Scaling up measurement noise scaling laws

Igor Sadalski [1] **Dan Raviv** [1] **Jonathan Rosenfeld** [1] **Allon Klein** [2] **Gokul Gowri** [2]

Abstract

Learning meaningful representations of cellular states is a key problem in computational biology. Yet, the scaling behavior of single-cell representation learning models remains poorly understood. While recent work has proposed that model performance scales predictably with measurement noise, this hypothesis has only been validated with relatively small models and datasets. In this work-in-progress, we present the first empirical evidence supporting measurement noise scaling laws at large scales using datasets on the order of 10^7 cells and transformer-based models with $> 10^7$ parameters. We demonstrate that previously observed noise-scaling behavior again consistently emerge in these large-scale models and datasets. Our results provide further evidence that measurement noise is an important scaling axis for cellular representation learning.

1. Introduction

Across image and text domains it has been observed that scaling up deep learning model size and training dataset size often leads to predictable improvements in performance (Rosenfeld et al., 2019; Kaplan et al., 2020). Inspired by this, several efforts have been made to train large generative models on increasingly large collections of single-cell transcriptomes (e.g. Geneformer, scFoundation, scGPT), aiming to learn "universal" representations of cellular states (Theodoris et al., 2023; Hao et al., 2024; Cui et al., 2024).

However, the information contained in a single-cell RNA sequencing (scRNA-seq) dataset depends not only on the number of cells it contains, but also the accuracy of the measurements. Quantifying transcript abundance in a cell is fundamentally challenging due to low copy number, and molecular undersampling results in significant measurement noise. Recent work (Gowri et al., 2025) suggests that representation quality improves logarithmically with per-cell transcript counts. In particular, for an information theoretic model quality metric $\mathcal{I}$ and molecular counts per cell u,

$$\mathcal{I}(u) = \mathcal{I}_{\max} - \frac{1}{2}\log\frac{1 + u/\bar{u}}{u/\bar{u} + 2^{-2\mathcal{I}_{\max}}} \tag{1}$$

However, this noise-scaling behavior has been verified empirically only in datasets up to 10^5 cells and with relatively small models. Whether these noise-scaling laws persist in far larger models and for datasets with tens of millions of cells remains unexplored.

A major barrier is that validating measurement noise scaling laws requires an independent "external signal" (protein measurements, spatial context, etc.) against which to quantify representation quality. Yet the large majority of scRNA-seq data lack such auxiliary data. To address this, we propose developmental time as an external signal. This allows us to build upon a 10^7-cell mouse developmental atlas consisting of several embryos from gastrulation to birth (Qiu et al., 2024), where we can evaluate a cellular representation by its ability to capture information about developmental stage.

In this work, we provide the first empirical evidence that measurement noise scaling laws extend to large-scale foundation models. We show that even with 10^7 cells and $> 10^7$ parameter transformer models, the mutual information between learned representations and developmental time continues to follow the same logarithmic scaling law with transcript counts. This shows that measurement noise remains a fundamental axis for improving single-cell generative models.

2. Experimental setup

We closely follow the experimental setup introduced in Gowri et al. (2025), while introducing a new external signal and studying a large Geneformer model. Below, we briefly review the key details and highlight the contributions of this work.

2.1. Review: Information-theoretic probing of representation quality

As in Gowri et al. (2025), we evaluate the quality of a representation by measuring the mutual information between the

[1]Somite AI, Boston, MA [2]Department of Systems Biology, Harvard University. Correspondence to: Gokul Gowri <ggowri@g.harvard.edu>.

Preliminary work.

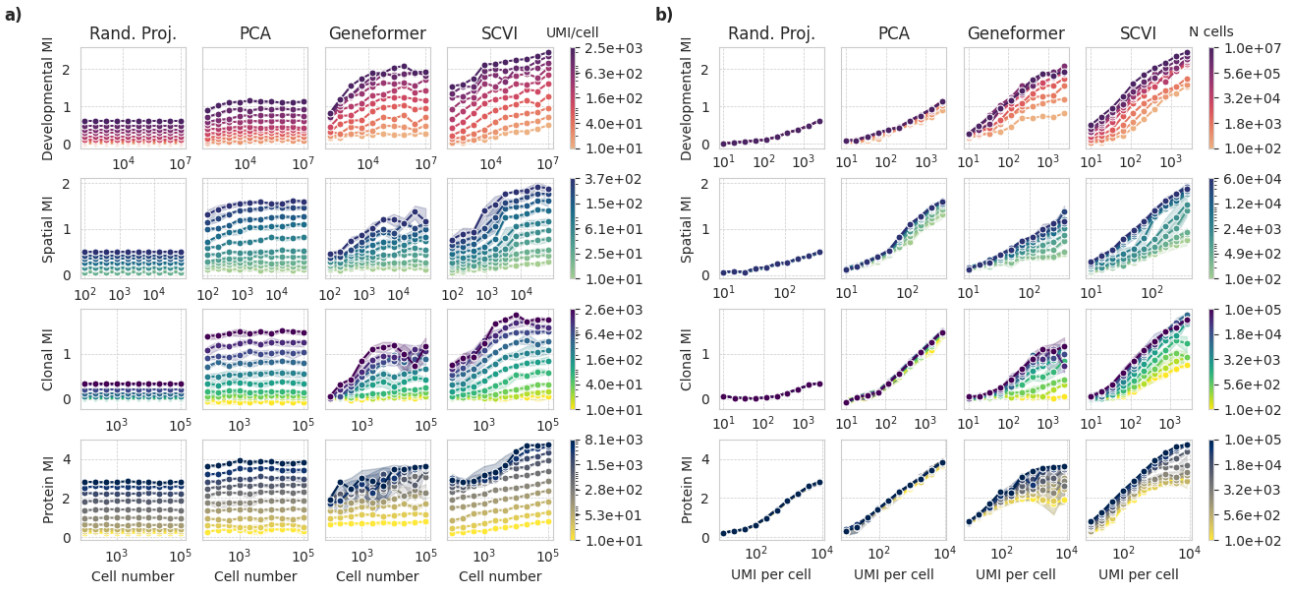

Figure 1. Cell subsampling and count downsampling experiments. a) Representation information visualized as a function of cell number in training dataset. *b)* Representation information visualized as a function of mean UMI per cell in dataset. Confidence bands show min and max over three replicates (varying random seed) for spatial, clonal, and protein signals. Replicates for developmental signal are omitted due to resource constraints.

representation and an external signal. As opposed to model-specific loss functions, this allows comparisons across models and datasets. In order to estimate mutual information in high dimensions, we use an approximation approach introduced in Gowri et al. (2024).

To understand scaling behavior of cellular representation learning models, we measure representation quality after training on datasets which are artificially subsampled to different numbers of cells, and artificially downsampled to different numbers of transcripts captured, or unique molecular identifiers (UMI), per cell.

2.2. Settings introduced in this work

The key contributions of this work are (1) the use of developmental time as an external signal and (2) the exploration of a large-scale transformer based model.

Scaling analysis with a developmental atlas The use of developmental time as an external signal enables the observation of noise scaling behavior in a dataset with 10^7 cells. Recent work (Qiu et al., 2024) has introduced an atlas of cell states across 74 mouse embryos from 45 developmental timepoints between gastrulation and birth. We learn representations using transcriptional profiles, and measure information contained about developmental timepoint. We treat the developmental timepoint as a discrete variable.

In addition to this large scale dataset, we additionally replicate scaling analyses on smaller datasets studied in Gowri et al. (2025) in order to understand the scaling behavior of large Geneformer models trained on these datasets. In particular, we study the human peripheral mononuclear blood cell CITE-seq dataset from Hao et al. (2021), the lineage-traced scRNA-seq dataset of mouse hematopoeitic stem cells from Weinreb et al. (2020), and a spatially resolved transcriptomic dataset of a mouse brain from Vizgen (2021).

Scaling analysis for Geneformer Another key contribution of this work is to explore large-scale transformer based models. While previous work (Gowri et al., 2025) considered a small scale transformer based model, here we use an implementation of Geneformer with 13 million parameters. Architecture details can be found in Sec. A. In addition to this Geneformer implementation, we also include the previously studied baselines of random projection, PCA, and scVI (Lopez et al., 2018).

3. Results

Our experimental results are shown in Fig. 1. We will first briefly compare the performance of each modelling approach, then show agreement between our experimental results and the previously proposed noise-scaling form (Eqn. 1).

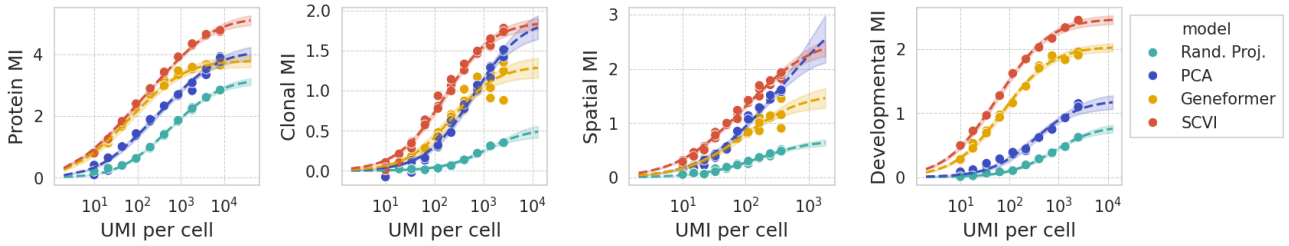

Figure 2. Measurement noise scaling laws describe observed behavior. Scatterplot points indicate experimental results when using the largest training dataset size for each signal. Dotted line indicates fit of Eqn. 1. Shaded bands indicate 2σ confidence interval.

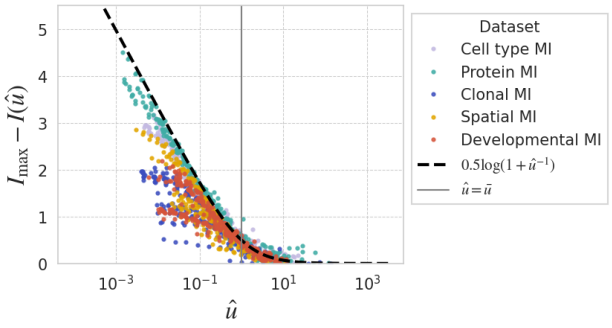

Figure 3. Observations collapse onto a universal scaling curve. For all PCA, VAE, and Geneformer models, rescaling UMI per cell and $I(u)$ falls onto a single curve when u is sufficiently large.

3.1. Model comparison

Overall performance In almost all settings considered, scVI shows the highest quality representations (with respect to information probing). PCA and Geneformer perform relatively similarly to each other, with lower performance than scVI. In the large-scale developmental dataset, Geneformer outperforms PCA, while PCA generally outperforms Geneformer in the smaller 10^5 cell datasets.

Cell number scaling behavior As expected, random projections do not improve with larger training datasets: these representations are independent of the training data. PCA also benefits minimally from increasing training dataset size beyond 10^3 cells, in line with the intuition that linear models converge quickly. Geneformer and scVI both benefit much more significantly from increasing training dataset size.

3.2. Noise scaling behavior

We next show that the noise-scaling form (Eqn. 1) closely fits our observations, even for large datasets and large models. We first show that when measurements of $I(u)$ and u are rescaled, all experimental observations collapse onto an asymptotic form of Eqn. 1 for $u \gg 2^{-2I_{\max}}$, deviating only when u or $I_{\max}$ is small (Fig. 3).

We also directly fit the noise-scaling form for individual curves with fixed cell number. For the largest dataset sizes, we show the direct fits in Fig. 2.

4. Discussion

This work demonstrates that measurement noise scaling laws persist in large models trained on large scale scRNA-seq datasets of millions of cells. However, many threads of this work remain in progress.

First, while we have shown that measurement noise scaling laws describe the relationship between model performance and transcripts captured, Eqn. 1 does not govern how performance scales with dataset size. An important future direction is to understand how these scaling axes interact – to what extent can noisy data be compensated for by an increase in training samples?

Another interesting future direction is to more carefully study the role of model scale. Here, we used a fixed model sizes due to practical constraints. However, there may be rich joint scaling behavior between model size, dataset size, and measurement noise.

A related limitation of this work is the lack of extensive hyperparameter search for the scVI and Geneformer models. In Fig. 1, models are trained in a consistent fashion between dataset sizes, although optimal hyper-parameters may vary.

Overall, this work extends the empirical evidence for measurement noise scaling laws. By understanding the role of measurement noise in deep learning model performance, this work-in-progress points toward practical implications for designing and generating large scale scRNA-seq datasets for deep learning.

Software and Data

All data and code in this paper are publicly available. Source transcriptomic datasets can be found associated with their respective publications (Hao et al., 2021; Weinreb et al., 2020; Vizgen, 2021; Qiu et al., 2024), and code to reproduce the results in this paper can be found at this link . This work benefits from several open-source software development efforts (Virshup et al., 2024; Pedregosa et al., 2011; Gayoso et al., 2022).

Acknowledgements

We thank Yair Hoffman for sharing with us efficient algorithms for sampling large datasets. This work was supported by NIH grant R01GM153805, and by an Edward Mallinckrodt Jr Scholar Award to AMK.

References

Cui, H., Wang, C., Maan, H., Pang, K., Luo, F., Duan, N., and Wang, B. scGPT: toward building a foundation model for single-cell multi-omics using generative AI. *Nat. Methods*, 21(8):1470–1480, August 2024.

Gayoso, A., Lopez, R., Xing, G., Boyeau, P., Valiollah Pour Amiri, V., Hong, J., Wu, K., Jayasuriya, M., Mehlman, E., Langevin, M., Liu, Y., Samaran, J., Misrachi, G., Nazaret, A., Clivio, O., Xu, C., Ashuach, T., Gabitto, M., Lotfollahi, M., Svensson, V., da Veiga Beltrame, E., Kleshchevnikov, V., Talavera-López, C., Pachter, L., Theis, F. J., Streets, A., Jordan, M. I., Regier, J., and Yosef, N. A python library for probabilistic analysis of single-cell omics data. *Nature Biotechnology*, Feb 2022. ISSN 1546-1696. doi: 10.1038/s41587-021-01206-w. URL https://doi.org/10.1038/s41587-021-01206-w.

Gowri, G., Lun, X., Klein, A., and Yin, P. Approximating mutual information of high-dimensional variables using learned representations. *Advances in Neural Information Processing Systems*, 37:132843–132875, 2024.

Gowri, G., Yin, P., and Klein, A. M. Measurement noise scaling laws for cellular representation learning, 2025. URL https://arxiv.org/abs/2503.02726.

Hao, M., Gong, J., Zeng, X., Liu, C., Guo, Y., Cheng, X., Wang, T., Ma, J., Zhang, X., and Song, L. Large-scale foundation model on single-cell transcriptomics. *Nat. Methods*, 21(8):1481–1491, August 2024.

Hao, Y., Hao, S., Andersen-Nissen, E., Mauck, 3rd, W. M., Zheng, S., Butler, A., Lee, M. J., Wilk, A. J., Darby, C., Zager, M., Hoffman, P., Stoeckius, M., Papalexi, E., Mimitou, E. P., Jain, J., Srivastava, A., Stuart, T., Fleming, L. M., Yeung, B., Rogers, A. J., McElrath, J. M., Blish, C. A., Gottardo, R., Smibert, P., and Satija, R. Integrated analysis of multimodal single-cell data. *Cell*, 184(13):3573–3587.e29, 24 June 2021. ISSN 0092-8674,1097-4172. doi: 10.1016/j.cell.2021.04.048.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., and Amodei, D. Scaling laws for neural language models. 2020.

Lopez, R., Regier, J., Cole, M. B., Jordan, M. I., and Yosef, N. Deep generative modeling for single-cell transcriptomics. *Nat. Methods*, 15(12):1053–1058, December 2018.

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

Qiu, C., Martin, B. K., Welsh, I. C., Daza, R. M., Le, T.-M., Huang, X., Nichols, E. K., Taylor, M. L., Fulton, O., O'Day, D. R., Gomes, A. R., Ilcisin, S., Srivatsan, S., Deng, X., Disteche, C. M., Noble, W. S., Hamazaki, N., Moens, C. B., Kimelman, D., Cao, J., Schier, A. F., Spielmann, M., Murray, S. A., Trapnell, C., and Shendure, J. A single-cell time-lapse of mouse prenatal development from gastrula to birth. *Nature*, 626(8001):1084–1093, February 2024.

Rosenfeld, J. S., Rosenfeld, A., Belinkov, Y., and Shavit, N. A constructive prediction of the generalization error across scales, 2019. URL https://arxiv.org/abs/1909.12673.

Theodoris, C. V., Xiao, L., Chopra, A., Chaffin, M. D., Al Sayed, Z. R., Hill, M. C., Mantineo, H., Brydon, E. M., Zeng, Z., Liu, X. S., and Ellinor, P. T. Transfer learning enables predictions in network biology. *Nature*, 618(7965):616–624, June 2023.

Virshup, I., Rybakov, S., Theis, F. J., Angerer, P., and Wolf, F. A. anndata: Access and store annotated data matrices. *J. Open Source Softw.*, 9(101):4371, September 2024.

Vizgen. Vizgen Data Release V1.0, May 2021. Title of the publication associated with this dataset: Mouse Brain Receptor Map.

Weinreb, C., Rodriguez-Fraticelli, A., Camargo, F. D., and Klein, A. M. Lineage tracing on transcriptional landscapes links state to fate during differentiation. *Science*, 367(6479), 14 February 2020. ISSN 0036-8075,1095-9203. doi: 10.1126/science.aaw3381.

A. Geneformer architecture details

In all experiments, we use a Geneformer model with the following architectural choices: context length of 512, embeddings with 256 dimensions, feedfoward size of 512, 4 attention heads, and 3 encoder units. Full implementation details and training code can be found at this link.