A measurement noise scaling law for cellular representation learning

Gokul Gowri[1, 2, 3*†], Igor Sadalski[4†], Dan Raviv[4], Peng Yin[1, 2], Jonathan Rosenfeld[4], Allon Klein[1*]

[1]Department of Systems Biology, Harvard Medical School, Boston, MA, USA.
[2]Wyss Institute for Biologically Inspired Engineering, Harvard Medical School, Boston, MA, USA.
[3]Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, Cambridge, MA, USA.
[4]Cellular Intelligence, Boston, 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 train models that learn meaningful representations of cellular systems. Across domains, model performance improves predictably with dataset size, providing a basis for allocating data and computation. Scientific data, however, is also limited by noise arising from factors such as molecular undersampling, sequencing errors, and image resolution. By learning 1,670 representation models across three data modalities (gene expression, sequence, and image data), we show that noise defines a distinct axis along which performance improves. Noise scaling follows a logarithmic law. We derive the law from a model of noise propagation, and use it to define noise robustness and saturation as model benchmarking metrics. Protein sequence representations are noise-robust; single cell transcriptomics models are not, with a Transformer-based model showing greater noise robustness but lower saturating performance than a variational autoencoder model. Noise scaling metrics may support future model evaluation and experimental design.

Keywords: representation learning, scaling laws, foundation models, single-cell transcriptomics

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 and other scientific domains, 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 has increased substantially 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, scattering, out-of-focus light, autofluorescence, lens optical resolution and detector sensitivity and pixel 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, predicted to persist 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 support the model evaluation, and allow rational planning of data-generation campaigns.

Here, we show evidence for a general and quantitative scaling relationship between measurement noise and model performance across representation learning models of scRNA-Seq, spatial transcriptomics, protein sequences, and imaging data. We show that the noise-scaling law can be derived by analogy to additive Gaussian noise channels, and that the resulting analytical form of the law can be used to define model benchmarks and identify noise-sensitivity and saturation that may guide experimental design. We then critically evaluate several models.

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 [21] 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 (**Fig. 1**). Formally, this approach is a generalization of linear probing [22, 23], which can be viewed as estimating 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 [24]. This approach provides a performance metric that is comparable between model types and noise levels in a given data set.

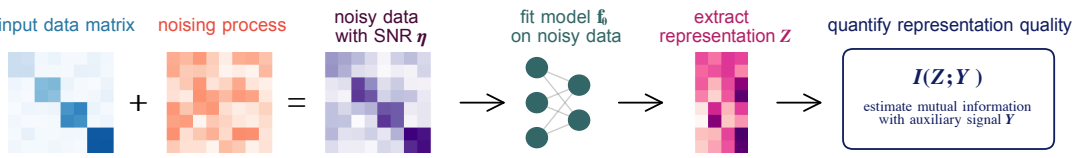
We first evaluated representation model performance for four single cell transcriptomic data sets, each of which provides an additional auxiliary signal as follows:

1. *Developmental time* of $\sim 10^7$ cells profiled by scRNA-seq across mouse development, where developmental time is quantified by embryonic stage [25].
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, with an explicit treatment of gene expression noise. Geneformer instead uses a Transformer-based language model that maps gene expression vectors to sequences by ordering gene-specific tokens based on expression level. Our implementations of scVI and Geneformer have between 1.3-116.8 million and 1.9-13 million parameters respectively (depending on gene number in each dataset), although we note that parameter counts across different model families are not directly comparable. We trained Geneformer across multiple GPUs using Deepspeed [30]. Model implementation and data preprocessing details can be found in **Supplemental Text** (sections S.I, S.II).

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.

a



b

Data modalities	Auxiliary signals	Model families	Noising process
Single-cell and spatial transcriptomics	Spatial neighbor expression (MERFISH) Clonal progenitor expression (LT-scSeq) Protein abundance (CITE-seq) Embryo stage (time series scRNA-seq)	scVI Geneformer PCA Random projection	Transcript count downsampling
Images	Kidney cell type (fluorescence microscopy) Image label (Caltech101 benchmark)	Mobilenetv2	Pixel-wise additive Gaussian Pixelation
Protein sequences	Collection month (SARS-CoV2 spike protein)	ESM2	Amino-acid substitution error

Fig. 1: **Workflow and data sets for studying noise in biological representation learning.** **(a)** The workflow used to evaluate representation model quality as a function of noise, with model performance quantified by auxiliary mutational information (see text). The process is repeated across many noise levels (signal-to-noise ratios η) to generate empirical noise-scaling curves. **(b)** Summary of the auxiliary signals, model families and noising processes used for each of the three data modalities in this study.

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, the number of single cells) used in training. In large language models, performance scales as a power of sample number [14]. We indeed found that I is well-described by a saturating power-law across all neural network-based models and auxiliary tasks, $I(N) = I_\infty - (N/N_{\text{sat}})^{-s}$, where the parameters $I_\infty, N_{\text{sat}}, s$ characterize how each model learns from new data (fit residual sum of squares $R^2 = 0.918 \pm 0.009$ across $n = 36$ model and task combinations). The fits are shown collectively across models and datasets in **Figs. 2a, b**, with parameter values and model comparisons in **Supplementary Fig. E2**. The differences in scaling between models is in agreement with prior work (**Supplementary Text S.V** and [31]), and establishes auxiliary-MI as suitable for studying the impact of noise on model performance.

Noise scaling for cellular representations

Although deep learning models are colloquially thought to be strong denoisers [32], 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. 2c**. As expected, reducing the depth per cell degrades the performance of all models, across all datasets.

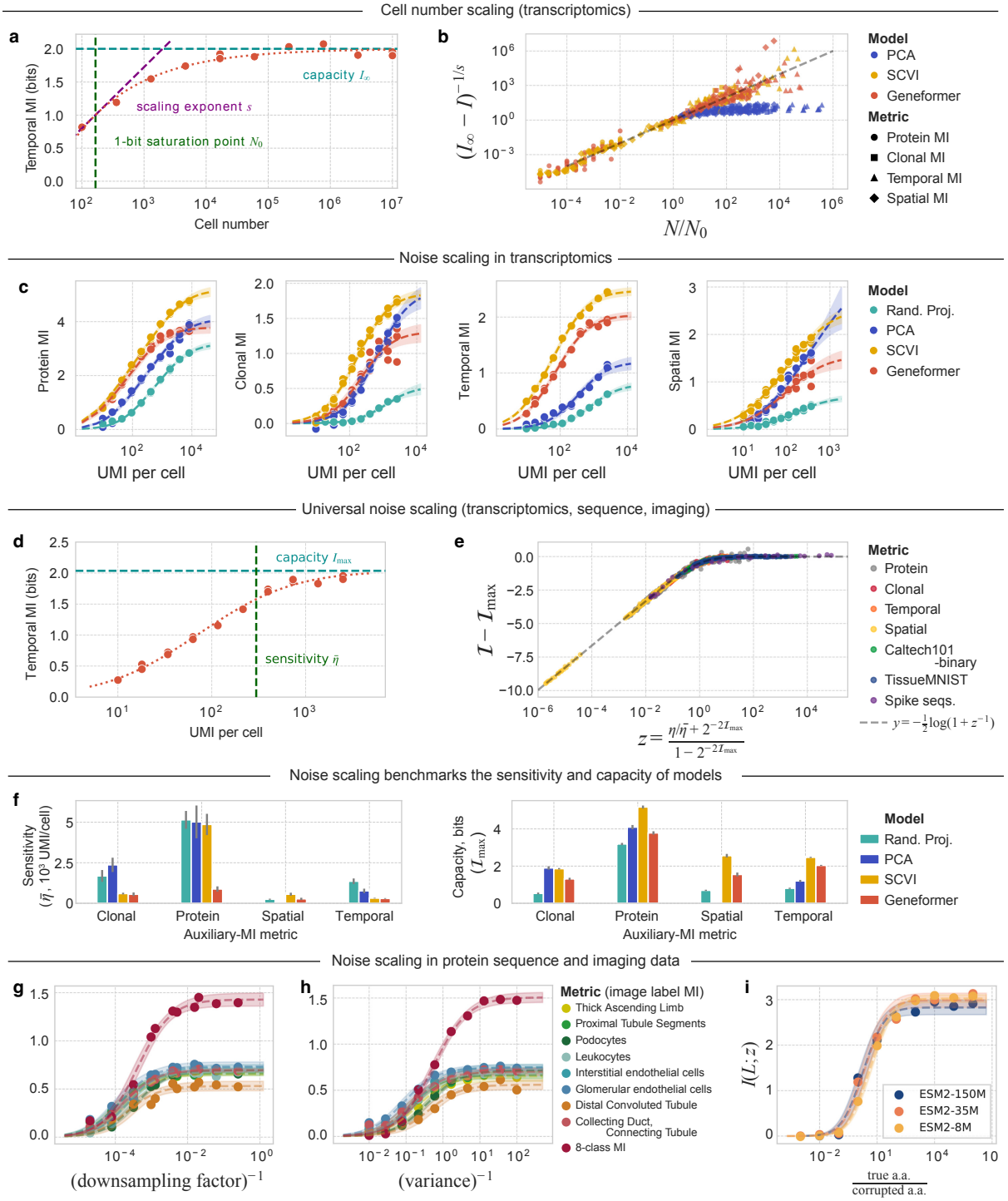


Fig. 2. Noise scaling in representation learning. (a,b) Representation model performance shows saturating power-law scaling with cell number. In (a), one example is shown of the Geneformer model applied to a 12 million cell single-cell transcriptomic data set [25]. In (b), the data is replotted with rescaled axes and extended to three model families and four datasets (see Fig. 1b), to show common scaling behavior. **(c)** Representation model performance as a function of noise, by progressive downsampling of transcripts detected per cell. Panels show different datasets. UMI=Unique Molecular Identifier. Dashed curves show fits to Eq. (1), with 2σ confidence intervals shown by the shaded regions. **(d)** Annotation of model capacity and noise-robustness parameters $I_{\max}$, $\bar{\eta}$ on a typical noise scaling plot. **(e)** Scaling plot showing collapse of 193 different noise scaling curves derived from measurements of 1,670 trained models spanning transcriptomic data, image data, and sequence data. Theory curves fit with $R^2 = 0.979 \pm 0.003$ (mean±SD). **(f)** Benchmarking models by noise-sensitivity and saturating performance. Parameters for PCA on the spatial metric are unconstrained and are omitted from the plot. **(g-i)** The underlying representation model performance plots from (e), for kidney cortex cell microscopy image Mobilenetv2 representations after progressive noising by downsampling (h) and additive noise (i); and for COVID19 protein spike sequence ESM2 representations with progressive amino acid substitutions (g). Confidence bands show 2σ interval. Image data is from [33].

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. 2c**). A subset of the curves show a 'hockey-stick' shape, indicating negligible robustness to downsampling noise, even at full transcript levels.

The smooth, sigmoidal performance curves as a function of downsampling noise suggest that a simple analytical 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. By inspection, it is evidently not a power-law. To define the noise–performance relationship, we turned to a simple model of information loss in noisy communication channels (see **Box 1**). By extending established information-theoretic results [34], we derived the following relationship between the signal-to-noise ratio of a measurement, $\eta = \text{SNR}$, and the mutual information preserved about an underlying external variable (see **Box 1**):

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

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). These parameters are annotated on an empirical curve for Geneformer and temporal information in the mouse embryogenesis dataset in **Fig. 2d**. To connect this general relationship to cellular measurements, we note that the signal-to-noise ratio introduced by molecular undersampling follows Poisson statistics, $\eta^2 = \text{CV}^{-2} \propto \text{UMI}$ [35], and $\bar{\eta}^2$ then takes on units of UMIs per cell.

In **Fig. 2c**, 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, 160 empirically measured curves from all four model families collapse onto a single universal relationship (**Fig. 2e**), 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. 2f**, 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. 2c**). 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. 2d 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 [34, 36] (proof in Appendix S.VI):

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

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 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 the same scaling relationship generalizes to other models, we examined noise–performance relationships in image representation models, and then in protein sequence models.

For image representation, we used MobileNetV2, a lightweight convolutional architecture designed for image classification [33]. We trained and evaluated this model on two different image datasets (1) a 5-class subset of the Caltech101 dataset [37], consisting of 240 × 240 pixel images with 2,707 total images and (2) a fluorescence microscopy dataset of 236,386 human kidney cortex cells annotated with one of eight cell type labels [38]. 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 [39]. 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. For both types of noise, we found that Eq. 1 accurately reproduced the observed noise–performance curves for all classification tasks (**Figs. 2g,h**, $R^2 = 0.984 \pm 0.004$ for Caltech101 and $R^2 = 0.979 \pm 0.005$ for kidney cortex models).

A similar pattern emerged in representation learning models of protein sequence. We finetuned ESM2 models (8M, 35M, and 150M parameter variants) [40] on a set of ∼63,000 SARS-CoV-2 spike protein sequences spanning the course of the pandemic [41], 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 number of months since the pandemic outbreak in January 2020 (with sequences up to April 2025). Despite the discrete and highly structured nature of protein sequences, the resulting noise–performance curves again closely match the form predicted by Eq. 1, with increasing substitution rates driving systematic and predictable declines in mutual information (**Fig. 2i**).

The shared behavior of these protein sequence and imaging models is demonstrated by collapsing the 33 additional image and sequence curves by appropriate rescaling in **Fig. 2e**. 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 Appendix S.VII).

For transcriptomic data, η^2 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 G6**. 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.

The same experimental-design questions can be asked for image and protein-sequence representation learning. For image classification (**Fig. 2g,h**), the fitted $\bar{\eta}$ values indicate, for example, the image resolution necessary for a given task. For cell type annotation, under pixelation noise, $\eta_{90} \approx 3 \cdot 10^{-3}$ corresponding to an effective resolution threshold of $\sim 15 \times 15$ pixels: images downsampled beyond this point lose more than ~ 0.15 bits of label information, corresponding to 10% information loss. Certain classes (e.g., Podocyte) exhibited a steeper performance decay (larger $\bar{\eta}$), indicating that their recognition relies on higher-resolution features. 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 simple information-theoretic model captures this relationship and recovers

the empirical scaling form observed across more than 10^3 learned 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. 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 suggests a more complete picture of representation learning in 'measurement-bound' fields such as biology, complementing the well-established roles of dataset and model size in neural scaling. Noise imposes a predictable, quantifiable constraint that can be analytically modeled and experimentally addressed with suitable trade-offs. Often, there may be a trade-off between sample number and measurement sensitivity. One can design assays that sit on or near the optimal learning curve for a given task.

Several questions remain. First, 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. Second, we have still only demonstrated noise scaling in a small number of modeling tasks. Third, even for the tasks at hand, we have only evaluated a small number of model architectures. The high cost of training model 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. 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.

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. G.G. acknowledges support from the Tayebati Postdoctoral Fellowship Program.

Supplemental Text

S.I Data preprocessing methods

In this section, we summarize key details of our data preprocessing methods. While the descriptions should be sufficient to reproduce our results, we also provide an annotated codebase in the supplemental files. First, we will describe our general data preprocessing pipeline applied to all single cell datasets, then we will describe dataset specific methods below. Finally, we will describe our image preprocessing pipeline.

S.I.1 MERFISH mouse brain dataset for spatial information probing

We use the 67,821 single cell transcriptomes measured in coronal section 1 of replicate 1 in the Vizgen mouse brain data release [28]. We remove "blank" measurements from the dataset, leaving 649 measurement dimensions. We define cell location by the center coordinate of the cell segmentation mask (which is provided in the metadata of the dataset). We construct a paired dataset of neighboring cells by randomly selecting one the 5 nearest cells as the neighbor pair for each cell in the dataset. Information probing then measures the information each cell representation contains about its neighbor pair.

S.I.2 LARRY hematopiesis dataset for clonal information probing

We pair clonally related cells as follows. In brief, we randomly pair clonally related cells between early and late timepoints. We first separate the dataset into cells profiled at early timepoints (day 2 and day 4), and final day 6 timepoint. Then, we subset the dataset for cells whose clonal barcodes appear in both early and late timepoints. Then, for each clone, we randomly select a cell from the early timepoint and pair it with a randomly selected cell from a late timepoint. Information probing then measures the information each cell representation contains about its clonally related pair.

S.I.3 CITE-seq PBMC dataset for protein and cell type probing

We use the CITE-seq PBMC dataset from [26] as distributed by `scvi-tools` [42].

S.I.4 Caltech101

We use the Caltech101 [37] as distributed by Pytorch [43]. We rescale pixel intensity values to [-1, 1], and crop images to 240×240 pixels. We then select the 5 classes with the largest number of images and subset only images from those 5 classes. This leaves a total of 2707 images. To downsample resolution by factor f, we tile the image in $240/\sqrt{f} \times 240/\sqrt{f}$ and each pixel is reassigned with the mean pixel value within its respective tile, in effect pixelating the image. To add Gaussian noise, we sample a 240×240 matrix i.i.d Gaussians with 0 mean and specified variance for each image and add it to the pixel values.

S.I.5 Kidney Cortex

Kidney tissue nuclear stain images were obtained from the MedMNIST dataset [38] at size 224×224 pixels and were preprocessed through a standardized transformation pipeline. First, pixel values normalized to the range [0, 1]. As the images contain a single channel (DAPI), channels were replicated to create 3-channel RGB-format images by repeating the single channel three times. Then, the noising process – either pixel-wise additive Gaussian noise or patch-based pooling (as in the Caltech101 experiments) to simulate pixelation – was applied.

S.I.6 SARS-CoV2 sequences

We obtain SARS-CoV-2 spike protein sequences from the GISAID database [41] until the month of 04/2025. For practicality, we use a subsample of sequences. As sequences from certain months (e.g., early 2021) are highly overrepresented, we chose not to uniformly randomly subsample the sequences. Instead, we sampled with a cap of 1000 sequences per collection month. This results in a total of 63,374 sequences across 71 months. We then randomly split this data into 75% training sequences and 25% test sequences. Then for each noise level, we randomly replace amino acids with a new amino acid uniformly sampled from the alphabet with a rate according to the noise level. We then tokenize the noised sequence using the ESM2 tokenizer distributed by huggingface [44] (`esm2_t6_8M_UR50D`).

S.II Model implementation details

Below we summarize the implementation details of the models we study in this work. While all details necessary for re-implementation are provided here, we also provide all code necessary for reproducibility in the supplemental files.

S.II.1 Random projection implementation

Random projection was implemented using `sklearn.random_projection.GaussianRandomProjection`[45] with 16 components and random state 42. The method operates on raw counts without normalization and provides 16-dimensional embeddings.

S.II.2 PCA implementation

We first further preprocess the count matrix by rescaling counts to 10^4 per cell, then log transorming, and unit-variance zero-mean standardizing each gene, and subsetting to the 750 highly variable genes. We then compute principal components using the truncated SVD method implemented in `sklearn` [45].

S.III Geneformer

Geneformer was implemented using the original repository [6] as a BERT-based transformer (`BertForMaskedLM`). Training was performed using the HuggingFace Trainer API with length-grouped batching to improve efficiency by grouping cells with similar numbers of expressed genes. To ensure consistent training across diverse dataset sizes, epochs scale dynamically as $\max(1, \lfloor 10 \times (10,000,000/\text{dataset\_size}) \rfloor)$, ensuring smaller datasets receive proportionally more training while larger datasets train for fewer epochs, thus maintaining a constant effective training budget. Early stopping prevents overfitting by monitoring validation loss, and the model uses masked language modeling with 15% token masking. Parameter count scales linearly with vocabulary size (number of genes) due to the embedding layer and language model head. Measured parameter counts across datasets range from 1.9M to 13.5M depending on the number of genes in each dataset (Table D3). Complete hyperparameters are provided in Table C1.

Supplemental Fig. C1 shows the training and validation loss curves for the Geneformer model trained on the developmental task using the full dataset of 10 million cells at full quality level. The model demonstrates stable convergence with both the training and validation losses decreasing monotonically, and the validation loss closely tracks the training loss, indicating good generalization without overfitting. Early stopping was triggered based on validation loss.

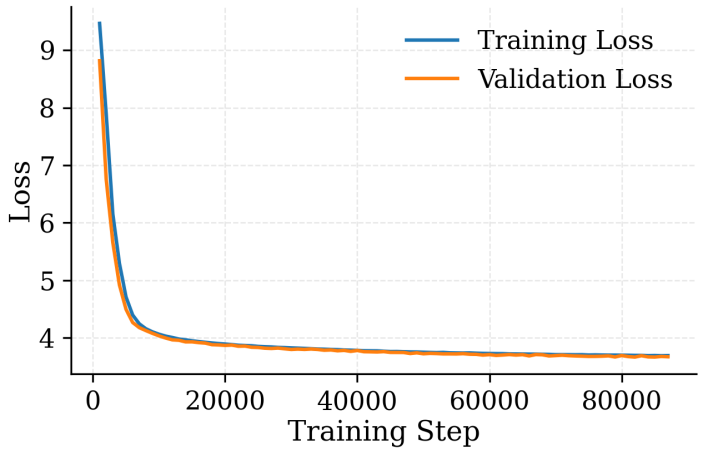


Fig. C1: Training and validation loss curves for Geneformer trained on the developmental dataset [25] with $\sim 10^7$ cells at full transcript count (no artificial noising). The model shows stable convergence.

S.IV scVI

scVI [29] was implemented using the distributed implementation from scverse [42] (`scvi.model.SCVI`) with a shallow architecture. The model uses gene-specific dispersion and zero-inflated negative binomial (ZINB) likelihood. Parameter count scales with input genes due to gene-specific parameters in

Table C1: Hyperparameters and Implementation Details for Geneformer

Parameter	Value
Model Architecture	
Number of layers	3
Hidden dimension	256
Attention heads	4
Intermediate size	512
Max sequence length	512 tokens
Activation function	ReLU
Attention dropout	0.02
Hidden dropout	0.02
Initializer range	0.02
Layer norm epsilon	1×10^{-12}
Training Hyperparameters	
Learning rate	1×10^{-3}
Optimizer	AdamW
Weight decay	0.001
Train batch size (per device)	64
Eval batch size (per device)	100
Learning Rate Schedule	
LR scheduler type	Linear
Warmup steps	5,000
Training Configuration	
Evaluation strategy	Steps
Evaluation steps	1,000
Max epochs	Dynamic (depends on dataset size)
Early Stopping	
Early stopping	Based on validation loss
Patience	5 steps

the decoder. Measured parameter counts across datasets range from 1.3M to 116.8M depending on the number of genes in each dataset (Table D3). Complete hyperparameters are provided in Table D2.

Similar to Geneformer, epochs scale dynamically to maintain consistent number of training steps across dataset sizes. KL divergence warmup is used to prevent the model from collapsing to the prior early in training. SCVI operates on raw count data without preprocessing.

S.IV.1 MobileNet for image classification

We finetune the ImageNet pretrained MobileNetv3 architecture distributed with Pytorch [43]. To adapt it to our 5-class subset of Caltech101, the final classification layer is replaced with a fully connected layer with a 5 dimensional output. To adapt it to the 8-class kidney cortex dataset, we similarly replace the classification layer with a 8 dimensional output. For Caltech101, we use a $1:1$ train-test split, and optimize a cross-entropy loss for 5 epochs. For the kidney cortex dataset, we use the train-test split provided with MedMNIST [38] (165,466 and 47,280 images respectively), and train for 30 epochs. Training and implementation details for the kidney cortex dataset experiments are provided in Table D4.

Table D2: Hyperparameters and Implementation Details for SCVI

Parameter	Value
Model Architecture	
Hidden size	512
Latent dimension	16
Number of hidden layers	1
Dropout rate	0.1
Dispersion	Gene-specific
Gene likelihood	Zero-inflated negative binomial (ZINB)
Latent distribution	Normal
Training Hyperparameters	
Learning rate	1×10^{-3}
Optimizer	Adam
Weight decay	1×10^{-6}
Adam epsilon	0.01
Batch size	512
Learning Rate Schedule	
LR scheduler	Reduce on plateau
LR scheduler metric	Validation ELBO
Minimum LR	1×10^{-6}
KL Annealing	
KL warmup epochs	1
Max KL weight	1.0
Min KL weight	0.0
Training Configuration	
Train/Val split	80% / 20%
Shuffle split	True
Max epochs	Dynamic (depends on dataset size)
Early Stopping	
Early stopping	Based on validation ELBO
Patience	5 epochs
Min delta	0.01

Table D3: Parameter counts for SCVI and Geneformer models across different datasets (millions of parameters)

Dataset	SCVI (M)	Geneformer (M)
Larry	64.9	8.3
MERFISH	1.3	1.9
PBMC	53.2	7.1
Shendure	116.8	13.5

S.IV.2 ESM2 for SARS-CoV2 spike protein sequence representations

We obtained pretrained ESM2 models of three different sizes (8M, 35M, 150M) as distributed with the huggingface `transformers` library [44]. We initialize models at the pretrained weights, and train them for a single epoch on the collection of SARS-CoV2 protein sequences curated from GISAID [41] (preprocessed as described in Appendix S.I. Training details including hyperparameters are summarized in Table D5.

Table D4: Hyperparameters and Implementation Details for Kidney Cortex Cell Type Annotation with MobileNetV3

Parameter	Value
Model Architecture	
Base model	MobileNetV3-Small
Weight initialization	ImageNet pretrained (`IMAGENET1K_V1`)
Input Channels	3 (grayscale converted by repeating)
Image Size	224×224
Training Hyperparameters	
Optimizer	Adam
Learning rate	1×10^{-3}
Finetuning objective	CrossEntropyLoss
Batch size	512
Epochs	30
Embedding Extraction	
Representation	Last layer before classifier
Information probing	
Method	Latent mutual information with 16 latent dimensions
Auxiliary signal	Class label (either one-vs.-all binary, or one-hot 8-way)

Table D5: Hyperparameters and Implementation Details for Sequence Experiments with ESM

Parameter	Value
Model Architecture	
Base models	ESM2 (8M, 35M, 150M parameters)
Finetuning objective	Masked Language Modeling (15% masking)
Max sequence length	1024
Hyperparameters	
Optimizer	AdamW
Learning rate	1×10^{-4}
Batch size	8
Epochs	1
Embedding Extraction	
Pooling method	Mean pooling of last hidden state
Information Probing	
Method	Latent mutual information with 16 latent dimensions
Auxiliary signal	Collection month (months since 01/2020)

S.V Additional details on cell number scaling behavior

As introduced in Section 4, we find that model performance scales with number of training samples (single cells) as:

$$I(N) = I_\infty - (N/N_{\text{sat}})^{-s} \tag{E1}$$

where $I_\infty, N_{\text{sat}}, s$ are parameters fit to observations. The saturating performance, I_∞, measures the capacity of a model to capture auxiliary information in the limit of infinite data. The saturation scale parameter N_{sat} quantifies the number of cells required to approach saturation, specifically, to be within 1 bit of I_∞. The scaling exponent s describes the model's sensitivity to new data (when $N \lesssim N_{\text{sat}}$). A representative fit identifying these parameters is shown in **Fig. 2a**, for the example of Geneformer

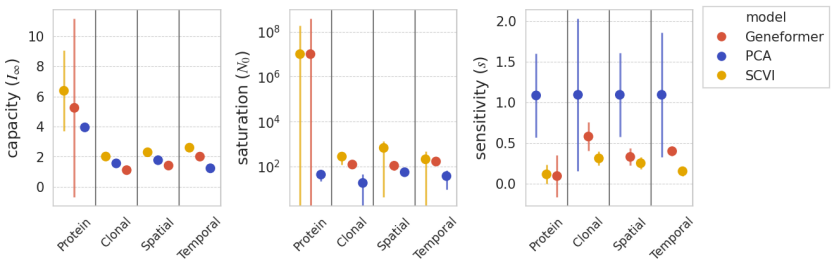


Fig. E2: Comparison of cell number scaling parameters across model families and representation quality metrics. Error bars denote 2σ confidence interval. Random projection is omitted due to lack of scaling behavior. All parameters are estimated from datasets without artificial noise.

capturing developmental time information in a mouse embryo atlas [25]. The generality of this law can be appreciated from collapsing scaling curves onto a single universal curve, **Fig. 2b** (across models and metrics) after appropriate rescaling.

The scaling of model performance with cell number reveals important differences between the models in their saturating performance, need for data, and sensitivity. The estimated parameters for each model and task are shown in **Fig. E2**

SCVI consistently showed the highest saturating performance I_∞, suggesting it learns the highest quality representations given sufficient data. The representations learned by Geneformer are less informative, with losses ranging from 0.6 to 1.1 bits across tasks compared to scVI – corresponding to approximately halving the complexity of the captured signal.

A simple PCA alone was competitive with Geneformer for some tasks: it showed lower saturating performance than Geneformer in capturing protein abundance and developmental time, but exhibited surpassed Geneformer in capturing clonal and spatial information. This suggests that Geneformer is not well-suited to these tasks.

The models also different considerably in their saturation scale. PCA learns with very little data – with N_0 in the tens of cells for all metrics. This indicates that PCA representation quality saturates almost immediately. This rapid convergence is expected of linear models [46] and suggests that PCA representations do not benefit from large dataset sizes. While Geneformer and scVI saturate much less rapidly, most models and metrics still have an $N_0 < 10^4$ cells. As this is lower than the actual dataset sizes, it suggests model performance on the evaluated metrics is largely saturated with respect to cell number.

Finally, the scaling exponent s describes the model's sensitivity to dataset size prior to saturation. PCA consistently demonstrates the largest s across all tasks. This indicates a high initial sensitivity to cell number, but this steep improvement reaches saturation quickly, as shown by N_0.

S.VI Analytical results for a toy model

Let X, Y be multivariate Gaussian random vectors representing signals distributed as follows

$$Y \sim \mathcal{N}(1, \Sigma_Y)$$

$$X = Y + U$$

where $U \sim \mathcal{N}(0, \Sigma_U)$.

Next, let Z be a random vector representing a noisy measurement of X with signal-to-noise ratio α:

$$Z = \alpha X + \mathcal{N}(0, I_n)$$

In our empirical results for transcriptomic data, X corresponds to the true transcript counts, Y corresponds to an auxiliary signal, and Z corresponds to the representation extracted from a noisy measurement of X. We are interested in how $I(Y; Z)$ scales as a function of α. We next show that in the above toy model, the relationship between α and $I(Y; Z)$ can be exactly specified.

Theorem 1 (Theorem 3.1) *For the three variable Gaussian noise model specified above,*

$$I(Y; Z) = \frac{1}{2} \log \frac{\det(\Sigma_Y + \Sigma_U + \alpha^{-2} I_n)}{\det(\Sigma_U + \alpha^{-2} I_n)} \tag{F2}$$

In the special case where $n = 1$, denoting the variances σ_Y^2, σ_U^2:

$$I(Y; Z) = \frac{1}{2} \log \frac{\alpha^2(\sigma_Y^2 + \sigma_U^2) + 1}{1 + \sigma_U^2 \alpha^2} \tag{F3}$$

Proof We will build on a basic result for Gaussian vector noise channels [34, 36] which states that for independent Gaussian random vectors X, N,

$$I(X; X + N) = \frac{1}{2} \log \frac{\det(\Sigma_x + \Sigma_N)}{\det(\Sigma_N)}$$

where Σ_X, Σ_N are the covariance matrices of X, N.

We will begin by rewriting Z in terms of Y. From definitions, we have

$$Z = \alpha X + \mathcal{N}(0, I_n)$$

$$= \alpha(Y + U) + \mathcal{N}(0, I_n)$$

$$= \alpha(Y + \mathcal{N}(0, \Sigma_U)) + \mathcal{N}(0, I_n)$$

Due to closure rules for Gaussians, we can rewrite

$$Z = \alpha Y + \mathcal{N}(0, \alpha^2 \Sigma_U + I_n)$$

Next, we observe that due to the scale invariance of mutual information [47]

$$I(Y;Z) = I(Y;\alpha^{-1}Z)$$

$$= I(Y;Y+\alpha^{-1}\mathcal{N}(0,\alpha^2\Sigma_U + I_n))$$

$$= I(Y;Y+\mathcal{N}(0,\Sigma_U + \alpha^{-2}I_n))$$

Now we can directly apply the Gaussian vector channel result:

$$I(Y;Z) = I(Y;Y+\mathcal{N}(0,\Sigma_U + \alpha^{-2}I_n))$$

$$= \frac{1}{2}\log\frac{\det(\Sigma_Y + \Sigma_U + \alpha^{-2}I_n)}{\det(\Sigma_U + \alpha^{-2}I_n)}$$

And in the special case where $n = 1$, we have that

$$I(Y;Z) = \frac{1}{2}\log\frac{\sigma_Y^2 + \sigma_U^2 + \alpha^{-2}}{\sigma_U^2 + \alpha^{-2}}$$

$$= \frac{1}{2}\log\frac{\alpha^2\sigma_Y^2 + \alpha^2\sigma_U^2 + 1}{1 + \sigma_U^2\alpha^2}$$

□

S.VII Inverting the noise scaling form

Measurement noise scaling laws can be used to determine the data quality or sample quantity necessary to obtain a model with a specified amount of information. The fit parameter $\bar{\eta}$ relates to the measurement sensitivity necessary to reach at least $\frac{1}{2}$ bits below $\mathcal{I}_{\max}$. More generally, one can invert Eq. 1 to obtain a function, $\eta(\mathcal{I})$, which estimates the acceptable SNR necessary to learn a representation with a given information content with respect to a specified external signal.

$$\mathrm{UMI}(\mathcal{I}) = \bar{u}\frac{2^{2\mathcal{I}} - 1}{2^{2\mathcal{I}_{\max}} - 2^{2\mathcal{I}}} \tag{G4}$$

Using Eqn. G4, we compute the UMI90 required to saturate auxiliary-MI across all the model-task pairs studied here, and provide these in Table G6.

Metric	Geneformer	PCA	Random Projection	SCVI	Actual UMIs
Temporal MI	2571.69 ± 201.05	5606.36 ± 993.97	8012.47 ± 1028.85	2815.46 ± 198.79	2500
Clonal MI	4212.89 ± 854.17	20190.04 ± 3764.73	7339.64 ± 1673.02	5073.33 ± 456.98	2580
Spatial MI	2212.04 ± 537.27	32107.48 ± 31922.48	1339.94 ± 177.00	5000.44 ± 830.70	367
Protein MI	8090.07 ± 1510.31	46266.58 ± 9330.52	46989.55 ± 5007.84	44958.50 ± 6071.71	8100

Table G6: η_{90} values by metric and model family, with $\pm 2\sigma$.

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] 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]

[22] 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

[23] 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]

[24] 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)

[25] 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

[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] Rasley, J., Rajbhandari, S., Ruwase, O., He, Y.: DeepSpeed: System optimizations enable training deep learning models with over 100 billion parameters. In: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. ACM, New York, NY, USA (2020). https://doi.org/10.1145/3394486.3406703

[31] 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

[32] Bengio, Y., Courville, A., Vincent, P.: Representation learning: A review and new perspectives. arXiv [cs.LG] (2012) arXiv:1206.5538 [cs.LG]

[33] 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]

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

[35] 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

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

[37] 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)

[38] Yang, J., Shi, R., Wei, D., Liu, Z., Zhao, L., Ke, B., Pfister, H., Ni, B.: MedMNIST v2 - A large-scale lightweight benchmark for 2D and 3D biomedical image classification. Scientific Data **10**(1), 41 (2023) https://doi.org/10.1038/s41597-022-01721-8

[39] 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]

[40] Rives, A., Meier, J., Sercu, T., Goyal, S., Lin, Z., Liu, J., Guo, D., Ott, M., Zitnick, C.L., Ma, J., Fergus, R.: Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. Proceedings of the National Academy of Sciences of the United States of America **118**(15) (2021) https://doi.org/10.1073/pnas.2016239118

[41] Shu, Y., McCauley, J.: GISAID: Global initiative on sharing all influenza data - from vision to reality. Euro surveillance : bulletin Europeen sur les maladies transmissibles [Euro surveillance : European communicable disease bulletin] **22**(13) (2017) https://doi.org/10.2807/1560-7917.ES.2017.22.13.30494

[42] Virshup, I., Bredikhin, D., Heumos, L., Palla, G., Sturm, G., Gayoso, A., Kats, I., Koutrouli, M., Scverse Community, Berger, B., Pe'er, D., Regev, A., Teichmann, S.A., Finotello, F., Wolf, F.A., Yosef, N., Stegle, O., Theis, F.J.: The scverse project provides a computational ecosystem for single-cell omics data analysis. Nature biotechnology **41**(5), 604–606 (2023) https://doi.org/10.1038/s41587-023-01733-8

[43] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S.: PyTorch: An Imperative Style, High-Performance Deep Learning Library. arXiv [cs.LG] (2019) arXiv:1912.01703 [cs.LG]

[44] Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T.,

Gugger, S., Drame, M., Lhoest, Q., Rush, A.: Transformers: State-of-the-art natural language processing. In: Liu, Q., Schlangen, D. (eds.) Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pp. 38–45. Association for Computational Linguistics, Stroudsburg, PA, USA (2020). https://doi.org/10.18653/v1/2020.emnlp-demos.6

[45] 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., Duchesnay, E.: Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research **12**(85), 2825–2830 (2011)

[46] Cai, T.T., Ma, Z., Wu, Y.: Sparse PCA: Optimal rates and adaptive estimation. arXiv [math.ST] (2012) https://doi.org/10.48550/arXiv.1211.1309 arXiv:1211.1309 [math.ST]

[47] Cover, T.M., Thomas, J.A.: Elements of Information Theory. Wiley & Sons, Incorporated, John, ??? (2006)