple cell types. There are two variants of the F1 score that measure performance when there are more than two classes. The **Micro F1 Score** computes TP, FN, and FP globally across all cell types. The **Macro F1 Score** computes the binary F1 score for each individual label and then averages these F1 scores (which weights each cell type equally, regardless of if they are present in unequal proportions).

Batch integration task and evaluation metrics

We used two simple embeddings as baselines for zero-shot batch integration: (1) highly variable genes and (2) the cell loadings of the first 50 principal components computed on the dataset. For an integration task with cell type labels $\mathbf{y} = (y_1, \ldots, y_N)$, model embeddings are used to obtain cluster labels $\hat{\mathbf{y}} = (\hat{y}_1, ..., \hat{y}_N)$. Here, we use the following metrics: NMI (cluster/label), ARI (cluster/label), ASW (label), ASW (batch), and Average Bio as implemented by Luecken et al. [29] in the `scib` package.

Normalized mutual information (NMI). For two sets of cell type labels, the normalized mutual information (NMI) captures the similarity between the two labels. In the zero-shot integration task, the two labels are the annotated cell type assignments and the clusters derived from the model embeddings. The NMI is computed as follows:

$$\mathrm{NMI}(\mathbf{y}, \hat{\mathbf{y}}) = \frac{2I(\mathbf{y}, \hat{\mathbf{y}})}{H(\mathbf{y}) + H(\hat{\mathbf{y}})} \tag{4}$$

where $I(\mathbf{y}, \hat{\mathbf{y}})$ is the mutual information of $\mathbf{y}$ and $\hat{\mathbf{y}}$, and $H(\bullet)$ is the Shannon entropy.

Adjusted Rand index (ARI). For two sets of cell type labels, the adjusted Rand index (ARI) captures the similarity between the two labels. In the zero-shot integration task, the two labels are the annotated cell type assignments and the clusters derived from the model embeddings. It is based on the Rand index (RI) which is computed as the following

$$\mathrm{RI} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FP} + \mathrm{FN} + \mathrm{TN}}. \tag{5}$$

The ARI corrects for the RI measurement's sensitivity to chance via permutation [37] where

$$\mathrm{ARI} = \frac{\mathrm{RI} - \mathbb{E}[\mathrm{RI}]}{1 - \mathbb{E}[\mathrm{RI}]}. \tag{6}$$

Here, $\mathbb{E}[\mathrm{RI}]$ denotes the expectation of the Rand index.

Average silhouette width (ASW). We compute the average silhouette width considering cell type labels (ASW_L), in order to measure how well embeddings are retaining biological information. We then normalize this score via the following

$$\mathrm{ASW}_{\mathrm{label}} = \frac{\mathrm{ASW}_L + 1}{2}. \tag{7}$$

The metric $\mathrm{ASW}_{\mathrm{label}}$ ranges from 0 to 1 where higher scores indicate better cell type clustering. We also compute the average silhouette width considering batch labels (ASW_B), in order to measure how well embeddings are retaining batch effects. We then invert this score to measure how well embeddings are removing batch effects where

$$\mathrm{ASW}_{\mathrm{batch}} = 1 - |\mathrm{ASW}_B|. \tag{8}$$

The metric $\mathrm{ASW}_{\mathrm{batch}}$ also ranges from 0 to 1 where higher scores indicate better batch-mixing performance.

Average BIO (AvgBIO). The AvgBIO score is the arithmetic mean of ASW, NMI, and ARI where

$$\mathrm{AvgBIO} = \mathrm{ARI}_{\mathrm{cell}} + \mathrm{NMI}_{\mathrm{cell}} + \mathrm{ASW}_{\mathrm{cell}}. \tag{9}$$

This metric was introduced by Cui et al. [8] and is intended to measure the biological conservation of the embeddings being evaluated. AvgBIO is normalized to the unit scale with 1 indicating complete agreement between clusters and cell type labels.

Data Diversity Measures

We used three different diversity measures to assess the impact of the various downsampling schemes on the diversity of the data used for model pre-training. The Shannon index and Gini-Simpson index are based on cell metadata, while the Vendi Score is based on the gene expression values directly. For the two metadata-based measures, we use the following notation: let $p = (p_1, \ldots, p_K)$ be the proportion of each cell type in the dataset where K is the total number of cell types.

Shannon index. The Shannon index is a popular diversity measure in ecology [22] and is the application of Shannon entropy [21] to group proportions. In particular, the Shannon index is given by

$$H' = \sum_i p_i \log_2 p_i. \tag{10}$$

While the Shannon index can be defined using any base for the logarithm, in this study we use base 2. For the Shannon index, a higher value indicates increased diversity.

Simpson index and Gini-Simpson index. The Simpson index is also a popular diversity index in ecology [38] and represents the probability that two randomly chosen cells belong to the same cell type

$$\gamma = \sum_i p_i^2. \tag{11}$$

For the Simpson index, a higher value (counter-intuitively) indicates decreased diversity. This leads to the definition of the Gini-Simpson index which is simply the complement of the Simpson index

$$\text{Gini-Simpson Index} = 1 - \gamma \tag{12}$$

For the Gini-Simpson index, a higher value indicates increased diversity.

Vendi Score. The Vendi Score is a measure of the intrinsic diversity of a dataset, without reference to any metadata [23]. In this study, we calculate the Vendi Score of the gene expression matrix for each dataset. When using a linear kernel, the Vendi Score has computational complexity $O(\min(J^3, N^3))$ where N is the number of samples in the dataset and J is the number of features in the dataset. Due to this, we pre-processed each dataset by

1. Noramalizing cells to 10,000 counts per cell.
2. Logarithmizing the counts with a pseudocount of 1 (i.e., using the function $f(x) = \log(x + 1)$).
3. Restricting to highly variable genes using the `scanpy` function `sc.pp.highly_variable_genes` with default parameters.
4. Centering and scaling the variable genes that remained.

Suppose that we have a collection of cells $(\mathbf{x}_1, \ldots, \mathbf{x}_N)$ and define a positive semidefinite similarity function k that satisfies $k(\mathbf{x}, \mathbf{x}) = 1$. Next let $\mathbf{K}$ be an $N \times N$ kernel matrix where each entry is $K_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$, and let $\lambda_1, \ldots, \lambda_N$ be the eigenvalues of $\mathbf{K}/N$. The Vendi Score with respect to the function k is then the exponential of the Shannon entropy of the eigenvalues of $\mathbf{K}/N$

$$VS_k(\mathbf{x}_1, \ldots, \mathbf{x}_N) = \exp\left(-\sum_i \lambda_i \log \lambda_i\right). \tag{13}$$

The Vendi Score can also be viewed as the effective rank of a matrix and was previously introduced in the signal processing literature [39].

Code and data availability

All code is available under an open-source MIT license at https://github.com/microsoft/scFM-dataselection. Specifically, we used scVI [17] which was downloaded from https://scvi-tools.org/ [40] (release 1.1.2), the version of SSL [18] corresponding to commit a149b403fb2f7fe976cb8bd30fcb3db7a51e9500 (May 2, 2024) from the GitHub repository https://github.com/theislab/ssl_in_scg, and the version of Geneformer [7] corresponding to commit b2bbd7ccc856e3b3c1a9199c7cca07d888d87663 (Jun 29, 2024) from the Hugging Face repository https://huggingface.co/ctheodoris/Geneformer/tree/main/geneformer. The clonal hematopoiesis, placenta, intestine-on-chip, and periodontitis are all available in the CELLxGENE census and can be found online at https://cellxgene.cziscience.com/. The lung dataset from Kim et al. [28] is available in the Curated Cancer Cell Atlas (3CA) and can be found online at https://www.weizmann.ac.il/sites/3CA/lung. Instructions for downloading the scTab corpus from the CELLxGENE census are provided at https://github.com/microsoft/scFM-dataselection.

Acknowledgments

This research was conducted by using a combination of computational resources and services provided by Microsoft Research and the Center for Computation and Visualization at Brown University. Parts of Figures 1 and 3 were created using BioRender.com. This research was also supported in part by a David & Lucile Packard Fellowship for Science and Engineering awarded to LC. SR acknowledges funding support from NCI K08 CA260442. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of any of the funders.

Author contributions

AD, APA, and LC conceived the study. AD, MH, and AT developed the software and performed the analyses. AG and AWN conducted secondary analyses. NF, SR, PSW, APA, and LC provided resources. SR, PSW, APA, and LC supervised the project. AD, MH, AT, APA, and LC wrote the initial draft. All authors interpreted the results and revised the manuscript.

Declaration of interests

SR holds equity in Amgen. SR and PSW receive research funding from Microsoft. MH, NF, APA, and LC are employees of Microsoft and own equity in Microsoft. PSW receives research funding from Microsoft and reports compensation for consulting/speaking from Engine Ventures and AbbVie unrelated to this work. All other authors have declared that no competing interests exist.

References

1. Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ B. Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri S. Chatterji, Annie S. Chen, Kathleen Creel, Jared Quincy Davis, Dorottya 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 D. 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 S. Krass, Ranjay Krishna, Rohith Kuditipudi, and et al. On the opportunities and risks of foundation models. *CoRR*, abs/2108.07258, 2021. URL `https://arxiv.org/abs/2108.07258`.

2. John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A. A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, Aug 2021. ISSN 1476-4687. doi: 10.1038/s41586-021-03819-2. URL `https://doi.org/10.1038/s41586-021-03819-2`.

3. Minkyung Baek, Frank DiMaio, Ivan Anishchenko, Justas Dauparas, Sergey Ovchinnikov, Gyu Rie Lee, Jue Wang, Qian Cong, Lisa N. Kinch, R. Dustin Schaeffer, Claudia Millán, Hahnbeom Park, Carson Adams, Caleb R. Glassman, Andy DeGiovanni, Jose H. Pereira, Andria V. Rodrigues, Alberdina A. van Dijk, Ana C. Ebrecht, Diederik J. Opperman, Theo Sagmeister, Christoph Buhlheller, Tea Pavkov-Keller, Manoj K. Rathinaswamy, Udit Dalwadi, Calvin K. Yip, John E. Burke, K. Christopher Garcia, Nick V. Grishin, Paul D. Adams, Randy J. Read, and David Baker. Accurate prediction of protein structures and interactions using a three-track neural network. *Science*, 373(6557):871–876, 2021. doi: 10.1126/science.abj8754. URL `https://www.science.org/doi/abs/10.1126/science.abj8754`.

4. Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023.

5. Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D 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 Ziegler, Jeffrey Wu, Clemens Winter, Chris 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. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020. URL `https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf`.

6. 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, Oct 2022. ISSN 2522-5839. doi: 10.1038/s42256-022-00534-z. URL `https://doi.org/10.1038/s42256-022-00534-z`.

7. 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, Jun 2023. ISSN 1476-4687. doi: 10.1038/s41586-023-06139-9. URL https://doi.org/10.1038/s41586-023-06139-9.

8. 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*, Feb 2024. ISSN 1548-7105. doi: 10.1038/s41592-024-02201-0. URL https://doi.org/10.1038/s41592-024-02201-0.

9. 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.

10. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *CoRR*, abs/1706.03762, 2017. URL http://arxiv.org/abs/1706.03762.

11. Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

12. Kasia Z. Kedzierska, Lorin Crawford, Ava P. Amini, and Alex X. Lu. Assessing the limits of zero-shot foundation models in single-cell biology. *bioRxiv*, 2023. doi: 10.1101/2023.10.16.561085. URL https://www.biorxiv.org/content/early/2023/11/05/2023.10.16.561085.

13. Rebecca Boiarsky, Nalini M. Singh, Alejandro Buendia, Ava P. Amini, Gad Getz, and David Sontag. Deeper evaluation of a single-cell foundation model. *Nature Machine Intelligence*, Dec 2024. ISSN 2522-5839. doi: 10.1038/s42256-024-00949-w. URL https://doi.org/10.1038/s42256-024-00949-w.

14. Sachin Goyal, Pratyush Maini, Zachary C. Lipton, Aditi Raghunathan, and J. Zico Kolter. Scaling laws for data filtering – data curation cannot be compute agnostic, 2024.

15. Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. Training compute-optimal large language models, 2022.

16. Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models, 2023.

17. 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, Dec 2018. ISSN 1548-7105. doi: 10.1038/s41592-018-0229-2. URL https://doi.org/10.1038/s41592-018-0229-2.

18. Till Richter, Mojtaba Bahrami, Yufan Xia, David S. Fischer, and Fabian J. Theis. Delineating the effective use of self-supervised learning in single-cell genomics. *bioRxiv*, 2024. doi: 10.1101/2024.02.16.580624. URL https://www.biorxiv.org/content/early/2024/02/18/2024.02.16.580624.

19. Felix Fischer, David S. Fischer, Roman Mukhin, Andrey Isaev, Evan Biederstedt, Alexandra-Chloé Villani, and Fabian J. Theis. sctab: Scaling cross-tissue single-cell annotation models. *Nature Communications*, 15(1):6611, Aug 2024. ISSN 2041-1723. doi: 10.1038/s41467-024-51059-5. URL https://doi.org/10.1038/s41467-024-51059-5.

20. Brian Hie, Hyunghoon Cho, Benjamin DeMeo, Bryan Bryson, and Bonnie Berger. Geometric sketching compactly summarizes the single-cell transcriptomic landscape. *Cell Systems*, 8(6):483–493.e7, 2019. ISSN 2405-4712. doi: https://doi.org/10.1016/j.cels.2019.05.003. URL https://www.sciencedirect.com/science/article/pii/S2405471219301528.

21. C. E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27 (3):379–423, 1948. doi: https://doi.org/10.1002/j.1538-7305.1948.tb01338.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/j.1538-7305.1948.tb01338.x.

22. Michael Roswell, Jonathan Dushoff, and Rachael Winfree. A conceptual guide to measuring species diversity. *Oikos*, 130(3):321–338, 2021.

23. Dan Friedman and Adji Bousso Dieng. The vendi score: A diversity evaluation metric for machine learning. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=g97OHbQyk1.

24. J. Brett Heimlich, Pawan Bhat, Alyssa C. Parker, Matthew T. Jenkins, Caitlyn Vlasschaert, Jessica Ulloa, Joseph C. Van Amburg, Chad R. Potts, Sydney Olson, Alexander J. Silver, Ayesha Ahmad, Brian Sharber, Donovan Brown, Ningning Hu, Peter van Galen, Michael R. Savona, Alexander G. Bick, and P. Brent Ferrell. Multiomic profiling of human clonal hematopoiesis reveals genotype and cell-specific inflammatory pathway activation. *Blood Advances*, 8(14):3665–3678, 07 2024. ISSN 2473-9529. doi: 10.1182/bloodadvances.2023011445. URL https://doi.org/10.1182/bloodadvances.2023011445.

25. Renée Moerkens, Joram Mooiweer, Aarón D. Ramírez-Sánchez, Roy Oelen, Lude Franke, Cisca Wijmenga, Robert J. Barrett, Iris H. Jonkers, and Sebo Withoff. An ipsc-derived small intestine-on-chip with self-organizing epithelial, mesenchymal, and neural cells. *Cell Reports*, 43(7), Jul 2024. ISSN 2211-1247. doi: 10.1016/j.celrep.2024.114247. URL https://doi.org/10.1016/j.celrep.2024.114247.

26. Regina Hoo, Elias R. Ruiz-Morales, Iva Kelava, Mukul Rawat, Cecilia Icoresi Mazzeo, Elizabeth Tuck, Carmen Sancho-Serra, Sara Chelaghma, Alexander V. Predeus, Simon Murray, David Fernandez-Antoran, Ross F. Waller, Damiana Álvarez-Errico, Marcus C.S. Lee, and Roser Vento-Tormo. Acute response to pathogens in the early human placenta at single-cell resolution. *Cell Systems*, 15(5):425–444.e9, May 2024. ISSN 2405-4712. doi: 10.1016/j.cels.2024.04.002. URL https://doi.org/10.1016/j.cels.2024.04.002.

27. Quinn T. Easter, Bruno Fernandes Matuck, Germán Beldorati Stark, Catherine L. Worth, Alexander V. Predeus, Brayon Fremin, Khoa Huynh, Vaishnavi Ranganathan, Zhi Ren, Diana Pereira, Brittany T. Rupp, Theresa Weaver, Kathryn Miller, Paola Perez, Akira Hasuike, Zhaoxu Chen,

Mandy Bush, Xufeng Qu, Janice Lee, Scott H. Randell, Shannon M. Wallet, Inês Sequeira, Hyun Koo, Katarzyna M. Tyc, Jinze Liu, Kang I. Ko, Sarah A. Teichmann, and Kevin M. Byrd. Single-cell and spatially resolved interactomics of tooth-associated keratinocytes in periodontitis. *Nature Communications*, 15(1):5016, Jun 2024. ISSN 2041-1723. doi: 10.1038/s41467-024-49037-y. URL https://doi.org/10.1038/s41467-024-49037-y.

28. Nayoung Kim, Hong Kwan Kim, Kyungjong Lee, Yourae Hong, Jong Ho Cho, Jung Won Choi, Jung-Il Lee, Yeon-Lim Suh, Bo Mi Ku, Hye Hyeon Eum, et al. Single-cell rna sequencing demonstrates the molecular and cellular reprogramming of metastatic lung adenocarcinoma. *Nature communications*, 11(1):2285, 2020.

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, Jan 2022. ISSN 1548-7105. doi: 10.1038/s41592-021-01336-8. URL https://doi.org/10.1038/s41592-021-01336-8.

30. Joseph M. Replogle, Reuben A. Saunders, Angela N. Pogson, Jeffrey A. Hussmann, Alexander Lenail, Alina Guna, Lauren Mascibroda, Eric J. Wagner, Karen Adelman, Gila Lithwick-Yanai, Nika Iremadze, Florian Oberstrass, Doron Lipson, Jessica L. Bonnar, Marco Jost, Thomas M. Norman, and Jonathan S. Weissman. Mapping information-rich genotype-phenotype landscapes with genome-scale perturb-seq. *Cell*, 185(14):2559–2575.e28, 2022. ISSN 0092-8674. doi: https://doi.org/10.1016/j.cell.2022.05.013. URL https://www.sciencedirect.com/science/article/pii/S0092867422005979.

31. Tomàs Montserrat-Ayuso and Anna Esteve-Codina. High content of nuclei-free low-quality cells in reference single-cell atlases: a call for more stringent quality control using nuclear fraction. *BMC genomics*, 25:1124, 2024.

32. Stephen R Quake. The cell as a bag of rna. *Trends in Genetics*, 37(12):1064–1068, 2021.

33. F William Townes, Stephanie C Hicks, Martin J Aryee, and Rafael A Irizarry. Feature selection and dimension reduction for single-cell rna-seq based on a multinomial model. *Genome biology*, 20:1–16, 2019.

34. Gerard A Bouland, Ahmed Mahfouz, and Marcel JT Reinders. Consequences and opportunities arising due to sparser single-cell rna-seq datasets. *Genome biology*, 24(1):86, 2023.

35. 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. URL https://www.biorxiv.org/content/early/2023/11/02/2023.10.30.563174.

36. Stefan Peidli, Tessa D Green, Ciyue Shen, Torsten Gross, Joseph Min, Samuele Garda, Bo Yuan, Linus J Schumacher, Jake P Taylor-King, Debora S Marks, et al. scperturb: harmonized single-cell perturbation data. *Nature Methods*, 21(3):531–540, 2024.

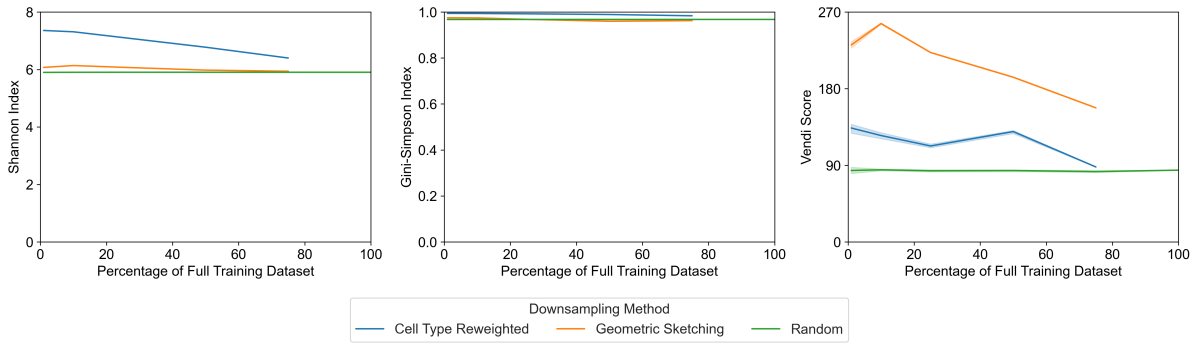
37. William M. Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66(336):846–850, 1971. ISSN 01621459. URL http://www.jstor.org/stable/2284239.

38. E. H. SIMPSON. Measurement of diversity. *Nature*, 163(4148):688–688, Apr 1949. ISSN 1476-4687. doi: 10.1038/163688a0. URL https://doi.org/10.1038/163688a0.

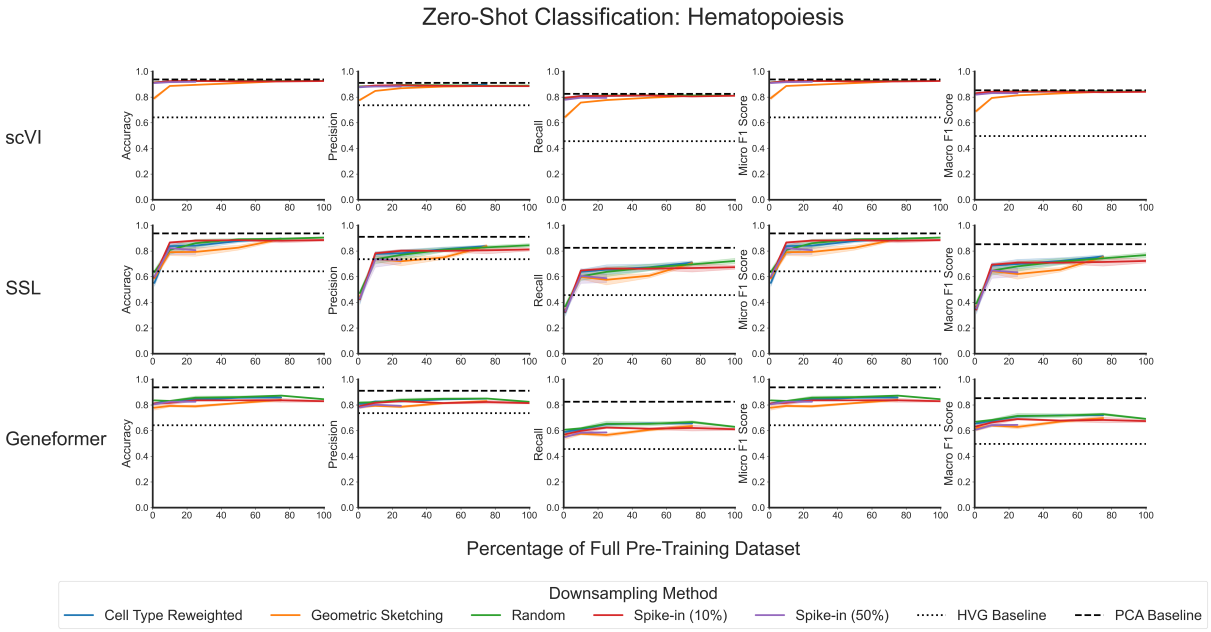
39. Olivier Roy and Martin Vetterli. The effective rank: A measure of effective dimensionality. In *2007 15th European Signal Processing Conference*, pages 606–610, 2007.

40. Adam Gayoso, Romain Lopez, Galen Xing, Pierre Boyeau, Valeh Valiollah Pour Amiri, Justin Hong, Katherine Wu, Michael Jayasuriya, Edouard Mehlman, Maxime Langevin, Yining Liu, Jules Samaran, Gabriel Misrachi, Achille Nazaret, Oscar Clivio, Chenling Xu, Tal Ashuach, Mariano Gabitto, Mohammad Lotfollahi, Valentine Svensson, Eduardo da Veiga Beltrame, Vitalii Kleshchevnikov, Carlos Talavera-López, Lior Pachter, Fabian J. Theis, Aaron Streets, Michael I. Jordan, Jeffrey Regier, and Nir Yosef. A python library for probabilistic analysis of single-cell omics data. *Nature Biotechnology*, 40(2):163–166, Feb 2022. ISSN 1546-1696. doi: 10.1038/s41587-021-01206-w. URL https://doi.org/10.1038/s41587-021-01206-w.

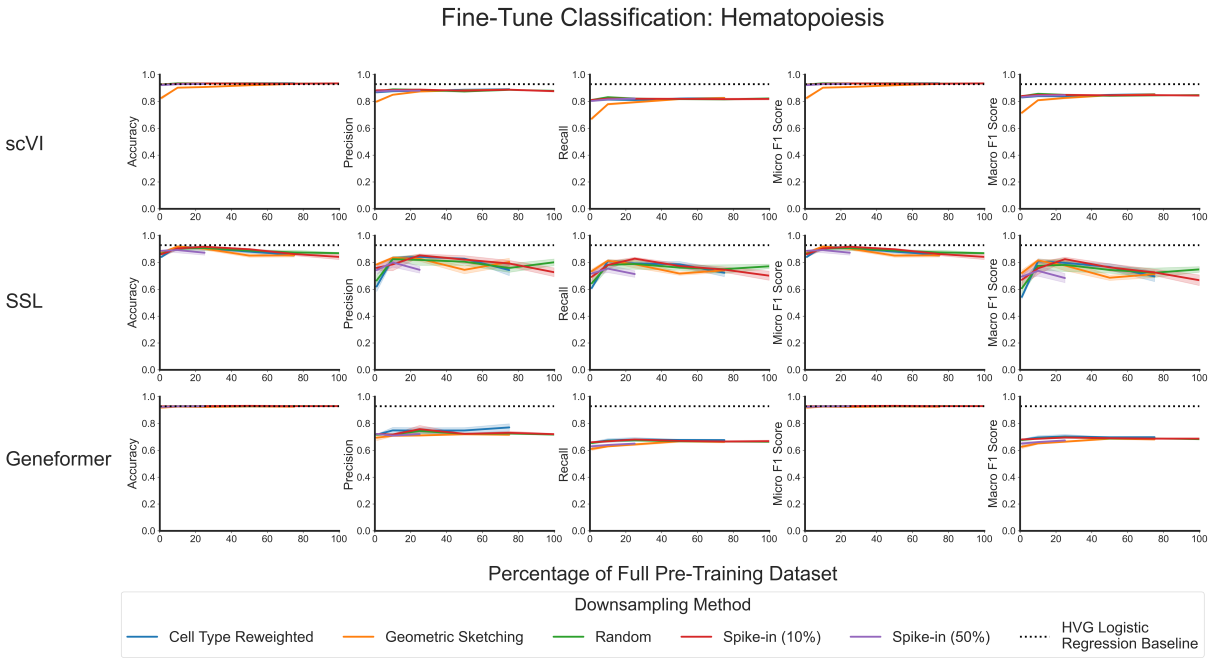
Supplementary Figures



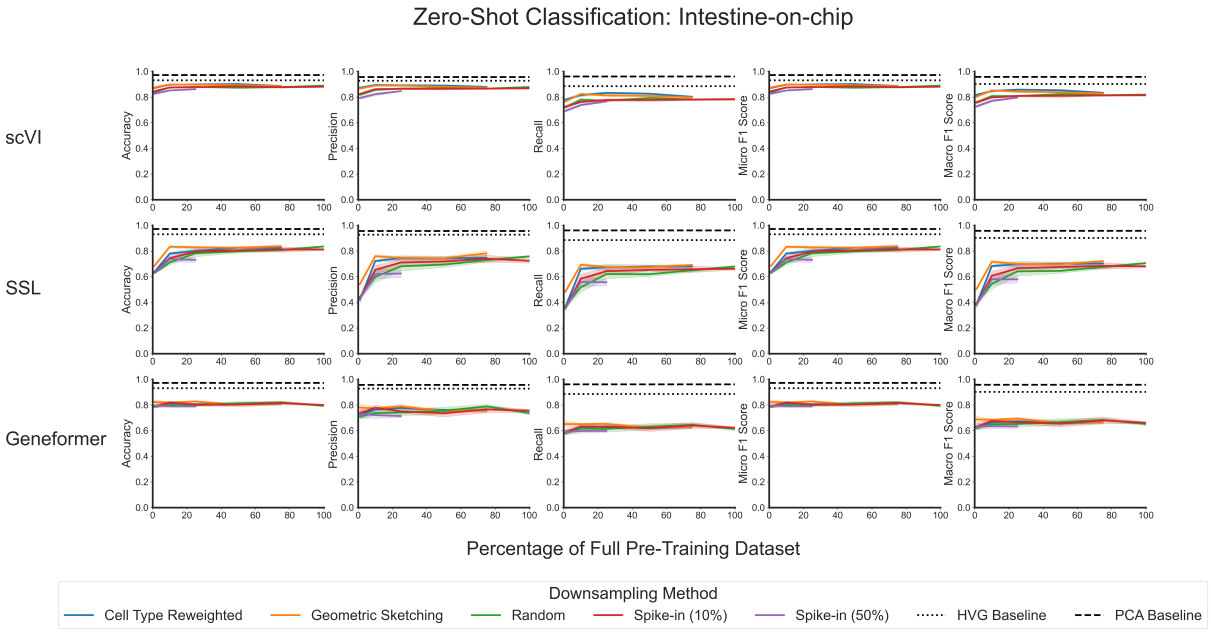
Supplemental Figure 1. Diversity of datasets used for pre-training as evaluated by intrinsic and extrinsic metrics. The Shannon index, Gini-Simpson index, and Vendi Score are shown for each of the downsampled pre-training datasets. Cell type re-weighting and geometric sketching have increased diversity relative to the randomly downsampled datasets. Cell type re-weighting (which re-weights based on cell type metadata) has the highest Shannon index and Gini-Simpson index (which both measure the diversity of cell type metadata). Geometric sketching (which samples evenly across transcriptional space) has the highest Vendi Score (which measures the diversity of the transcriptional data directly).



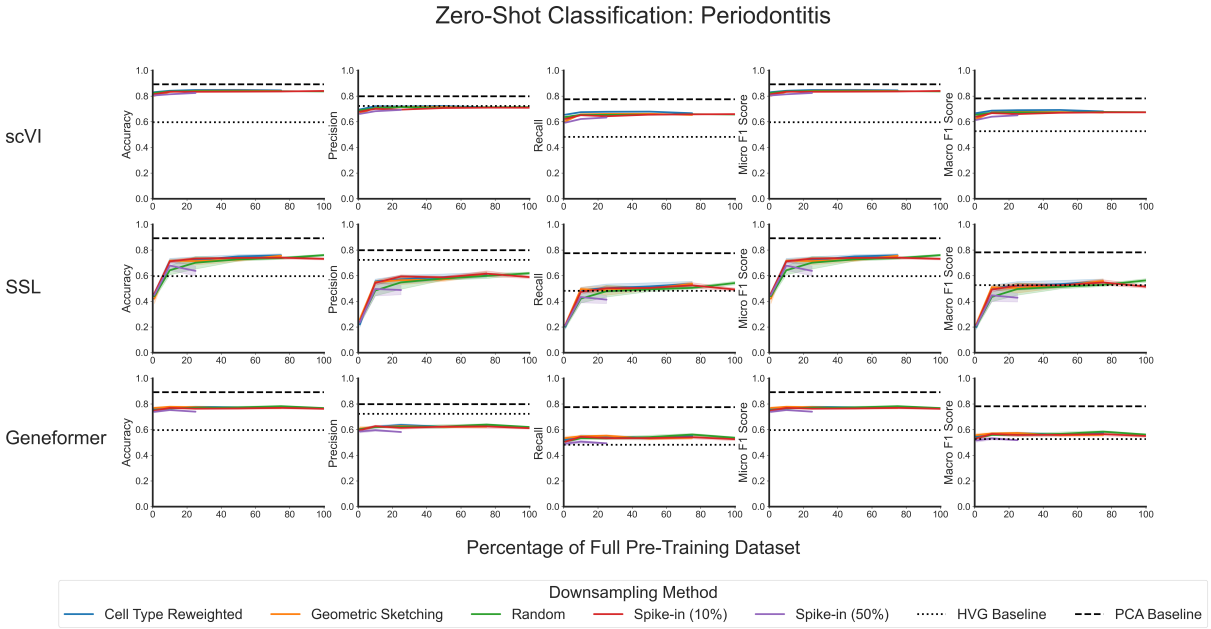
Supplemental Figure 2. Multiple metrics show that zero-shot model performance on classifying cells from a clonal hematopoiesis dataset plateau at a small fraction of the total data available for pre-training. Line plots showing zero-shot classification performance for each model's embeddings as evaluated by accuracy, precision, recall, micro F1 Score, and macro F1 score. For each model, the different colors correspond to the downsampling strategy used to generate the data used for pre-training. The dotted line shows the performance of simply using the highly variable genes as an embedding and the dashed line shows the performance of using principal component projections as an embedding.



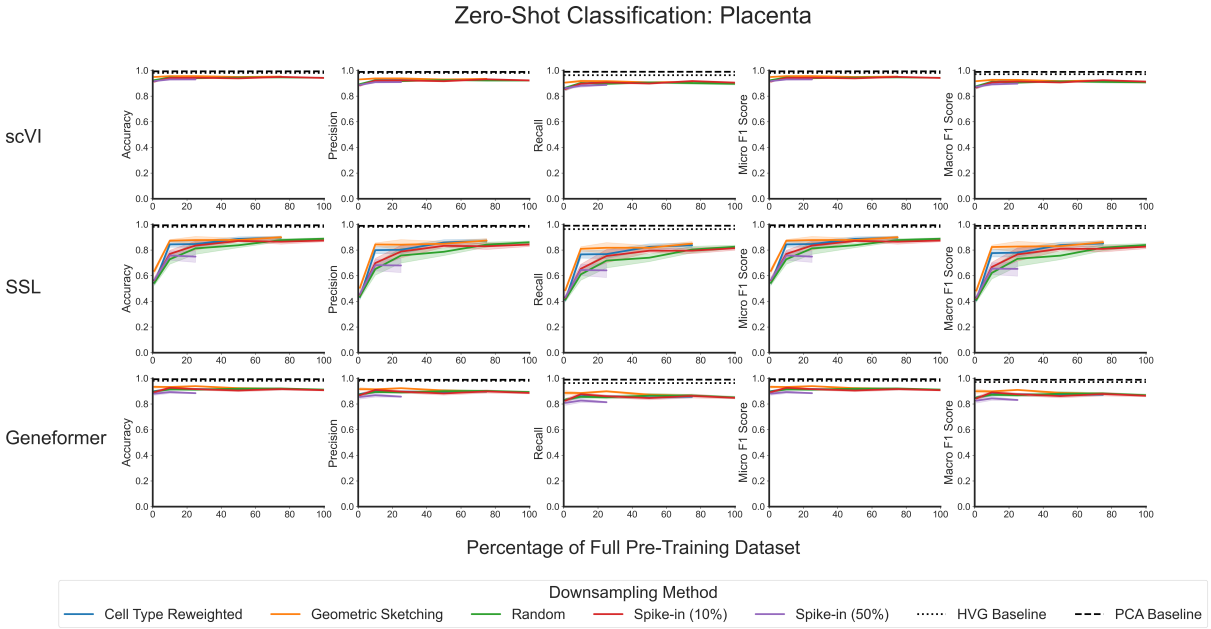
Supplemental Figure 3. Multiple metrics show that fine-tuned model performance on classifying cells from a clonal hematopoiesis dataset plateau at a small fraction of the total data available for pre-training. Line plots showing zero-shot classification performance for each model's embeddings as evaluated by accuracy, precision, recall, micro F1 Score, and macro F1 score. For each model, the different colors correspond to the downsampling strategy used to generate the data used for pre-training. The dotted line shows the performance of a regularized logistic classifier using the highly variable genes as input.



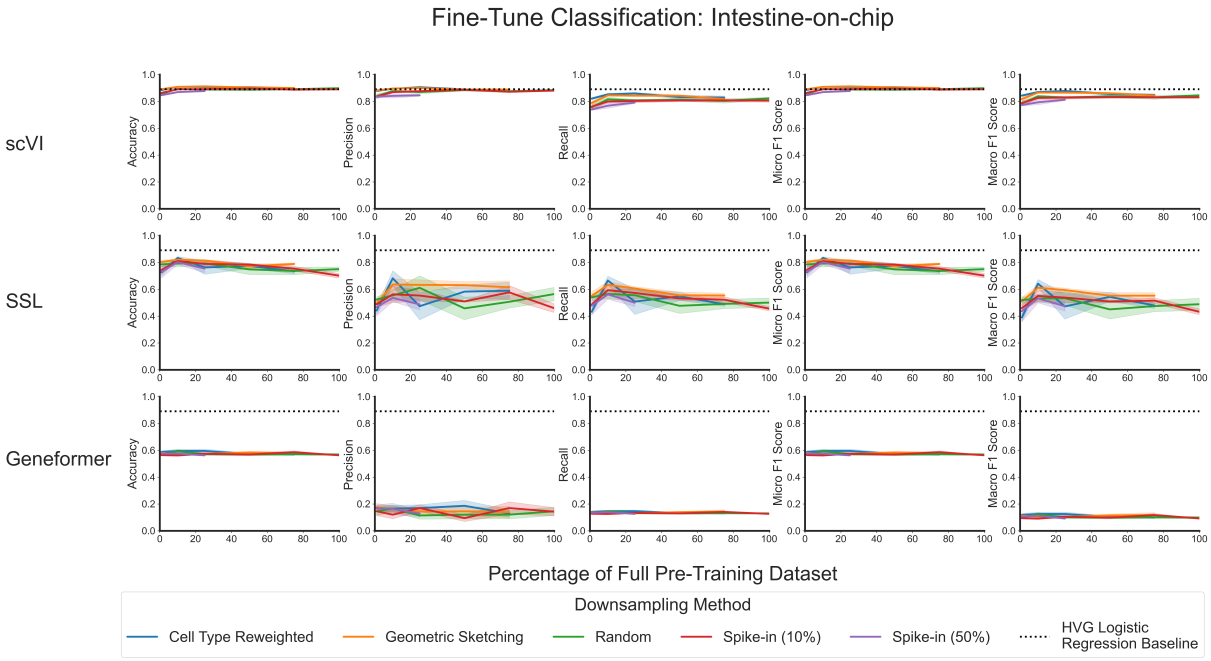
Supplemental Figure 4. Multiple metrics show that zero-shot model performance on classifying cells from an intestine-on-chip dataset plateau at a small fraction of the total data available for pre-training. Line plots showing zero-shot classification performance for each model's embeddings as evaluated by accuracy, precision, recall, micro F1 Score, and macro F1 score. For each model, the different colors correspond to the downsampling strategy used to generate the data used for pre-training. The dotted line shows the performance of simply using the highly variable genes as an embedding and the dashed line shows the performance of using principal component projections as an embedding.



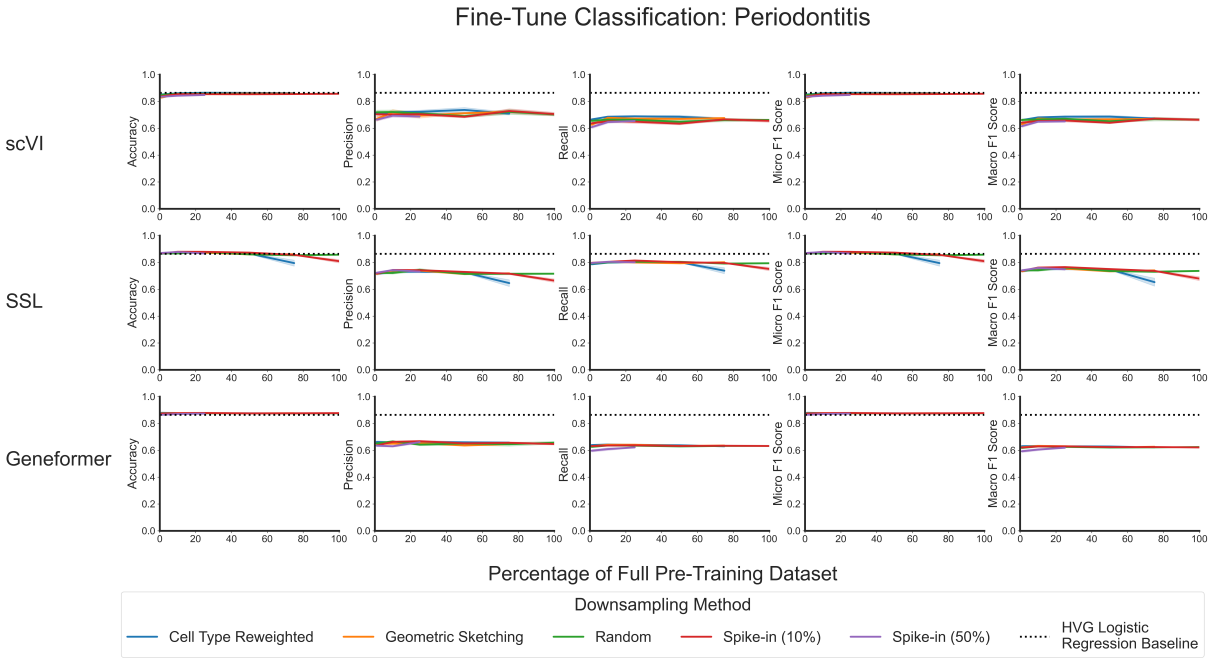
Supplemental Figure 5. Multiple metrics show that zero-shot model performance on classifying cells from a periodontitis dataset plateau at a small fraction of the total data available for pre-training. Line plots showing zero-shot classification performance for each model's embeddings as evaluated by accuracy, precision, recall, micro F1 Score, and macro F1 score. For each model, the different colors correspond to the downsampling strategy used to generate the data used for pre-training. The dotted line shows the performance of simply using the highly variable genes as an embedding and the dashed line shows the performance of using principal component projections as an embedding.



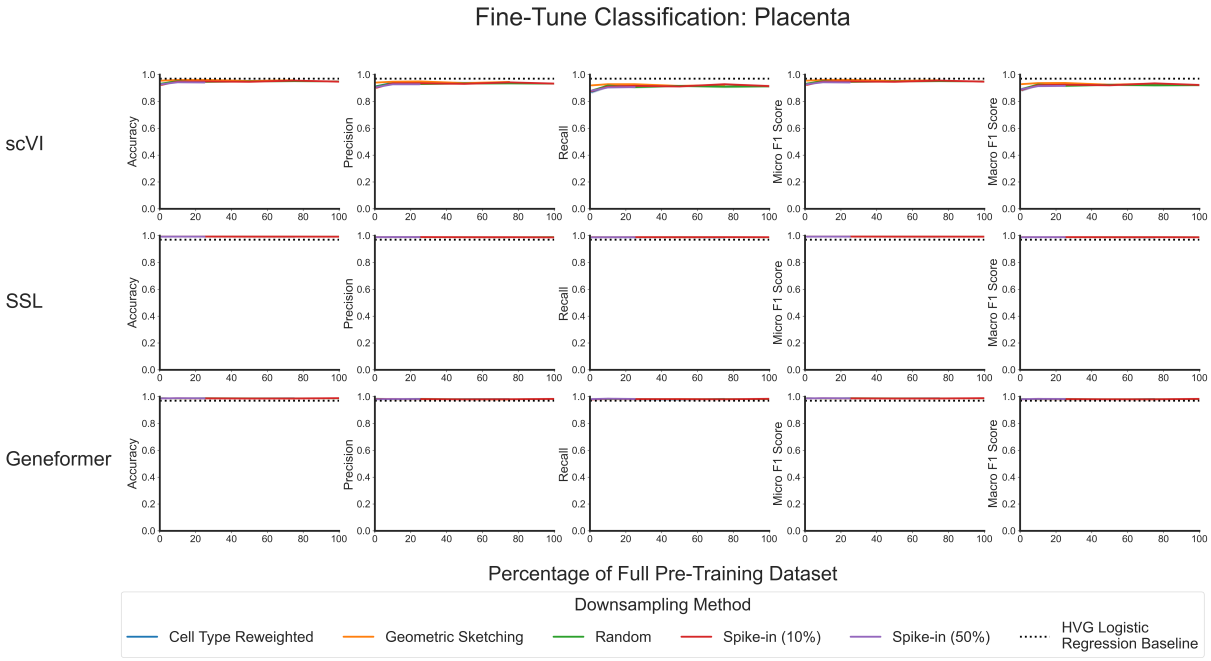
Supplemental Figure 6. Multiple metrics show that zero-shot model performance on classifying cells from a placenta dataset plateau at a small fraction of the total data available for pre-training. Line plots showing zero-shot classification performance for each model's embeddings as evaluated by accuracy, precision, recall, micro F1 Score, and macro F1 score. For each model, the different colors correspond to the downsampling strategy used to generate the data used for pre-training. The dotted line shows the performance of simply using the highly variable genes as an embedding and the dashed line shows the performance of using principal component projections as an embedding.



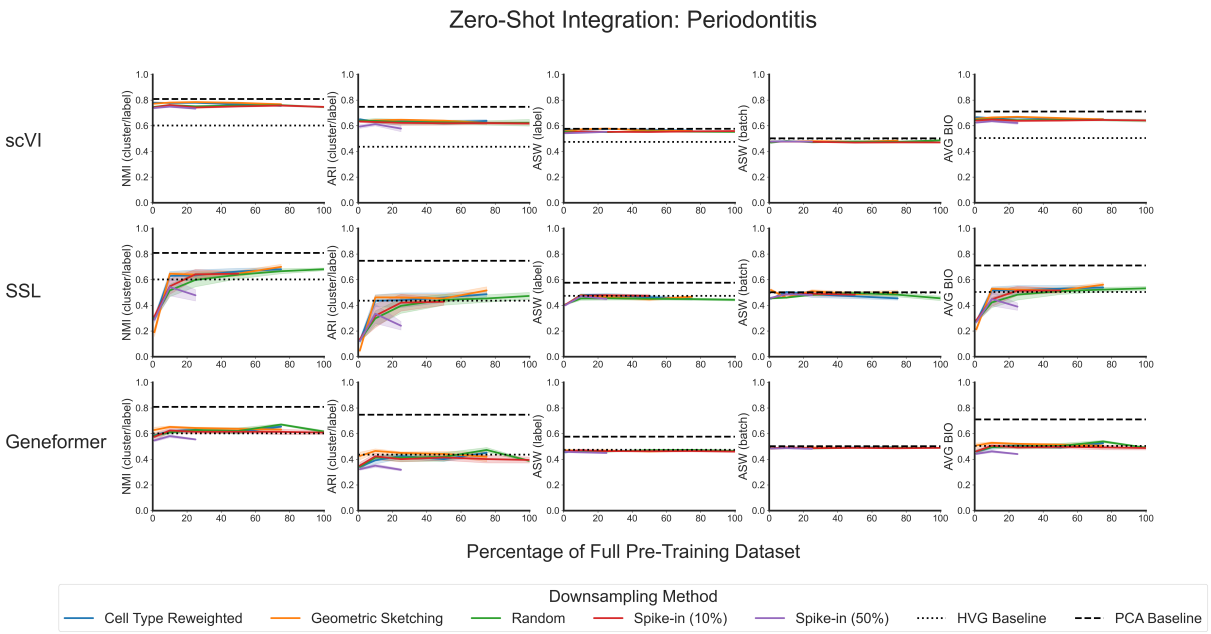
Supplemental Figure 7. Multiple metrics show that fine-tuned model performance on classifying cells from an intestine-on-chip dataset plateau at a small fraction of the total data available for pre-training. Line plots showing zero-shot classification performance for each model's embeddings as evaluated by accuracy, precision, recall, micro F1 Score, and macro F1 score. For each model, the different colors correspond to the downsampling strategy used to generate the data used for pre-training. The dotted line shows the performance of a regularized logistic classifier using the highly variable genes as input.



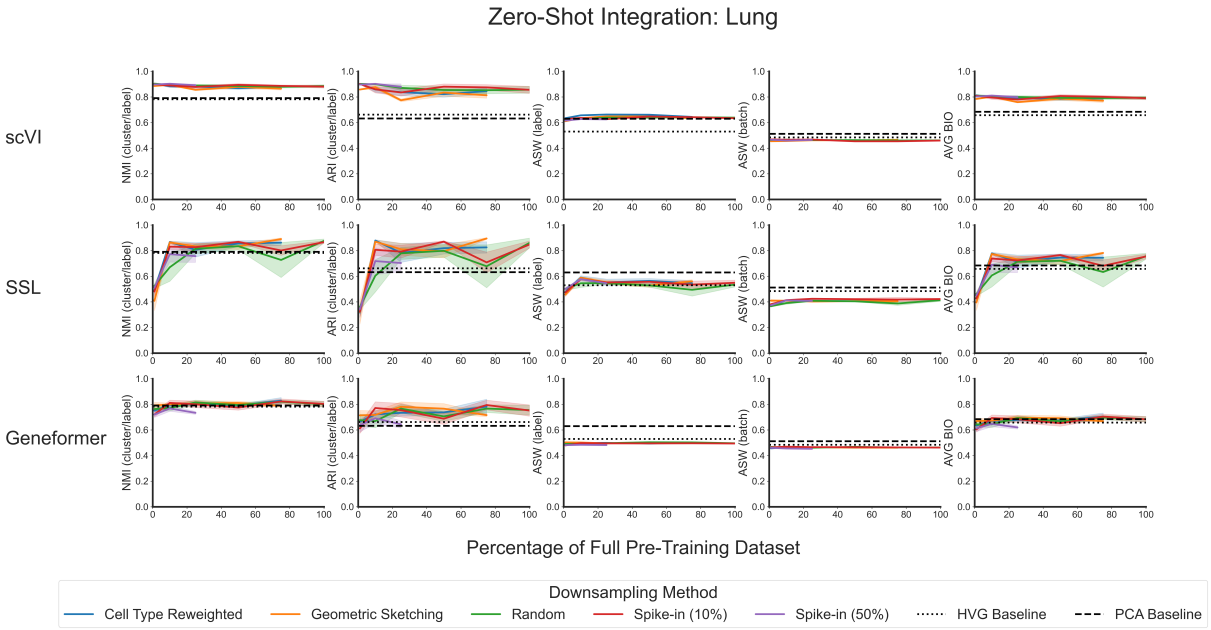
Supplemental Figure 8. Multiple metrics show that fine-tuned model performance on classifying cells from a periodontitis dataset plateau at a small fraction of the total data available for pre-training. Line plots showing zero-shot classification performance for each model's embeddings as evaluated by accuracy, precision, recall, micro F1 Score, and macro F1 score. For each model, the different colors correspond to the downsampling strategy used to generate the data used for pre-training. The dotted line shows the performance of a regularized logistic classifier using the highly variable genes as input.



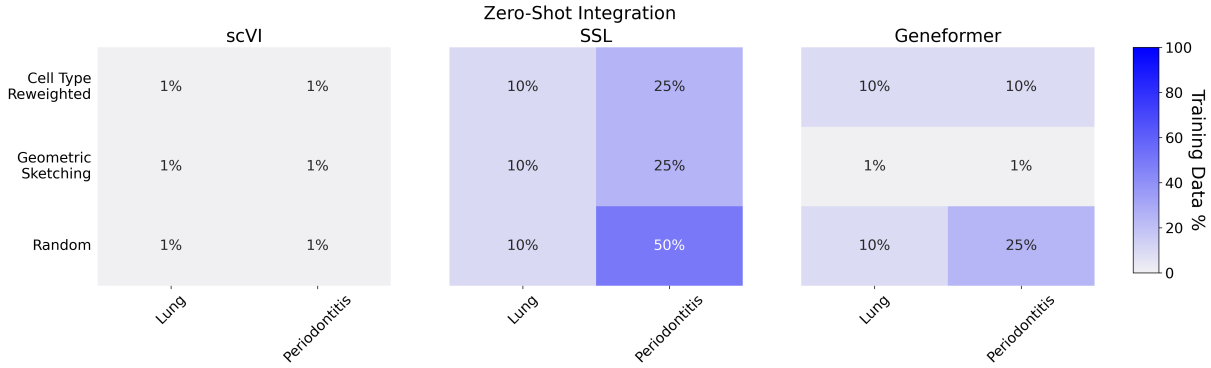
Supplemental Figure 9. Multiple metrics show that fine-tuned model performance on classifying cells from a placenta dataset plateau at a small fraction of the total data available for pre-training. Line plots showing zero-shot classification performance for each model's embeddings as evaluated by accuracy, precision, recall, micro F1 Score, and macro F1 score. For each model, the different colors correspond to the downsampling strategy used to generate the data used for pre-training. The dotted line shows the performance of a regularized logistic classifier using the highly variable genes as input.



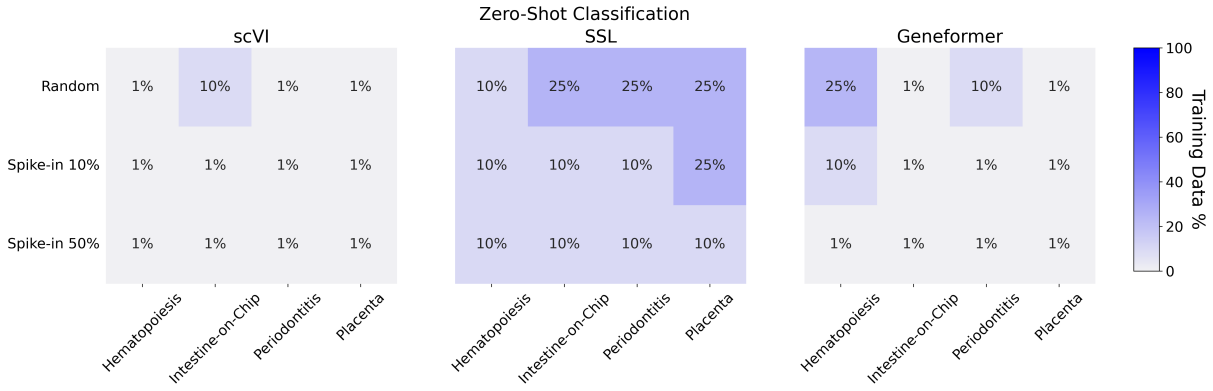
Supplemental Figure 10. Multiple metrics show that zero-shot model performance on integrating cells from a periodontitis dataset plateau at a small fraction of the total data available for pre-training. Line plots showing zero-shot integration performance for each model's embeddings as evaluated by normalized mutual information (NMI), adjusted Rand index (ARI), batch specific average silhouette width (ASW), cell type label ASW, and average BIO (AvgBIO). For each model, the different colors correspond to the downsampling strategy used to generate the data used for pre-training. The dotted line shows the performance of simply using the highly variable genes as an embedding and the dashed line shows the performance of using principal component projections as an embedding.



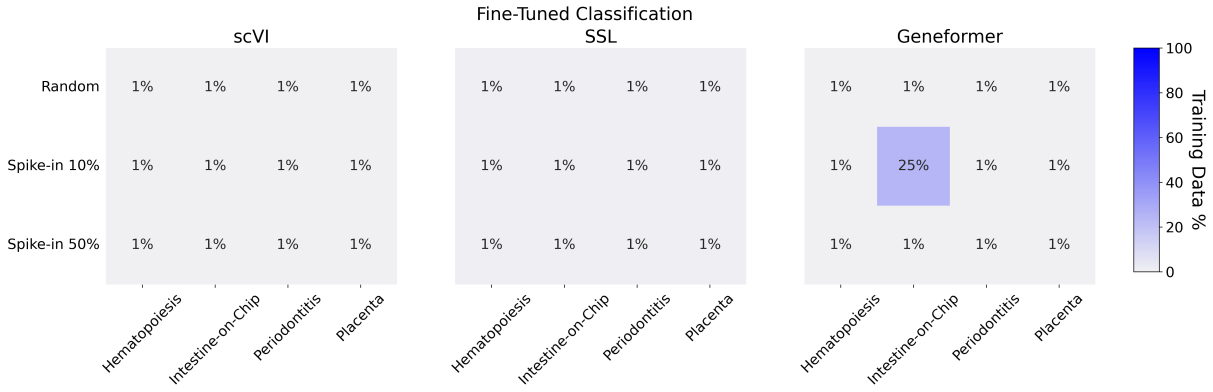
Supplemental Figure 11. Multiple metrics show that zero-shot model performance on integrating cells from a lung dataset plateau at a small fraction of the total data available for pre-training. Line plots showing zero-shot integration performance for each model's embeddings as evaluated by normalized mutual information (NMI), adjusted Rand index (ARI), batch-specific average silhouette width (ASW), cell type label ASW, and average BIO (AvgBIO). For each model, the different colors correspond to the downsampling strategy used to generate the data used for pre-training. The dotted line shows the performance of simply using the highly variable genes as an embedding and the dashed line shows the performance of using principal component projections as an embedding.



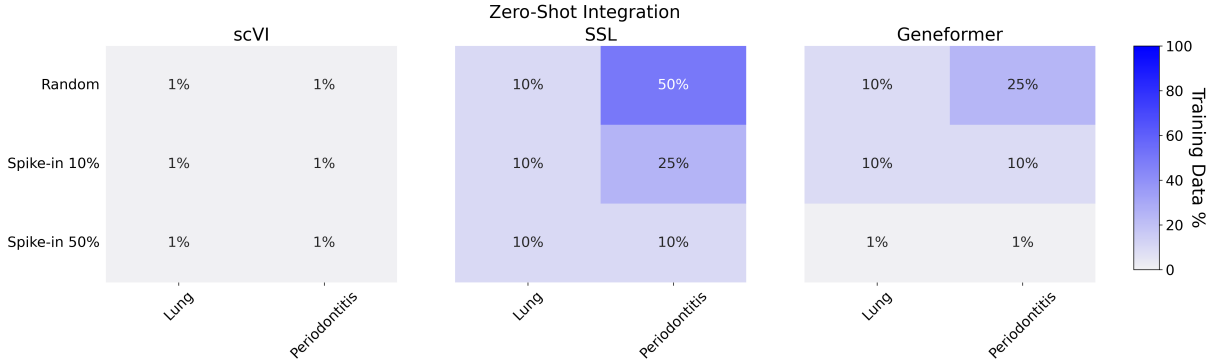
Supplemental Figure 12. Assessment of zero-shot integration performance saturation as a function of pre-training dataset size across multiple datasets. A heatmap showing the "learning saturation point" for the periodontitis dataset and the lung dataset for scVI, SSL, and Geneformer across each downsampling strategy when evaluated on zero-shot batch integration case. Each sub-panel corresponds to the model architecture, the x-axis corresponds to the dataset evaluated, and the y-axis corresponds to the downsampling strategy used to pre-train each model.



Supplemental Figure 13. Assessment of zero-shot classification performance saturation as a function of pre-training dataset size across multiple datasets. A heatmap showing the "learning saturation point" for the clonal hematopoiesis, intestine, periodontitis, and placenta datasets for scVI, SSL, and Geneformer across each downsampling strategy when evaluated on classification in the zero-shot case. Each sub-panel corresponds to the model architecture, the x-axis corresponds to the dataset evaluated, and the y-axis corresponds to the spike-in strategy used to pre-train each model. Thresholds for each model on uniform downsampled data (labeled as "random") are given as reference.



Supplemental Figure 14. Assessment of fine-tuned classification performance saturation as a function of pre-training dataset size across multiple datasets. A heatmap showing the "learning saturation point" for the clonal hematopoiesis, intestine, periodontitis, and placenta datasets for scVI, SSL, and Geneformer across each downsampling strategy when evaluated on classification in the fine-tuned case. Each sub-panel corresponds to the model architecture, the x-axis corresponds to the dataset evaluated, and the y-axis corresponds to the spike-in strategy used to pre-train each model. Thresholds for each model on uniform downsampled data (labeled as "random") are given as reference.



Supplemental Figure 15. Assessment of zero-shot integration performance saturation as a function of pre-training dataset size across multiple datasets. A heatmap showing the "learning saturation point" for the periodontitis dataset and the lung dataset for scVI, SSL, and Geneformer across each downsampling strategy when evaluated on zero-shot batch integration case. Each sub-panel corresponds to the model architecture, the x-axis corresponds to the dataset evaluated, and the y-axis corresponds to the spike-in strategy used to pre-train each model. Thresholds for each model on uniform downsampled data (labeled as "random") are given as reference.