Measurement noise scaling laws for cellular representation learning

Gokul Gowri[1, 3*†], **Igor Sadalski**[2†], Dan Raviv[2], Peng Yin[1], Jonathan Rosenfeld[2], Allon Klein[1*]

[1]Department of Systems Biology, Harvard Medical School, Boston, MA, USA.
[2]Somite Therapeutics, Boston, MA, USA.
[3]Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, Cambridge, MA, USA.

*Corresponding author(s). E-mail(s): gokulg@mit.edu; allon_klein@hms.harvard.edu;
†These authors contributed equally

Abstract

Large genomic and imaging datasets can be used to fit models that learn representations of cellular systems, extracting informative structure from data. In other domains, model performance improves predictably with dataset size, providing a basis for allocating data and computation. In biological data, however, performance is also limited by measurement noise arising from technical factors such as molecular undersampling or imaging variability. By learning representations of single-cell genomic and imaging data, we show that noise defines a distinct axis along which performance improves predictably across tasks. This scaling follows a simple logarithmic law that is consistent across model types, tasks, and datasets, and can be derived quantitatively from a model of noise propagation. We identify robustness to noise and saturating performance as properties that vary across models and tasks. Applied to a 12-million-cell mouse embryogenesis dataset, a large Transformer-based model shows greater robustness but lower saturating performance than a variational autoencoder-based model.

Keywords: scaling laws, single-cell analysis

Introduction

Cellular profiles obtained by single-cell RNA sequencing (scRNA-seq) and high-content imaging now span diverse tissues, developmental stages, disease states, and experimental perturbations [1, 2]. These large datasets (collectively $> 10^8$ samples) create opportunities to identify shared cellular states across experimental contexts and predict responses to novel perturbations [3, 4]. To realize these opportunities, representation learning models are used to capture biologically meaningful variation, while filtering out technical nuisance factors [5]. Several deep learning approaches underlie such models to date, including transformer-based architectures, autoencoder-based architectures, and contrastive losses [6–9].

In domains outside of biology including natural language processing, image processing and chemical informatics, large model development has been guided by the study of model scalability. Choices in architecture, data collection efforts, and training strategies are guided by deep learning scaling laws, which are empirical relationships that describe how model performance improves with increases in key resources like data, compute, and model parameters [10–15].

In biology, model performance can also be limited by noise in the data used for model training. A few specific data modalities, such as DNA sequence, exist in large repositories with reasonably low error rates ($< 10^{-2}$ errors/nucleotide, [16]) but the majority of biological data modalities are more prone to measurement noise. scRNA-Seq and spatially-resolved transcriptomics, for example, are methods fundamentally limited by the low numbers of mRNA molecules per gene per cell. Though measurement sensitivity is increasing with ongoing development of these methods [17], for many existing technologies the probability of detecting a given mRNA molecule is well below 50%, and in some cases the detection rate is further decreased by insufficient sequence depth [18]. As a result, measured transcript counts are subject to undersampling noise. Fluorescent microscopy imaging is also prone to noise of different types including background signal, quantum yield and resolution [19].

In contrast to the scaling of model performance with data set size and model size, much less is known about the role of measurement noise on the ability of a model to learn meaningful representations. In textual representation by large language models (LLMs), errors in training data lead to degraded performance, even in the limit of infinite data [20]. However, textual data used in LLM training are much less noisy than biological data. As representation models are being developed for diverse biological tasks, understanding how noise alters the learning rate of models could be important.

Here, we recapitulate sample-size scaling in the quality of learned representations of scRNA-Seq, spatial transcriptomics and image data, and we show evidence for a general and quantitative scaling relationship between measurement noise and model performance. To show this law, we introduce an information-theoretic framework for studying the scalability of representation learning models with respect to changes in measurement noise and dataset size. We show that the noise-scaling law can be derived by analogy to additive Gaussian noise channels, and that the resulting theoretical framework can be used to guide experimental design. When applied to a 12-million-cell mouse embryogenesis dataset [21], our framework suggests that a Transformer-based model is more robust to noise, but has lower saturating performance than a varational autoencoder-based model.

Results

A metric for representation-learning model performance

In neural scaling analyses, it is typical to evaluate the quality of models directly by evaluating their loss in reconstructing test data [10–12, 14]. However, model loss is not comparable between model families, or even for a single model applied to data with different statistical properties [22] such as different noise properties. Therefore, to study the effect of noise on representation learning model performance, we introduced an alternative approach to measuring representation quality, by estimating the mutual information (MI) between the representations learnt by a model and some information about each sample that remains hidden until after learning is completed. Formally, this approach is a generalization of linear probing [23, 24], which estimates MI between a representation and a classification label. Our generalization uses a neural network-based estimator of MI that accommodates high-dimensional and continuous auxiliary signals[25]. This approach provides a performance metric that is comparable between model types and noise levels in a given data set.

We evaluated representation model performance for four test data sets, each of which provides single cell transcriptional state with some additional auxiliary signal as follows:

1. **Developmental time**, using an atlas of $\sim 10^7$ cells profiled by scRNA-seq across mouse development where developmental time is quantified by embryonic stage [21].
2. **Surface protein abundances** of $\sim 10^5$ peripheral mononuclear blood cells (PBMCs) measured by an antibody panel through CITE-seq [26].
3. **Transcriptional profile of a clonally related cell** in $\sim 10^5$ mouse hematopoietic stem cells measured using lineage-traced scRNA-seq [27].
4. **Transcriptional profile of a spatially adjacent cell** in a coronal mouse brain section of $\sim 10^5$ cells measured using MERFISH [28].

For each of these, we evaluated two linear baselines: random projection and dimensionality reduction by principal component analysis (PCA), and we compared these to two modern generative models: single cell variational inference (SCVI) [29] and Geneformer [6]. SCVI is a variational autoencoder designed to compress high-dimensional gene expression information into a low-dimensional latent space. Geneformer instead uses a Transformer-based language model that maps gene expression vectors to sequences by ordering gene-specific tokens based on expression level.

In all cases, to facilitate consistent comparisons, we learned representations on one data subset, and then evaluated performance in a separate, fixed held-out subset. This approach ensures that observed differences in mutual information are attributable to variations in the representations themselves, rather than estimation artifacts.

Cell number scaling for cellular representations

As a baseline for understanding the impact of noise on model learning, we first tested whether auxiliary-MI performance, I, shows expected scaling behavior with the number of samples N – here, single cells – used in training. In large language models, performance scales as a power of sample number [14], and a similar law is seen for cell representation models [30]. We indeed found that I is well-described by a saturating power-law across all deep learning models and auxiliary tasks, $I(N) = I_\infty - (N/N_{sat})^{-s}$, where the parameters I_∞, N_{sat}, s characterize how each model learns from new data (fit residual sum of squares $R^2 = 0.913 \pm 0.008$ across $n = 44$ model and task combinations). The fits are shown collectively across models and datasets in **Fig. 1xx**, with parameter values and model comparisons in **Figs. S-XX,YY**, **Table S-xx**, and Supplementary Text **S-I**.

Noise scaling for cellular representations

Although deep learning models are colloquially thought to be strong denoisers, the degree to which cellular representation learning models are robust to noise in their training data is unknown. A noise robust model would exhibit a regime in which the informativity of learned representations remains stable despite increasing noise levels. We evaluated the extent to which models are noise robust by simulating increasing measurement noise through downsampling observed transcript counts, and subsequently evaluating the quality of the learned representations using auxiliary-MI performance.

The dependence of model performance I on the degree of downsampling noise is shown in **Fig. 1d**. As expected, reducing the depth per cell degrades the performance of all models, across all datasets. Of note, no model or dataset exhibits large regimes of noise robustness. Instead, many of the measured performance curves are sigmoidal, indicating only limited robustness at the transcript levels present in the original datasets before performance steadily deteriorates (**Fig. 1d**). A subset of the curves show a 'hockey-stick' shape, indicating negligible robustness to downsampling noise, even at full transcript levels.

Neural scaling laws provide expectations for how model performance improves with additional computational or data resources. Loss of representation quality as a function of downsampling noise produces families of smooth, sigmoidal performance curves, suggesting that a similarly simple quantitative relationship might capture how measurement noise constrains biological representation learning. Such a relationship would be valuable for experimental design, enabling principled allocation of sequencing depth and cell numbers in the same way that neural scaling laws guide resource decisions in large-scale machine learning.

To investigate whether the observed noise–performance behavior is predictable, we turned to a classical model of information loss in noisy communication channels (see **Box 1**). This framework extends established information-theoretic results [31] to derive an analytical relationship between the signal-to-noise ratio of a measurement, $\eta =$SNR, and the mutual information preserved about an underlying external variable. This analysis (Eq. 2 in Box 1) yields a closed-form prediction for how auxiliary-MI performance should depend on sequencing depth:

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

where $\mathcal{I}_{\max}$ is the maximal information that can be extracted from a noiseless measurement at a fixed sample size, and $\bar{\eta}$ serves as a measure of noise robustness (specifically, the signal-to-noise ratio at which a model can gain at most 1/2 a bit of information by increasing measurement sensitivity). To connect this general relationship to cellular measurements, we note that the signal-to-noise ratio introduced by molecular undersampling follows Poisson statistics, $\eta^2 = \mathrm{CV}^{-2} \propto \mathrm{UMI}$ [32], and $\bar{\eta}$ then takes on units of UMIs per cell.

In Fig. 1d, the theoretical curves defined by Eq. (1) (dotted lines) closely match the empirical performance curves (scatter points) across models and datasets. Strikingly, the noise-scaling relationship holds across model architectures and across single-cell datasets spanning nearly five orders of magnitude in sample size. When rescaled by their fitted $\mathcal{I}_{\max}$ and $\bar{\eta}$ values, XXX empirically measured curves from all the model families collapse onto a single universal relationship (Fig. 1e), indicating that a shared principle governs how measurement noise in transcriptomic data limits representation learning.

The fitted noise-scaling parameters from Eq. 1 provide a compact summary of the noise-robustness of a model. In particular, $\bar{\eta}$ reflects each model's effective noise tolerance, while $\mathcal{I}_{\max}$ captures its asymptotic capacity in the absence of measurement noise. For a given auxiliary task, models that combine low $\bar{\eta}$ with high $\mathcal{I}_{\max}$ are therefore preferred.

In Fig. 1f, we compare inferred $\bar{\eta}$ and $\mathcal{I}_{\max}$ values across models. Geneformer consistently shows the greatest robustness to noise: across all tasks, it approaches within 0.5 bits of its asymptotic performance at fewer than 1,000 UMI per cell. scVI displays similarly low noise sensitivity for three of the four tasks, but in the protein-abundance task it becomes noise-sensitized at $\sim$ 4,000 UMI per cell. PCA, by contrast, shows far greater sensitivity to noise, with $\bar{\eta}$ values 2.5–12.8-fold larger than those of Geneformer, consistent with the limited denoising capacity of linear methods.

Despite its robustness to noise, Geneformer is not a strong model in terms of its capacity. Across all tasks, its capacity, $\mathcal{I}_{\max}$, is lower than those of scVI by 0.4–1.4 bits – corresponding to approximately halving the complexity of the captured signal. This difference in performance is not only in its asymptotic capacity, but also at the noise level present in the datasets (**Fig. 1d**). Thus, scVI ultimately extracts more auxiliary information in the limit of low noise. It is possible that other models may simultaneously show noise robustness and higher information capacity.

Box 1: A model of noise scaling in representation learning

The empirical noise–performance curves in Fig. 1d suggest that a simple theoretical relationship underlies how measurement noise limits the information extractable by representation models. A classical setting in which such limits are analytically tractable is a Gaussian noise channel, where both the signal and the noise are modeled as Gaussian random variables. Although simplified, this framework captures the essential effect of diminishing returns: as measurement quality improves, each additional increment in signal-to-noise ratio (SNR) conveys progressively less new information. We use it to derive the scaling form in Eq. (1).

Let X, Y be multivariate Gaussian random vectors representing the system state and an auxiliary signal, and let Z be a noisy measurement of X with SNR η:

$$Y \sim \mathcal{N}(0, \Sigma_Y), \qquad X = Y + \mathcal{N}(0, \Sigma_U), \qquad Z = \eta X + \mathcal{N}(0, I_n).$$

The mutual information between Y and Z—the amount of auxiliary signal retained after measurement—follows the standard expression for Gaussian vector noise channels [31, 33] (proof in Appendix **??**):

$$I(Y;Z) = \tfrac{1}{2}\log\frac{\det(\Sigma_Y + \Sigma_U + \eta^{-2} I_n)}{\det(\Sigma_U + \eta^{-2} I_n)}.$$

For the scalar case ($n = 1$), where $\Sigma_Y = \sigma_Y^2$ and $\Sigma_U = \sigma_U^2$,

$$I(Y;Z) = \tfrac{1}{2}\log\frac{\eta^2(\sigma_Y^2 + \sigma_U^2) + 1}{1 + \sigma_U^2 \eta^2}. \tag{2}$$

Two characteristic quantities govern this scaling:

$$\mathcal{I}_{\max} = \lim_{\alpha\to\infty} I(Y;Z) = \tfrac{1}{2}\log\frac{\sigma_Y^2 + \sigma_U^2}{\sigma_U^2},$$

the maximal achievable information, and $\bar{\eta} = 1/\sigma_U^2$, an effective noise scale. Substituting $\mathcal{I}_{\max}$ and $\bar{\eta}$ into $I(Y;Z)$ recovers precisely the empirical noise-scaling relationship of Eq. 1. Despite its simplicity, this model captures the universal shape of the performance–noise curves observed across datasets and architectures.

Generalization of noise scaling

The noise scaling observed in single-cell representation learning may extend to other data modalities. The scaling law (Eq. 1) depends only on the signal-to-noise ratio η, and a model that explains this law (**Box 1**) is not specific to transcriptomic data. To test whether this framework generalizes, we examined whether Eq. 1 quantitatively predicts noise–performance relationships in image representation models, and then in a protein sequence model.

For image representation, we used MobileNetV2, a lightweight convolutional architecture designed for image classification [34]. We trained and evaluated this model on a 5-class subset of the Caltech101 dataset [35], consisting of 240×240 pixel images with 1,354 total images. Images were perturbed with two distinct forms of degradation: additive Gaussian noise and reduced spatial resolution. Pixel-wise Gaussian noise is common in imaging measurements [36]. We then used auxiliary-MI to evaluate model performance under both forms of image noise. As with the transcriptomic models, we quantified performance using auxiliary-MI, here measuring the mutual information between the learned representations and the true image labels. We trained the MobileNetV2 models and computed the auxiliary-MI between predicted and true labels on held-out images, assessing performance for two tasks: classification of all class labels, as well as multiple one-vs-all problems. We introduced Gaussian noise with $\eta = 1/\sigma_N^2$, where σ_N is the noise standard deviation, while resolution degradation was introduced by averaging local pixel neighborhoods, with $\eta = 1/f$ for downsampling factor f. Downsampling introduces a noise SNR f (**Fig. Sxxx**). For both types of noise, we found that Eq. 1 accurately reproduced the observed noise–performance curves for all classification tasks (**Fig. 1g**, $R^2 = xxx$). [MODIFY THE ABOVE PARAGRAPHS TO INCLUDE MORE IMAGE DATA SETS (NO NEED TO ADD A NEW PARAGRAPH].

A similar pattern emerged in a representation learning model of protein sequence. We finetuned a small ESM2 foundation model (8M parameters) on a set of ~20,000 SARS-CoV-2 spike protein sequences spanning the course of the pandemic, after introducing controlled levels of amino-acid substitution to simulate increasingly corrupted measurements. We then quantified auxiliary-MI between the learned representations and the collection date of each sequence, measured as time since the pandemic outbreak in January 2020. Despite the discrete and highly structured nature of protein sequences, the resulting noise–performance curves again conformed closely to the form predicted by Eq. 1, with increasing substitution rates driving systematic and predictable declines in mutual information (**Fig. X**). The shared behavior of the models is demonstrated by collapsing the XXX curves by appropriate rescaling in **Fig. 1f**. Together, this analysis adds to the evidence that noise in training data is a systematic determinant of representation quality—one that can be modeled alongside sample size when characterizing learning behavior.

Noise scaling and experimental design

Measurement noise scaling laws can be used to determine the data quality or sample quantity required to achieve a specified level of representation performance. The parameter $\bar{\eta}$ from Eq. 1 directly reports the measurement sensitivity at which model performance reaches within 0.5 bits (or approximately 70%) of its asymptotic value. More generally, inverting Eq. 1 yields a function $\eta(\mathcal{I})$ that predicts the minimum signal-to-noise ratio needed for a model to achieve a desired information content with respect to a given auxiliary variable (see **Supp. Text 1**).

For transcriptomic data, η is proportional to the total UMIs per cell, enabling an estimate of the sequencing depth needed for a representation to reach, for example, 90% of its maximum informativity. These depth requirements (UMI90) are reported for all model–task pairs in Table ??. Several clear patterns emerge. Geneformer consistently operates below its UMI90 on all datasets examined, indicating that its performance is already near its asymptotic limit under typical sequencing depths. In contrast, UMI90 for scVI exceeds the observed UMI counts for protein abundance, spatial information, and clonal information tasks—suggesting that these tasks remain sensitivity-limited and would benefit from deeper sequencing. These examples illustrate how noise-scaling relationships can guide the choice of models and allocation of sequencing depth across tasks with different intrinsic difficulty.

To assess whether these experimental-design conclusions extend beyond transcriptomics, we applied the same analysis to image and protein-sequence representation learning. For image classification (**Fig. 1xx**), the fitted $\mathcal{I}_{\max}$ values closely matched the theoretical maximum determined by the label entropy, and the fitted $\bar{\eta}$ values provided interpretable design guidance. [**REPLACE THE FOLLOWING WITH η_90 FOR CONSISTENCY**]. In the 5-way task, $e\bar{t}a \approx 0.13$ corresponds to an effective resolution threshold of $\sim 31 \times 31$ pixels: images downsampled beyond this point lose more than 0.5 bits of label information. Certain classes (e.g., "watch") exhibited a steeper performance decay, indicating that their recognition relies on higher-resolution features. These results parallel the transcriptomic findings and show that noise tolerance thresholds can be used to rationally plan training of image models as well. For protein sequence, auxiliary-MI remained stable up to substitution rates of approximately 1 in 1,000 amino acids—a noise level well above what is typical in modern sequencing. This is consistent with DNA and protein sequence models to date being able to largely ignore measurement noise.

Discussion

Noise in training data inevitably affects model performance, but it has remained unclear whether there exist predictable, quantitative rules governing how representation quality degrades as noise increases. Across single-cell transcriptomics, imaging, and protein sequence data, we find that auxiliary-task performance follows a characteristic sigmoidal scaling curve. Models retain robust performance above a modality-specific noise threshold, after which representation quality declines approximately logarithmically with increasing noise. A remarkably simple information-theoretic model captures this relationship and recovers the empirical scaling form observed across more than 10^3 trained representations. These results suggest that predictable noise-dependent learning curves may be a common feature across diverse biological data modalities.

This work provides practical guidance for designing and evaluating biological representation models. The fitted scaling parameters $\bar{\eta}$ and $\mathcal{I}_{\max}$ jointly characterize model behavior: $\bar{\eta}$ reflects noise robustness, while $\mathcal{I}_{\max}$ represents the maximal task-relevant information that a model can encode [THE ISSUE IS THAT THIS IS ALSO THE INFORMATION IN TEST DATA, WHICH CHANGES BETWEEN NOISE LEVELS]. Nonlinear models such as Geneformer and scVI exhibit substantially greater robustness to measurement noise than PCA, consistent with the expectation that nonlinear architectures more effectively denoise sparse molecular measurements. However, robustness alone is insufficient. Geneformer, despite its stability under noise, often attains a relatively low $\mathcal{I}_{\max}$, capturing less auxiliary information than scVI and, for certain tasks, even linear baselines. These results emphasize that noise robustness and representational capacity must be jointly optimized in model design.

Noise scaling has implications for experimental design, particularly for large-scale single-cell profiling. Our analysis shows that some tasks, such as our test tasks of predicting surface-protein or spatial information from scRNA-seq, remain sensitivity-limited even in current datasets. These tasks would benefit substantially from higher per-cell transcript counts. Conversely, for tasks such as predicting developmental stage in the mouse embryo atlas, existing sequencing depth is already sufficient to approach the representational limit. These distinctions highlight that improvements in measurement quality, rather than cell number alone, may be the most impactful direction for next-generation atlases and molecular profiling initiatives.

More broadly, considering measurement noise as an additional scaling axis, parallel to well-established roles of dataset and model size in neural scaling, suggests a more complete picture of representation learning in 'measurement-bound' fields such as biology. Noise imposes a predictable, quantifiable constraint that can be analytically modeled and experimentally manipulated. This creates opportunities for joint optimization of dataset size and measurement sensitivity, and for designing assays that sit on or near the optimal learning curve for a given task.

Several questions remain. First, we have still only demonstrated noise scaling in a small number of modeling tasks. Second, even for the tasks at hand, we have only evaluated a small number of model architectures. The high cost of training moder foundation models makes it impractical for us to evaluate additional models. It is possible that finetuning of pre-trained models may provide a faithful probe of noise-tolerances of a model, allowing systematic evaluation of additional models. Third, an open theoretical question is to understand the origin of the scaling law. The model we introduce here (**Box 1**) is exact for scalar Gaussian channels, yet it fits high-dimensional biological data surprisingly well. Understanding why this is the case, and under what conditions noise scaling breaks down, represents a theoretical direction. Finally, our analysis has treated measurement noise and sample size separately; developing a joint scaling law that unifies both axes would further clarify how to allocate resources to build predictive models of high-dimensional biological systems.

In sum, our findings suggest that measurement noise is a predictable and actionable determinant of representation model performance, one that can be optimized alongside dataset size to guide both model development and experimental design across biological modalities.

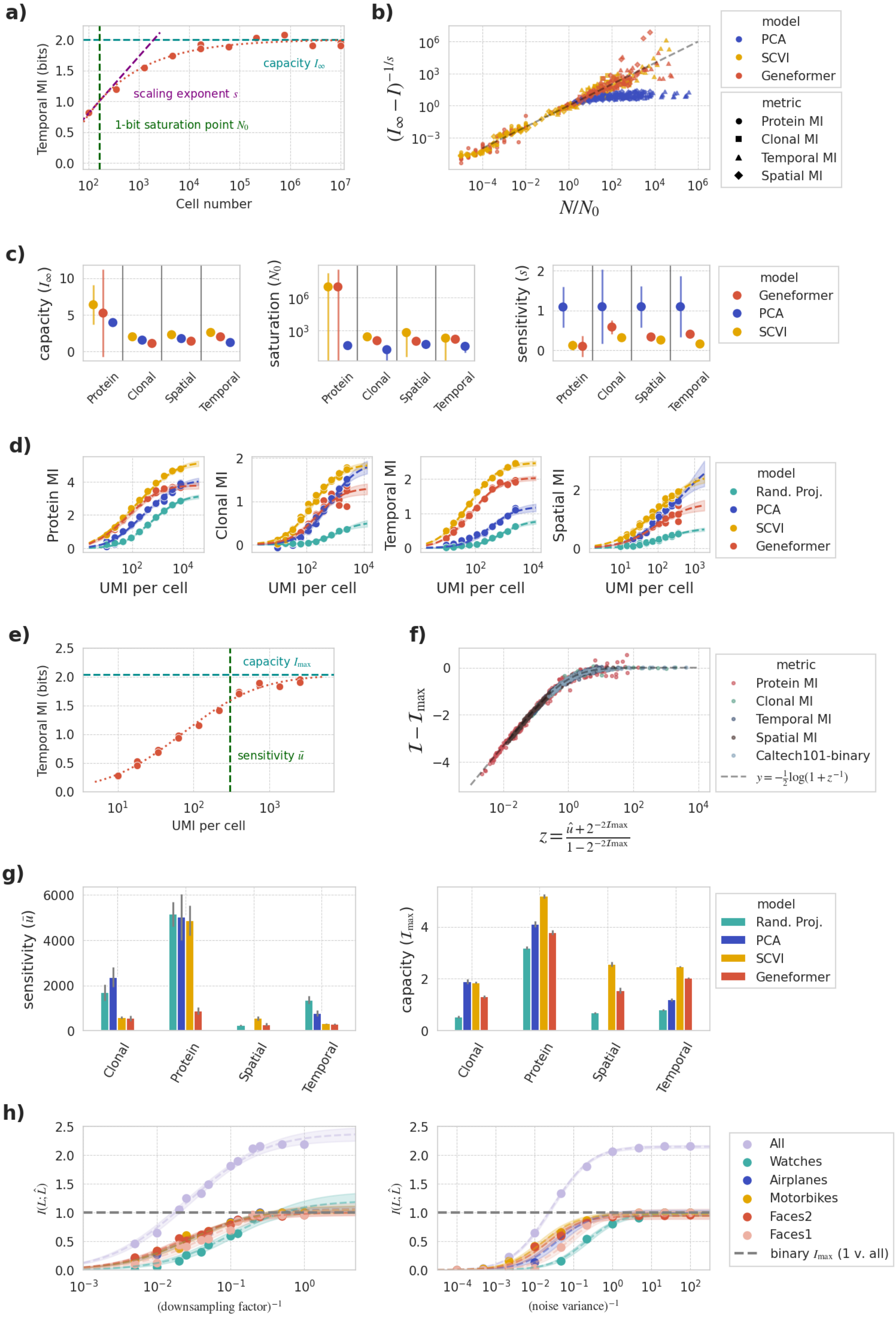
a)
Temporal MI (bits)
capacity I_∞
scaling exponent s
1-bit saturation point N_0
Cell number
b)
$(I_\infty - I)^{-1/s}$
N/N_0
model
PCA
SCVI
Geneformer
metric
Protein MI
Clonal MI
Temporal MI
Spatial MI
c)
capacity (I_∞)
saturation (N_0)
sensitivity (s)
Protein
Clonal
Spatial
Temporal
model
Geneformer
PCA
SCVI
d)
Protein MI
Clonal MI
Temporal MI
Spatial MI
UMI per cell
model
Rand. Proj.
PCA
SCVI
Geneformer
e)
Temporal MI (bits)
capacity $I_{\max}$
sensitivity $\tilde{u}$
UMI per cell
f)
$I - I_{\max}$
metric
Protein MI
Clonal MI
Temporal MI
Spatial MI
Caltech101-binary
$y = -\frac{1}{2}\log(1+z^{-1})$
$z = \frac{\hat{u} + 2^{-2I_{\max}}}{1 - 2^{-2I_{\max}}}$
g)
sensitivity ($\tilde{u}$)
capacity ($I_{\max}$)
Clonal
Protein
Spatial
Temporal
model
Rand. Proj.
PCA
SCVI
Geneformer
h)
$I(L;\hat{L})$
(downsampling factor)$^{-1}$
(noise variance)$^{-1}$
All
Watches
Airplanes
Motorbikes
Faces2
Faces1
binary $I_{\max}$ (1 v. all)

Fig. 1 *(previous page)*: **Scaling laws for cellular representation learning. (a)** Geneformer representation quality, measured by information about developmental time, as a function of number of training data points drawn from a mouse embryo development atlas [21]. Theory curve is shown with a dashed line. Cell number scaling parameters $I_\infty,\ s,\ N_0$ are annotated on the theory curve. **(b)** Scaling collapse of 54 different cell number scaling curves across three model families and four datasets. Datasets with transcript counts downsampled by more than one order of magnitude are omitted. **(c)** Comparison of cell number scaling parameters across model families and representation quality metrics. **(d)** Observations and noise scaling law fits for representation quality as a function of molecules detected per cell. Confidence bands show 2σ interval. **(e)** Geneformer representation quality, measured by information about developmental time, as a function of number of transcripts captured per cell. Theory curve is shown with a dashed line. Noise scaling parameters $I_{\max},\ \bar{u}$ are annotated on the theory curve. **(f)** Scaling collapse of 112 different noise scaling curves across four model families and five datasets. Curves with unconstrained $\mathcal{I}_{\max}$ are omitted. **(g)** Comparison of noise scaling parameters across model families with fixed-size unsubsampled datasets. Parameters for PCA on the spatial metric are unconstrained and omitted from the plot. **(h)** Observations and noise scaling law fits for image classification performance of Mobilenetv2 models [34]. Confidence bands show 2σ interval.

Acknowledgements. This work is supported by funding from NIH Pioneer Award DP1GM133052, R01HG012926 to P.Y., and Molecular Robotics Initiative at the Wyss Institute. A.M.K. acknowledges support of an Edward Mallinckrodt Jr. Scholar Award.

References

[1] Yao, Z., Velthoven, C.T.J., Kunst, M., Zhang, M., McMillen, D., Lee, C., Jung, W., Goldy, J., Abdelhak, A., Aitken, M., Baker, K., Baker, P., Barkan, E., Bertagnolli, D., Bhandiwad, A., Bielstein, C., Bishwakarma, P., Campos, J., Carey, D., Casper, T., Chakka, A.B., Chakrabarty, R., Chavan, S., Chen, M., Clark, M., Close, J., Crichton, K., Daniel, S., DiValentin, P., Dolbeare, T., Ellingwood, L., Fiabane, E., Fliss, T., Gee, J., Gerstenberger, J., Glandon, A., Gloe, J., Gould, J., Gray, J., Guilford, N., Guzman, J., Hirschstein, D., Ho, W., Hooper, M., Huang, M., Hupp, M., Jin, K., Kroll, M., Lathia, K., Leon, A., Li, S., Long, B., Madigan, Z., Malloy, J., Malone, J., Maltzer, Z., Martin, N., McCue, R., McGinty, R., Mei, N., Melchor, J., Meyerdierks, E., Mollenkopf, T., Moonsman, S., Nguyen, T.N., Otto, S., Pham, T., Rimorin, C., Ruiz, A., Sanchez, R., Sawyer, L., Shapovalova, N., Shepard, N., Slaughterbeck, C., Sulc, J., Tieu, M., Torkelson, A., Tung, H., Valera Cuevas, N., Vance, S., Wadhwani, K., Ward, K., Levi, B., Farrell, C., Young, R., Staats, B., Wang, M.-Q.M., Thompson, C.L., Mufti, S., Pagan, C.M., Kruse, L., Dee, N., Sunkin, S.M., Esposito, L., Hawrylycz, M.J., Waters, J., Ng, L., Smith, K., Tasic, B., Zhuang, X., Zeng, H.: A high-resolution transcriptomic and spatial atlas of cell types in the whole mouse brain. Nature **624**(7991), 317–332 (2023) https://doi.org/10.1038/s41586-023-06812-z

[2] Zhang, J., Ubas, A.A., Borja, R., Svensson, V., Thomas, N., Thakar, N., Lai, I., Winters, A., Khan, U., Jones, M.G., Tran, V., Pangallo, J., Papalexi, E., Sapre, A., Nguyen, H., Sanderson, O., Nigos, M., Kaplan, O., Schroeder, S., Hariadi, B., Marrujo, S., Salvino, C.C.A., Gallareta Olivares, G., Koehler, R., Geiss, G., Rosenberg, A., Roco, C., Merico, D., Alidoust, N., Goodarzi, H., Yu, J.: *Tahoe-100M*: A Giga-scale single-cell perturbation atlas for context-dependent gene function and cellular modeling. bioRxiv, 2025–0220639398 (2025) https://doi.org/10.1101/2025.02.20.639398

[3] Bunne, C., Roohani, Y., Rosen, Y., Gupta, A., Zhang, X., Roed, M., Alexandrov, T., AlQuraishi, M., Brennan, P., Burkhardt, D.B., Califano, A., Cool, J., Dernburg, A.F., Ewing, K., Fox, E.B., Haury, M., Herr, A.E., Horvitz, E., Hsu, P.D., Jain, V., Johnson, G.R., Kalil, T., Kelley, D.R., Kelley, S.O., Kreshuk, A., Mitchison, T., Otte, S., Shendure, J., Sofroniew, N.J., Theis, F., Theodoris, C.V., Upadhyayula, S., Valer, M., Wang, B., Xing, E., Yeung-Levy, S., Zitnik, M., Karaletsos, T., Regev, A., Lundberg, E., Leskovec, J., Quake, S.R.: How to build the virtual cell with artificial intelligence: Priorities and opportunities. arXiv [q-bio.QM] (2024) arXiv:2409.11654 [q-bio.QM]

[4] Wang, H., Leskovec, J., Regev, A.: Limitations of cell embedding metrics assessed using drifting islands. Nature biotechnology, 1–4 (2025) https://doi.org/10.1038/s41587-025-02702-z

[5] Gunawan, I., Vafaee, F., Meijering, E., Lock, J.G.: An introduction to representation learning for single-cell data analysis. Cell reports methods **3**(8), 100547 (2023) https://doi.org/10.1016/j.crmeth.2023.100547

[6] 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., Ellinor, P.T.: Transfer learning enables predictions in network biology. Nature **618**(7965), 616–624 (2023) https://doi.org/10.1038/s41586-023-06139-9

[7] Cui, H., Wang, C., Maan, H., Pang, K., Luo, F., Duan, N., Wang, B.: scGPT: toward building a foundation model for single-cell multi-omics using generative AI. Nature methods **21**(8), 1470–1480 (2024) https://doi.org/10.1038/s41592-024-02201-0

[8] Heimberg, G., Kuo, T., DePianto, D.J., Salem, O., Heigl, T., Diamant, N., Scalia, G., Biancalani, T., Turley, S.J., Rock, J.R., Corrada Bravo, H., Kaminker, J., Vander Heiden, J.A., Regev, A.: A cell atlas foundation model for scalable search of similar human cells. Nature, 1–3 (2024) https://doi.org/10.1038/s41586-024-08411-y

[9] Richter, T., Bahrami, M., Xia, Y., Fischer, D.S., Theis, F.J.: Delineating the effective use of self-supervised learning in single-cell genomics. Nature machine intelligence, 1–11 (2024) https://doi.org/10.1038/s42256-024-00934-3

[10] Hestness, J., Narang, S., Ardalani, N., Diamos, G., Jun, H., Kianinejad, H., Patwary, M.M.A., Yang, Y., Zhou, Y.: Deep Learning Scaling is Predictable, Empirically. arXiv [cs.LG] (2017) arXiv:1712.00409 [cs.LG]

[11] Rosenfeld, J.S., Rosenfeld, A., Belinkov, Y., Shavit, N.: A constructive prediction of the generalization error across scales. arXiv [cs.LG] (2019) arXiv:1909.12673 [cs.LG]

[12] Kaplan, J., McCandlish, S., Henighan, T., Brown, T.B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., Amodei, D.: Scaling laws for neural language models. arXiv [cs.LG] (2020) arXiv:2001.08361 [cs.LG]

[13] Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., Casas, D.d.L., Hendricks, L.A., Welbl, J., Clark, A., Hennigan, T., Noland, E., Millican, K., Driessche, G., Damoc, B., Guy, A., Osindero, S., Simonyan, K., Elsen, E., Rae, J.W., Vinyals, O., Sifre, L.: Training Compute-Optimal Large Language Models. arXiv [cs.CL] (2022) arXiv:2203.15556 [cs.CL]

[14] Bahri, Y., Dyer, E., Kaplan, J., Lee, J., Sharma, U.: Explaining neural scaling laws. Proceedings of the National Academy of Sciences of the United States of America **121**(27), 2311878121 (2024) https://doi.org/10.1073/pnas.2311878121

[15] Chen, D., Zhu, Y., Zhang, J., Du, Y., Li, Z., Liu, Q., Wu, S., Wang, L.: Uncovering neural scaling laws in molecular Representation Learning. Neural Information Processing Systems **abs/2309.15123**, 1452–1475 (2023) https://doi.org/10.48550/arXiv.2309.15123 2309.15123

[16] Stoler, N., Nekrutenko, A.: Sequencing error profiles of Illumina sequencing instruments. NAR genomics and bioinformatics **3**(1), 019 (2021) https://doi.org/10.1093/nargab/lqab019

[17] Hagemann-Jensen, M., Ziegenhain, C., Chen, P., Ramsköld, D., Hendriks, G.-J., Larsson, A.J.M., Faridani, O.R., Sandberg, R.: Single-cell RNA counting at allele and isoform resolution using Smart-seq3. Nature biotechnology **38**(6), 708–714 (2020) https://doi.org/10.1038/s41587-020-0497-0

[18] Svensson, V., Natarajan, K.N., Ly, L.-H., Miragaia, R.J., Labalette, C., Macaulay, I.C., Cvejic, A., Teichmann, S.A.: Power analysis of single-cell RNA-sequencing experiments. Nature methods **14**(4), 381–387 (2017) https://doi.org/10.1038/nmeth.4220

[19] Lichtman, J.W., Conchello, J.-A.: Fluorescence microscopy. Nature methods **2**(12), 910–919 (2005) https://doi.org/10.1038/nmeth817

[20] Bansal, Y., Ghorbani, B., Garg, A., Zhang, B., Krikun, M., Cherry, C., Neyshabur, B., Firat, O.: Data scaling laws in NMT: The effect of noise and architecture. arXiv [cs.LG] (2022) arXiv:2202.01994 [cs.LG]

[21] 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., Shendure, J.: A single-cell time-lapse of mouse prenatal development from gastrula to birth. Nature **626**(8001), 1084–1093 (2024) https://doi.org/10.1038/s41586-024-07069-w

[22] Brandfonbrener, D., Anand, N., Vyas, N., Malach, E., Kakade, S.: Loss-to-loss prediction: Scaling laws for all datasets. arXiv [cs.LG] (2024) arXiv:2411.12925 [cs.LG]

[23] Belinkov, Y.: Probing classifiers: Promises, shortcomings, and advances. Computational linguistics (Association for Computational Linguistics) **48**(1), 207–219 (2022) https://doi.org/10.1162/coli_a_00422

[24] Pimentel, T., Valvoda, J., Maudslay, R.H., Zmigrod, R., Williams, A., Cotterell, R.: Information-theoretic probing for linguistic structure. arXiv [cs.CL] (2020) arXiv:2004.03061 [cs.CL]

[25] Gowri, G., Lun, X.-K., Klein, A.M., Yin, P.: Approximating mutual information of high-dimensional variables using learned representations. In: Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J., Zhang, C. (eds.) Advances in Neural Information Processing Systems, vol. 37, pp. 132843–132875. Curran Associates, Inc., ??? (2024)

[26] Hao, Y., Hao, S., Andersen-Nissen, E., Mauck, W.M. 3rd, 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., Satija, R.: Integrated analysis of multimodal single-cell data. Cell **184**(13), 3573–358729 (2021) https://doi.org/10.1016/j.cell.2021.04.048

[27] Weinreb, C., Rodriguez-Fraticelli, A., Camargo, F.D., Klein, A.M.: Lineage tracing on transcriptional landscapes links state to fate during differentiation. Science **367**(6479) (2020) https://doi.org/10.1126/science.aaw3381

[28] Vizgen: Vizgen Data Release V1.0. Title of the publication associated with this dataset: Mouse Brain Receptor Map (2021)

[29] Lopez, R., Regier, J., Cole, M.B., Jordan, M.I., Yosef, N.: Deep generative modeling for single-cell transcriptomics. Nature methods **15**(12), 1053–1058 (2018) https://doi.org/10.1038/s41592-018-0229-2

[30] DenAdel, A., Hughes, M., Thoutam, A., Gupta, A., Navia, A.W., Fusi, N., Raghavan, S., Winter, P.S., Amini, A.P., Crawford, L.: Evaluating the role of pre-training dataset size and diversity on single-cell foundation model performance. bioRxiv, 2024–1213628448 (2024) https://doi.org/10.1101/2024.12.13.628448

[31] Guo, D.: Gaussian channels: Information, estimation and multiuser detection. PhD thesis (2004)

[32] Klein, A.M., Mazutis, L., Akartuna, I., Tallapragada, N., Veres, A., Li, V., Peshkin, L., Weitz, D.A., Kirschner, M.W.: Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. Cell **161**(5), 1187–1201 (2015) https://doi.org/10.1016/j.cell.2015.04.044

[33] Polyanskiy, Y., Wu, Y.: Information Theory: From Coding to Learning. Cambridge university press, ??? (2024)

[34] Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.-C.: MobileNetV2: Inverted residuals

and linear bottlenecks. arXiv [cs.CV] (2018) arXiv:1801.04381 [cs.CV]

[35] Fei-Fei, L., Fergus, R., Perona, P.: Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories, 178–178 (2004)

[36] Costa, G.B.P., Contato, W.A., Nazare, T.S., Neto, J.a.E.S.B., Ponti, M.: An empirical study on the effects of different types of noise in image classification tasks. arXiv [cs.CV] (2016) arXiv:1609.02781 [cs.CV]