nature machine intelligence

Article

https://doi.org/10.1038/s42256-026-01332-7

A measurement noise scaling law for cellular representation learning

Received: 22 January 2026

Accepted: 30 September 2026

Published online: 26 November 2026

Check for updates

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

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 and compute budget, providing a basis for allocating data and computation. Scientific data, however, are also limited by noise arising from factors such as molecular undersampling, sequencing errors and image resolution. Here we introduce mutual information with auxiliary signals as a metric of representation model performance and study its scaling with noise. By fitting more than 2,000 representation learning 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 sensitivity and model capacity as benchmarking metrics. We show that protein-sequence representations tolerate error rates in modern sequencing data, whereas single-cell transcriptomic models remain noise limited at current measurement signal-to-noise ratio levels. Within transcriptomic cell-state representation learning models, we find that transformer-based models are more noise robust but less performant than variational autoencoders. Noise scaling metrics may support future model evaluation and experimental design.

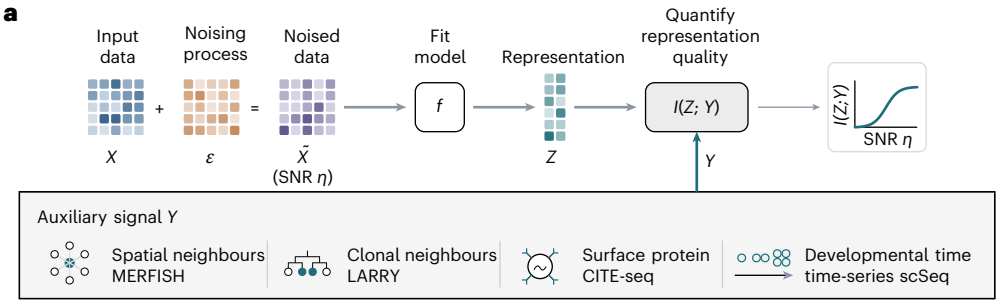
Cellular profiles obtained by single-cell RNA sequencing (scRNA-seq) and high-content imaging now span diverse tissues, developmental stages, disease states and experimental perturbations[1,2]. These large datasets (collectively, >10^8 samples) create opportunities to identify shared cellular states 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 and filter out technical nuisance factors[5]. Several deep learning approaches underlie such models so far, including transformer-based architectures, autoencoder-based architectures and models with 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 acquisition 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 per 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 in each cell. Though measurement sensitivity has increased substantially with the ongoing development of these

[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. ✉e-mail: gokulg@mit.edu; allon_klein@hms.harvard.edu

https://doi.org/10.1038/s42256-026-01332-7



b

Data modality	Tag	Auxiliary signal	*N* samples	Mean UMI per cell	Model families	Noising process
Single-cell and spatial transcriptomics	Spatial	Spatial neighbour expression (MERFISH)	7×10^4	3.7×10^2	scVI Geneformer STATE PCA Random projection	Transcript count downsampling
	Clonal	Clonal neighbour expression (LARRY)	1×10^5	2.6×10^3		
	Protein	Protein abundance (CITE-seq)	1×10^5	8.1×10^3		
	Temporal	Embryo stage	1×10^7	2.5×10^3		
Images	Tissue MNIST	Kidney cell type (fluorescence microscopy)	2.4×10^5	–	MobileNetV3	Pixel-wise additive Gaussian; pixelation
	Caltech101	Image label	2.7×10^3	–		
Protein sequences	Spike	Collection month (SARS-CoV-2 spike)	6.3×10^4	–	ESM2	Amino-acid substitution error

Fig. 1 | Workflow and datasets for studying noise in biological representation learning. a, Workflow used to evaluate representation model quality as a function of noise, with model performance quantified by auxiliary MI (see the main text). The process is repeated across many noise levels (SNR 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.

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 sequencing depth[18]. As a result, the measured transcript counts are subject to undersampling noise. Fluorescence microscopy is also prone to noise of different types including shot noise, scattering, out-of-focus light, autofluorescence, and limits in optical and camera resolution[19].

In contrast to the scaling of model performance with dataset 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, errors in training data lead to degraded performance, predicted to persist even in the limit of infinite data[20]. However, textual data used in large language model training can be curated and are subject to extensive quality control[21]. As representation learning models are being developed for diverse biological tasks, understanding how noise alters the performance of models could support model evaluation, and allow the 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. Noise scaling may also guide experimental design. Critically evaluating several models, we find that a transformer-based transcriptomics model is more robust to noise but has lower saturating performance than a variational autoencoder, illustrating how noise scaling can inform both model choice and the design of data generation efforts.

Results

A metric for representation learning model performance

In deep learning scaling analyses, it is typical to evaluate model quality directly by measuring loss on held-out validation data[10–12,14]. However, loss has two key limitations as a metric. First, it is not comparable between models with different loss functions, or even for a single model applied to data with different statistical properties (for example, different noise properties)[22]. Second, even where losses are comparable, a lower loss need not translate into more useful representations for downstream tasks[23,24].

To study the effect of noise on representation learning model performance and to avoid these issues, we introduce an alternative approach to measure the representation quality: estimating the mutual information (MI) between a model's learned representation Z and an auxiliary signal Y, which consists of some information about each sample that remains hidden until after learning is completed (Fig. 1). We use a neural-network-based MI estimator that accommodates continuous and high-dimensional auxiliary signals[25] and is consistent with well-established estimators on low-dimensional data (Supplementary Text VII). This yields a performance metric that is comparable across model types and noise levels within a given dataset. Because Z need not perfectly reconstruct Y, the approach tolerates some noise in Y at the cost of a lower maximum MI between Y and Z.

Formally, this approach generalizes linear probing[26,27], and indeed, the MI values we report later in the text agree well with goodness of fit from linear probing (average Kendall rank correlation, $\tau = 0.71$; Supplementary Text VI and Supplementary Fig. 8). We favour MI over linear probing for its general advantages as a metric: it captures nonlinear relationships and expresses representation quality in the standard units of bits. As with any probe-based approach, however, this metric is

agnostic to variation in the measurement unrelated to Y, and therefore, it is not a universal measure of representation quality, but rather, one tied to a particular downstream task.

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

1. Developmental time of ~10^7 cells profiled by scRNA-seq across mouse development, where developmental time is quantified by embryonic stage[28].
2. Surface protein abundances of ~10^5 peripheral blood mononuclear cells (PBMCs) measured by an antibody panel through CITE-seq[29].
3. Transcriptional profile of a clonally related cell in ~10^5 mouse hematopoietic stem cells measured using lineage-traced scRNA-seq[30].
4. Transcriptional profile of a spatially adjacent cell in a coronal mouse brain section of ~10^5 cells measured using MERFISH[31].

For each of these, we evaluated two linear baseline models: random projection, motivated as a representation learning baseline by the Johnson–Lindenstrauss lemma[32], and dimensionality reduction by principal component analysis (PCA). We then evaluated three modern generative models with distinct architectures: single-cell variational inference (scVI)[33], Geneformer[6] and STATE[34]. 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 the expression level. STATE is a more recent transformer-based model that uses a different token embedding framework built on protein language models. Our implementations of scVI and Geneformer comprised 1.3–116.8 million and 1.9–13.5 million parameters, respectively (depending on the gene number in each dataset), and our STATE models comprised 5.4 million trainable parameters, although we note that parameter counts across different model families are not directly comparable. We trained Geneformer across multiple GPUs using DeepSpeed[35]. Model implementation and data preprocessing details are provided in the Methods.

In all cases, to facilitate consistent comparisons, we learned representations on one data subset, and then evaluated the performance in a separate, fixed held-out subset. This approach ensures that the observed differences in MI 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 established that the auxiliary MI performance I shows expected scaling behaviour with the number of samples N (here the number of single cells) used in training. In image and language models, it has been empirically observed that performance scales as a saturating power law with respect to the sample number[14]. For scRNA-seq data, model performance has similarly been shown to saturate with the dataset size[36]. For neural-network-based models (scVI, Geneformer and STATE), we observe that I is empirically well-described by a saturating power law across auxiliary tasks, $I(N) = I_\infty - (N/N_{\text{sat}})^{-s}$, where the parameters I_∞, N_{sat} and s characterize how each model learns from new data (mean fit coefficient of determination $R^2 = 0.90$ across $n = 120$ models). The fits for these models are shown across datasets in Fig. 2a,b (Supplementary Fig. 4 shows the individually plotted curves), with parameter values and model comparisons shown in Supplementary Text II and Supplementary Fig. 3. The performance of PCA does not meaningfully fit the saturating power law as it saturates almost immediately with respect to the cell number, as also observed in prior work[36]. Thus, the MI performance metric that we will use recreates the well-studied sample size scaling: it recapitulates scaling and saturation, and recapitulates the observation that linear models have earlier saturation points than deep learning approaches.

Noise scaling for cellular representations

Although deep learning models are often thought to be strong denoisers[37], 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 of noisy data 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. Of note, no model or dataset exhibits large regimes of noise robustness. Instead, many of the measured performance curves are sigmoidal, with a short saturation span indicating only limited robustness at the transcript levels present in the original datasets before performance steadily deteriorates (Fig. 2c). A subset of the curves (for example, PCA representations for spatial information) remains unsaturated, 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 the principled allocation of sequencing depth and cell numbers in the same way that neural scaling laws guide resource decisions in large-scale machine learning, as well as clarifying how different sources of noise impact model performance. By inspection, it is evidently not a power law, unlike previous neural scaling laws[14].

To define the noise-performance relationship, we turned to a simple model of information loss in noisy communication channels (Box 1). By extending established information-theoretic results[38], we derived the following relationship between the signal-to-noise ratio (SNR) of a measurement η, and the MI preserved about an underlying external variable (Box 1 and Supplementary Text VIII):

$$\mathcal{I}(\eta) = \mathcal{I}_{\max} - \frac{1}{2}\log\left[\frac{\eta/\bar{\eta}+1}{\eta/\bar{\eta}+2^{-2\mathcal{I}_{\max}}}\right], \quad (1)$$

where $\mathcal{I}_{\max}$ is the maximum 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 SNR at which a model can gain at most 1/2 bit of information by increasing the measurement sensitivity). These parameters are annotated on an empirical curve for Geneformer and temporal information in the mouse embryogenesis dataset in Fig. 2d. For cellular measurements, the explicit SNR introduced by molecular undersampling follows Poisson statistics (Methods), $\eta = \text{CV}^{-2} \propto \text{UMI}$ (ref. 39) (where CV is the coefficient of variation), and $\bar{\eta}$ then takes on units of UMIs per cell. Equation (1) factorizes such that other independent sources of noise not captured by η manifest through the values of $\mathcal{I}_{\max}$ and $\bar{\eta}$ (Supplementary Text IX shows the expressions for combining noise sources).

In Fig. 2c, the theoretical curves defined by equation (1) (dashed lines) closely match the empirical performance curves (scatter points) across models and datasets. The noise scaling relationship holds across model architectures and across single-cell datasets spanning nearly five orders of magnitude in sample size. To illustrate the generality of the relationship, when the 200 empirically measured transcriptomic curves are rescaled by their fitted $\mathcal{I}_{\max}$ and $\bar{\eta}$ values, all models across five model architectures collapse onto a single universal relationship

https://doi.org/10.1038/s42256-026-01332-7

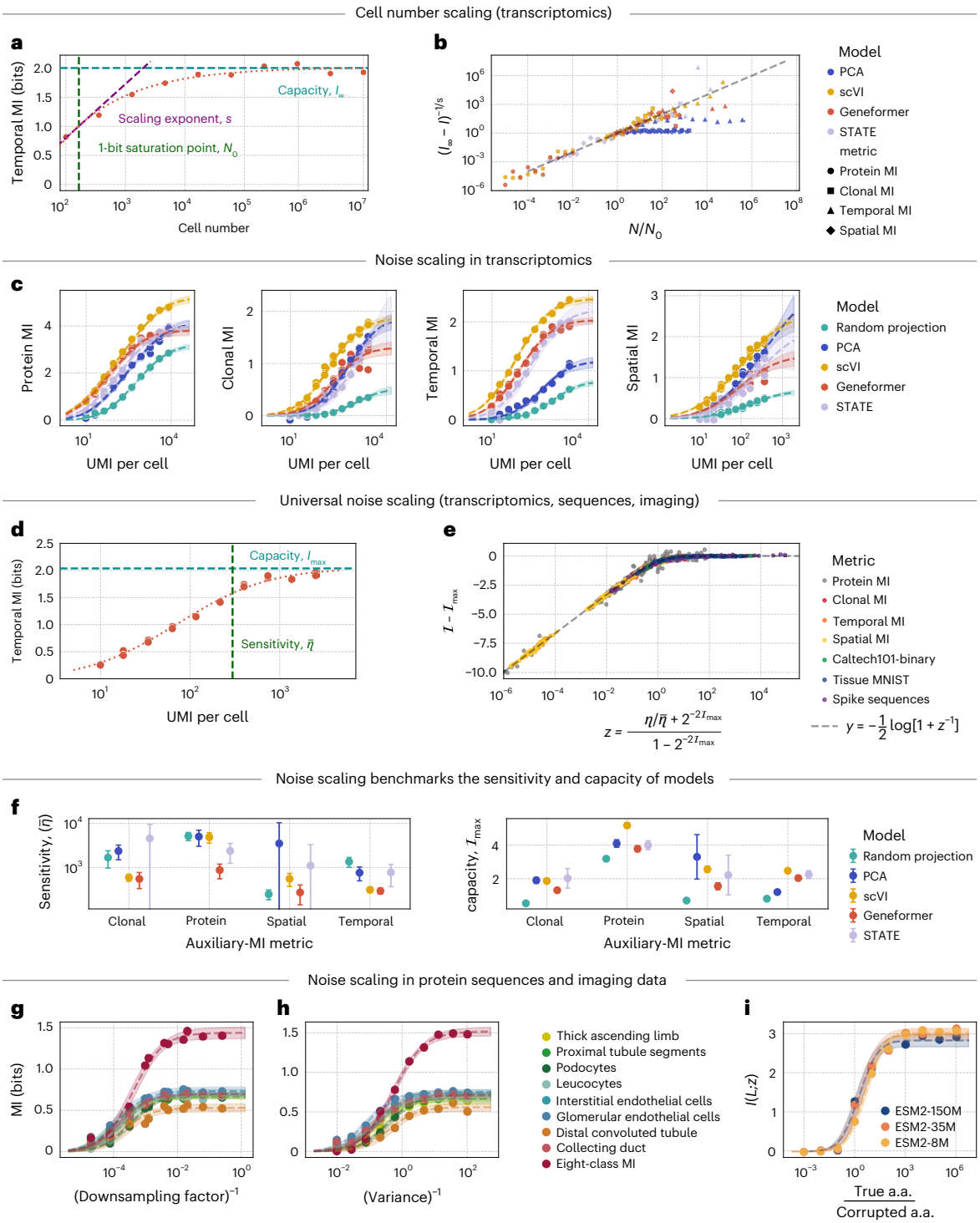


Fig. 2 | Noise scaling in representation learning. a,**b**, Representation model performance shows saturating power-law scaling with the cell number. In **a**, one example is shown of the Geneformer model applied to a 10-million-cell single-cell transcriptomic dataset[28]. In **b**, the data are replotted with rescaled axes and extended to four model families and four datasets (Fig. 1b), showing a common scaling behaviour. Random projection is omitted here owing to its lack of scaling with cell number. $n = 160$ models across four metrics and four model families. **c**, Representation model performance as a function of noise, by the progressive downsampling of transcripts detected per cell. Panels show different datasets. UMI, unique molecular identifier. Dashed curves show fits to equation (1), with 2σ confidence intervals shown by the shaded regions. The measure of the centre for bands is the least squares fit to equation (1). **d**, Annotation of model capacity and noise robustness parameters I_{max} and $\bar{\eta}$, respectively, on a typical noise scaling plot. **e**, Scaling plot showing a collapse of 233 different noise scaling curves (200, transcriptomic; 33, image and sequence) derived from measurements consisting of 2,328 fit models spanning transcriptomic data, image data and sequence data. Theory curves fit with $R^2 = 0.967 \pm 0.004$ (mean ± s.e.m.). **f**, Benchmarking models by noise sensitivity and saturating performance. Error bars denote the 2σ confidence interval around the least squares best fit parameter estimates, each fit using $n = 10$ points along a scaling curve. **g**–**i**, Underlying representation model performance plots from **e**, for kidney cortex cell microscopy image MobileNet representations after progressive noising by downsampling (**g**) and additive noise (**h**), and for COVID-19 protein spike sequence ESM2 representations with progressive amino acid substitutions (**i**). Across **g**–**i**, confidence bands show the 2σ interval from the least squares best fit curves. L denotes the image/sequence label, and z denotes the representation.

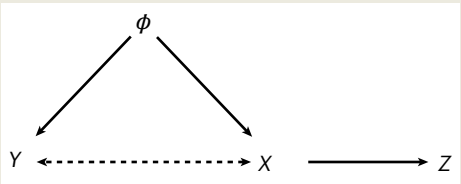
https://doi.org/10.1038/s42256-026-01332-7

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 signal and noise are modelled as Gaussian random variables. We use this Gaussian channel model as a minimal setting to derive the noise scaling form in equation (1).

Let ϕ denote the latent system state; X, the observed state (for example, transcript counts); and Y, an auxiliary signal. Let Z be a representation learned from a noisy measurement of X with SNR η.



Modelling all variables as a Gaussian with dimension n, we get

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

In this parameterization, $\mathcal{N}(0, \Sigma_U)$ represents the component of X independent of Y.

Here η corresponds to the power-ratio SNR when signals are normalized such that $\mathbb{E}[X^2] = 1$ (following the convention of ref. 38). The MI between Y and Z—the amount of auxiliary signal retained after measurement—follows from a standard expression for Gaussian vector noise channels[38,57] (Supplementary Text VIII shows the proof):

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

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

$$I(Y;Z) = \frac{1}{2}\log\left[\frac{\eta(\sigma_Y^2 + \sigma_U^2) + 1}{1 + \sigma_U^2\eta}\right]. \quad (2)$$

Two characteristic quantities govern this scaling:

$$\mathcal{I}_{\max} = \lim_{\eta\to\infty} I(Y;Z) = \frac{1}{2}\log\left[\frac{\sigma_Y^2 + \sigma_U^2}{\sigma_U^2}\right],$$

which reflects the maximum information achievable with a noiseless measurement, and $\bar{\eta} = 1/\sigma_U^2$, an effective noise scale. Substituting $\mathcal{I}_{\max}$ and $\bar{\eta}$ into $I(Y;Z)$ precisely recovers the empirical noise scaling relationship of equation (1). Despite its simplicity and derivation in a one-dimensional model, this model captures the shared shape of the performance-noise curves observed across non-Gaussian, high-dimensional data and architectures (Fig. 2e).

(Fig. 2e; Fig. 2f shows the fitting parameters for each model architecture and task). Within a single model architecture, the same noise scaling law is observed for models trained with different hyperparameter choices (for the STATE model), for models trained by fine-tuning (for the STATE model) and for different model sizes (for the Geneformer model; 0.71M to 54.8M parameters) (Supplementary Texts III–V). The noise scaling law also occurs in the presence of simulated structured, correlated measurement noise representative of technical batch effects[40], where the batch effect results in a lower effective SNR that depends on the quality of batch correction (Supplementary Text X and Supplementary Fig. 10).

Generalization of noise scaling

The scaling law (equation (1)) depends only on the SNR 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 modalities, we examined the noise-performance relationships in image representation models and then in protein sequence models. In these cases, unlike for transcriptomic data, we do not have access to the intrinsic SNR in the data and, therefore, η only reflects the scale of synthetic noise.

For image representation, we used MobileNet, a lightweight convolutional architecture designed for image classification[41]. We trained and evaluated this model on two different image datasets: (1) a five-class subset of the Caltech101 dataset[42], consisting of 240 × 240-pixel^2 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[43]. 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[44]. We then used auxiliary MI to evaluate model performance under both forms of image noise, here measuring the MI between the learned representations and the true image labels. We trained the MobileNetV3 models and computed the auxiliary MI between the predicted and true labels on held-out images, assessing performance for two tasks: classification of all class labels, as well as multiple one-versus-all problems. We introduced Gaussian noise with $\eta = 1/\sigma_N^2$, where σ_N is the noise standard deviation, whereas resolution degradation was introduced by averaging the local pixel neighbourhoods, with a downsampling factor f inducing an empirically measured SNR $\eta \approx 1/f^2$ ($R^2 = 0.995$; Supplementary Text XI and Supplementary Fig. 11). For both types of noise, we found that equation (1) accurately reproduced the observed noise-performance curves for all classification tasks (Fig. 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 fine-tuned ESM2 models (8M, 35M and 150M parameter variants)[45] on a set of ~63,000 SARS-CoV-2 spike protein sequences spanning the course of the pandemic[46], 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 the 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 equation (1), with increasing substitution rates driving systematic and predictable declines in MI (Fig. 2i).

The shared behaviour of these protein sequence and imaging models is demonstrated by collapsing the 33 additional image and sequence curves by appropriate rescaling (Fig. 2e). Together, this analysis adds to the evidence that noise in the training data is a systematic determinant of representation quality—one that can be modelled alongside sample size when characterizing learning behaviour.

Model evaluation and experimental design

The fitted noise scaling parameters from equation (1) serve two practical purposes: they give a compact basis for comparing models with one another on noise robustness and capacity, and they translate directly

into concrete guidance for experimental design when they can be estimated precisely from observations, which is often possible even without fully observing a wide SNR range (Supplementary Text XIII). We take each in turn below.

In Fig. 2f, we compare the inferred model sensitivity $\bar{\eta}$ and capacity $\mathcal{I}_{\max}$ across transcriptomic models and tasks, and find that noise robustness (low $\bar{\eta}$) and capacity do not consistently go together. Capacity favours scVI overall, though the margin depends on the task. Taken alone, noise robustness favours Geneformer, as it approaches within 0.5 bits of its asymptotic performance at fewer than 1,000 UMI per cell across all tasks. PCA is far more sensitive than any of the three, with $\bar{\eta}$ values 2.5–12.8-fold larger than Geneformer's, consistent with the limited denoising capacity of linear methods.

However, the case of Geneformer offers a clear illustration of the robustness/capacity trade-off: despite it being the most noise-robust model tested, its capacity is lower than that of scVI by 0.4–1.4 bits across all tasks—roughly half the complexity of the captured signal. This gap holds both asymptotically and already at the noise levels actually present in the datasets (Fig. 2c). This lower capacity may reflect a tendency to overfit, which limits the number of steps over which the model can be optimized: as noise increases, Geneformer's validation loss reaches its minimum progressively earlier in training (Supplementary Fig. 2). In summary, fitting $\bar{\eta}$ and $\mathcal{I}_{\max}$ offers a direct way to evaluate candidate models, and a future model may prove to have both high noise robustness and high capacity for any task.

The same fitted parameters also allow asking a complementary question: how sensitive does a measurement need to be for a model to reach a given level of performance? Although early work has used generative models such as scVI to explore this question for scRNA-seq experiment design[47], noise scaling offers a complementary tool for the purpose. Practically, inverting equation (1) yields a function $\eta(\mathcal{I})$ that predicts the minimum SNR needed to reach a desired information content with respect to a given auxiliary variable (Supplementary Text XII and equation (9) in the Supplementary Information).

For transcriptomic data, a resulting useful metric is UMI90, the sampling depth required to reach 90% of the maximum informativity. These depth requirements (UMI90) are reported for all model–task pairs in Supplementary Table 8. Several clear patterns emerge. Geneformer consistently operates above its UMI90, indicating that its performance is already near its asymptotic limit under typical sequencing depths. By contrast, UMI90 for scVI exceeds the observed UMI counts for protein abundance, spatial information and clonal information tasks. Similarly, across models, the spatial information metric has UMI90 > 10^3, far exceeding current measurement sensitivities, suggesting that spatial assays remain sensitivity limited and would benefit from optimizations that yield more UMIs per cell.

The same questions apply to image and protein-sequence representation learning. For the task of image classification (Fig. 2g,h), under pixelation noise, $\eta_{90} \approx 3 \times 10^{-3}$ corresponding to an effective resolution threshold of ~15 × 15 pixels[2]: images downsampled beyond this point lose more than ~0.15 bits of label information, corresponding to 10% information loss. Certain classes (for example, podocytes) exhibited a steeper performance decay (larger $\bar{\eta}$), indicating that their recognition relies on higher-resolution features. For protein sequences, 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 after consensus correction using multiple sequences[16]. In both cases, η is defined by the scale of noise we inject (for example, pixelation and amino acid substitution), rather than measured directly from an independent noise source in the image itself; as a result, the fitted resolution requirements should be interpreted as conservative, lower-bound estimates of true model sensitivity (Supplementary Text IX). Nonetheless, they are consistent with DNA and protein sequence models so far being able to largely ignore measurement noise.

Discussion

In conclusion, across single-cell transcriptomics, imaging and protein sequence data, we have shown that representation learning performance declines predictably with increasing noise. Model performance tolerates noise up to a threshold that depends on modality, task and model. The behaviour appears universal across the architectures and modalities tested, establishing measurement noise as a third, quantifiable scaling axis for representation learning, alongside dataset size and model size. Noise sensitivity can be measured and could inform project design, with practical consequences given the substantial resources now committed to training and evaluating large foundation models of biological data.

Two implications concern how representation learning models should be evaluated. Model loss (the conventional target of scaling analyses) lacks consistent units across architectures and data distributions, making it unsuitable for comparing noise scaling behaviour across models. MI probing, as used in this study, avoids this problem by providing a common evaluation metric. Relatedly, noise sensitivity is itself an axis of evaluation, and in some cases, it allows extrapolating to asymptotic capacity of a model. Doing so is not yet standard practice for representation learning or foundation models. As shown above (Fig. 2f), no model tested here combines the lowest noise sensitivity with the highest capacity on any task, underscoring that these properties do not co-occur naturally and could be measured and optimized jointly. Models with an inductive bias for noise invariance, as explored in ambient diffusion[48], are a promising direction for closing this gap.

Several open questions remain. Foremost among them is whether noisy measurements can still support an accurate representation of a population, even when they cannot support an accurate representation of individuals. In single-cell biology, this is relevant to a class of problems currently evaluated at single-cell resolution: predicting how a population of cells responds to a genetic or signalling perturbation[3]. If the representation target were instead a coarse-grained response (for example, at a population level), the relevant noise scale might be set by the total counts accumulated across multiple cells. In this case, increasing the cell number, not only per-cell sequencing depth, would offer a route to higher effective sensitivity for population tasks. Similarly, DNA and protein sequence representation models are fit on preprocessed consensus sequences rather than raw sequencing reads, which may have errors. Whether models operating at the population level indeed escape the noise limit in practice remains untested here.

Noise scaling also remains untested along three further axes: (1) further tasks and model architectures, (2) interaction with priors and (3) structured forms of measurement noise. We next elaborate on these. (1) We have still only demonstrated noise scaling in a small number of modelling tasks, and even for the tasks at hand, we have only evaluated a small number of model architectures. Further tests of noise scaling could require retraining foundation models, which may be expensive, but it may be that fine tuning pretrained models will provide a faithful and relatively inexpensive probe of noise tolerance. Indeed, fine tuning a pretrained STATE model reproduces the noise scaling law (Supplementary Fig. 5). (2) Here we have not considered models that incorporate external priors, or that use auxiliary information during model training. Such models may show greater resilience to noise. (3) Our observations have so far focused on measurement noise that is independent across cells and genes, whereas common errors in real data include batch effects or shifts in technology that introduce correlation structure. We found that simulated batch effects follow the same logarithmic scaling law (Supplementary Text X), but the study of real-world batch effects is left for future work.

Finally, a theoretical question concerns the origin of the scaling law itself. The model we introduce to derive the law (Box 1) is exact for scalar Gaussian channels, yet it fits high-dimensional biological data surprisingly well—why remains unclear. This simple model neither provides answers for how noise scaling and sample size scaling

interact nor how the diversity of sampled cellular states alters noise tolerance or how model size and compute budget impact the ability to tolerate measurement noise. These factors have been shown to influence performance scaling in other domains[14,49]. Empirically, the noise scaling law is preserved across different model sizes (0.71–54.8M parameters; Supplementary Fig. 6), but with different model capacities and noise sensitivities. Understanding why a simple one-dimensional model is sufficient to explain noise scaling in practice, and how these factors interact to alter $\mathcal{I}_{\max}$ and $\bar{\eta}$, would clarify the range of noise regimes and data modalities over which the scaling law can be safely extrapolated. It may also further clarify how to allocate resources to build predictive models of high-dimensional biological systems.

Even with these questions open, 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.

Methods

Data preprocessing methods

In this section, we summarize key details of our data preprocessing methods for each dataset. We also provide a link to an annotated codebase in the Code Availability.

For transcriptomic datasets, we chose auxiliary signals that are determined independently of the cell-state measurement itself—spatial position, clonal identity, surface protein abundance and developmental stage—rather than cell-type labels. Cell-type labels are themselves inferred from the cell state, typically by clustering a particular embedding, so MI with such labels is biased towards whichever representation was used to annotate them (by construction, it is maximum for that representation), conflating representation quality with similarity to the annotation pipeline.

Transcript downsampling. For the transcriptomic datasets, measurement noise was simulated by downsampling each cell's transcript counts: a number of molecules equal to a fixed fraction (the quality level) of the cell's total counts was drawn without replacement from that cell's molecules, reducing depth and preserving relative gene proportions in expectation. Each dataset was evaluated over ten quality levels spaced logarithmically from the full measured depth down to a mean of almost ten UMIs per cell, crossed with ten logarithmically spaced cell numbers.

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[31]. We remove 'blank' measurements from the dataset, leaving 649 measurement dimensions. We define the cell location by the centre coordinate of the cell segmentation mask (which is provided in the metadata of the dataset). We construct a paired dataset of neighbouring cells by randomly selecting one of the five nearest cells as the neighbour pair for each cell in the dataset. Information probing then measures the information each cell representation contains about its neighbour pair.

LARRY hematopoiesis dataset for clonal information probing. We use the in vitro differentiation dataset from ref. 30. We pair clonally related cells as follows. We first separated the dataset into cells profiled at early time points (day 2 and day 4), and the final day-6 time point. Then, we subset the dataset for cells whose clonal barcodes appear in both early and late time points. For each remaining clone, we randomly select a cell from the early time point and pair it with a randomly selected cell from a late time point. Information probing then measures the information each cell representation contains about its clonally related pair.

CITE-seq PBMC dataset for protein abundance information probing. We use the CITE-seq PBMC dataset from ref. 29, with the count matrix preprocessed and distributed by scvi-tools[50]. Information probing measures the MI between transcriptome representations and protein abundance vectors.

Caltech101. We use the Caltech101 dataset[42] as distributed by PyTorch[51]. We rescale the pixel intensity values to [−1, 1], and crop images to 240 × 240 pixels^2. We then select the five classes with the largest number of images and subset only images from those five classes. This leaves a total of 2,707 images. To downsample resolution by a factor f, we tile the image into $240/\sqrt{f} \times 240/\sqrt{f}$ 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 with i.i.d. Gaussians with 0 mean and specified variance for each image and add it to the pixel values.

Kidney cortex. Kidney tissue nuclear stain images were obtained from the MedMNIST dataset[43] at size 224 × 224 pixels^2 and were preprocessed through a standardized transformation pipeline. First, the pixel values were normalized to the range [0, 1]. As the images contain a single channel (DAPI), channels were replicated to create three-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.

SARS-CoV-2 sequences. We obtain SARS-CoV-2 spike protein sequences from the GISAID database[46] until April 2025. For practicality, we use a subsample of sequences. As sequences from certain months (for example, early 2021) are highly over-represented, we chose not to uniformly randomly subsample the sequences. Instead, we sampled with a cap of 1,000 sequences per collection month. This results in a total of 63,374 sequences across 64 months. We then randomly split these 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[52] (esm2_t6_8M_UR50D).

Model implementation details

Below, we summarize the implementation details of the models we study in this work. Further details are provided in Supplementary Texts I and XIV.

Random projection implementation. Random projection was implemented using sklearn.random_projection.GaussianRandomProjection[53] with 16 components and a fixed random seed. The method operates on gene expression without normalization (that is, without rescaling or highly variable gene selection).

PCA implementation. PCA was calculated using the dense PCA implementation from scikit-learn[53]. We first further preprocessed the count matrix by rescaling counts to 10^4 per cell, followed by log transformation. We then standardized each gene to zero mean and unit variance and subset to the 750 highly variable genes. Next, we computed principal components using the truncated singular value decomposition method, retaining the top 16 components. Note that we use highly variable genes for PCA because the full gene set is memory prohibitive at our largest sample sizes when using the scikit-learn package: a dense ~10^7 cells × >10^4 genes matrix occupies several hundred GB. Restricting the PCA input to highly variable genes makes the computation tractable.

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 whereas larger datasets train for fewer epochs, thereby maintaining a constant effective training budget. Early stopping prevents overfitting by monitoring validation loss, and the model uses masked language modelling with 15% token masking. Parameter count scales linearly with vocabulary size (number of genes) due to the embedding layer and language model head. The measured parameter counts across datasets range from 1.9M to 13.5M depending on the number of genes in each dataset (Supplementary Table 3). The complete hyperparameters are listed in Supplementary Table 1.

Supplementary Fig. 1 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 the full quality level. The model demonstrates stable convergence with both 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.

To probe whether Geneformer's lower saturating performance reflects limited capacity or active overfitting, we examined its validation loss trajectories as a function of measurement noise. As the data become noisier (fewer UMIs per cell), the validation loss reaches its minimum and begins to rise earlier in training, indicating earlier and more severe overfitting (Supplementary Fig. 2). The models evaluated in our scaling pipeline are the checkpoints with the lowest validation loss, that is, before this overfitting.

scVI. We use the scVI[33] implementation distributed by scverse[50] (`scvi.model.SCVI`) with a shallow architecture. The model uses gene-specific dispersion and zero-inflated negative binomial likelihood. Parameter count scales with the number of input genes due to gene-specific parameters in the decoder, and range from 1.3M to 116.8M (Supplementary Table 3). Complete hyperparameters are provided in Supplementary Table 2. Similar to Geneformer, epochs scale dynamically to maintain a consistent number of steps across dataset sizes. Kullback–Leibler divergence warm-up is used to prevent the model from collapsing to the prior early in training. scVI operates on raw count data without further preprocessing. We do not select for highly variable genes.

MobileNet for image classification. We fine-tune the ImageNet pretrained MobileNetV3 architecture distributed with PyTorch[51]. To adapt it to our five-class subset of Caltech101, the final classification layer is replaced with a fully connected layer having a five-dimensional output. To adapt it to the eight-class kidney cortex dataset, we similarly replace the classification layer with an eight-dimensional output. For Caltech101, we use a 1:1 train–test split, and optimize a cross-entropy loss for five epochs. For the kidney cortex dataset, we use the train–test split provided with MedMNIST[43] (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 Supplementary Table 4.

ESM2 for SARS-CoV-2 spike protein sequence representations. We obtained pretrained ESM2 models of three different sizes (8M, 35M and 150M) as distributed with the HuggingFace transformers library[52]. We initialize models at the pretrained weights, and train them for a single epoch on the collection of SARS-CoV-2 protein sequences curated from GISAID[46] (preprocessed as described earlier). Training details including hyperparameters are summarized in Supplementary Table 7.

STATE. The state embedding model SE[34] is a self-supervised transformer-based model in which each cell is represented as an ordered sequence of its most highly expressed genes (up to 2,048 tokens), with the expression magnitude incorporated via soft count binning added to gene token embeddings. A bidirectional transformer encoder processes this sequence to produce a fixed-dimensional cell embedding. During training, the model predicts expression for 1,280 target genes per cell—512 highly expressed, 512 unexpressed and 256 shared across the batch—using a dual-axis tabular energy loss: an energy distance computed gene wise within each cell and cell wise across cells for the shared gene subset.

We implemented the STATE embedding model following the published recipe. We represented gene tokens using ESM2 (ref. 45) protein language model embeddings—specifically, the same 2,560-dimensional, multispecies (*Mus musculus* + *Homo sapiens*) gene embedding table released alongside TranscriptFormer[54].

We do not use the exact SE-100M architecture, and instead choose a smaller and more tractable architecture (Supplementary Table 6). The full set of hyperparameters used for from-scratch training is reported in Supplementary Table 6.

Fine tuning an SE-100M model. In addition to training STATE from scratch, we evaluated fine tuning the publicly released pretrained SE-100M embedding model[34] on the PBMC CITE-seq dataset across noise levels. To fine-tune SE-100M, we first downloaded the pretrained model checkpoint and stripped all optimizer and scheduler states, and retained the model weights. For each noised dataset, the stripped checkpoint was then fine-tuned independently for a fixed three-epoch budget with early stopping disabled. The fine-tuning hyperparameters are summarized in Supplementary Table 7.

Relating molecular measurement sensitivity to SNR

In our analysis of transcriptomic data, we relate the SNR to the mean UMI per cell as $\eta \propto \text{UMI}$. Here we provide a statistical motivation for this relationship.

The definition of SNR that we introduce in equation (1) is the power ratio (as in refs. 38,55):

$$\text{SNR} := \frac{\mathbb{E}[S^2]}{\mathbb{E}[N^2]} \tag{3}$$

where S is a signal corrupted by additive noise N.

This definition aligns with the choice of SNR = η in the scalar case of the noise propagation model introduced in Box 1 and Supplementary Text VIII. The observed noisy signal is defined as $Z = \sqrt{\eta}X + \mathcal{N}(0,1)$; therefore, the power ratio is

$$\text{SNR}_{\text{model}} = \mathbb{E}[(\sqrt{\eta}X)^2] = \eta\mathbb{E}[X^2] \tag{4}$$

where η is equivalent to the power-ratio notion of SNR when X is normalized such that $\mathbb{E}[X^2] = 1$. This is the convention followed in ref. 38.

In the transcriptomic data, a measurement is a number of molecules detected in a cell, following $u \sim \text{Pois}(\lambda)$. We define a signal component as the mean λ and a noise component as fluctuations $u - \lambda$, which together comprise the noisy measurement. Now, we can define a power-ratio SNR for counts as

$$\text{SNR}_{\text{Pois}} = \frac{\lambda^2}{\mathbb{E}[(u-\lambda)^2]} = \text{CV}^{-2} \propto \lambda \tag{5}$$

As such, we take $\eta \propto \lambda$, where λ is the mean UMI count per cell.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

All datasets analysed in this study are previously published and publicly available. The mouse embryo development atlas is obtained from ref. 28. The CITE-seq PBMC dataset is obtained from ref. 29, and used as preprocessed and distributed by scvi-tools[50]. The LARRY hematopoiesis dataset is obtained from ref. 30. The MERFISH mouse brain data are obtained from the Vizgen mouse brain receptor map data release[31]. Caltech101[42] was used as distributed by PyTorch[51], and the kidney cortex images are the TissueMNIST subset of MedMNIST (v. 2) (ref. 43). SARS-CoV-2 spike protein sequences were obtained from GISAID[46].

Code availability

All code necessary to reproduce the results in this work is available via GitHub at https://github.com/g-kl/noise-scaling/ and archived via Zenodo at https://doi.org/10.5281/zenodo.22943757 (ref. 56).

References

1. Yao, Z. et al. A high-resolution transcriptomic and spatial atlas of cell types in the whole mouse brain. *Nature* **624**, 317–332 https://doi.org/10.1038/s41586-023-06812-z (2023).
2. Zhang, J. et al. Tahoe-100M: Mapping drug-induced molecular phenotypes at single-cell resolution. *Cell* **189**, 5945–5961.e8 (2026).
3. Bunne, C. et al. How to build the virtual cell with artificial intelligence: priorities and opportunities. *Cell* **187**, 7045–7063 (2024).
4. Wang, H., Leskovec, J. & Regev, A. Limitations of cell embedding metrics assessed using drifting islands. *Nat. Biotechnol.* **44**, 574–577 https://doi.org/10.1038/s41587-025-02702-z (2026).
5. Gunawan, I., Vafaee, F., Meijering, E. & Lock, J. G. An introduction to representation learning for single-cell data analysis. *Cell Rep. Methods* **3**, 100547 https://doi.org/10.1016/j.crmeth.2023.100547 (2023).
6. Theodoris, C. V. et al. Transfer learning enables predictions in network biology. *Nature* **618**, 616–624 https://doi.org/10.1038/s41586-023-06139-9 (2023).
7. Cui, H. et al. scGPT: toward building a foundation model for single-cell multi-omics using generative AI. *Nat. Methods* **21**, 1470–1480 https://doi.org/10.1038/s41592-024-02201-0 (2024).
8. Heimberg, G. et al. A cell atlas foundation model for scalable search of similar human cells. *Nature* **638**, 1085–1094 https://doi.org/10.1038/s41586-024-08411-y (2025).
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. *Nat. Mach. Intell.* **7**, 68–78 https://doi.org/10.1038/s42256-024-00934-3 (2025).
10. Hestness, J. et al. Deep learning scaling is predictable, empirically. Preprint at https://doi.org/10.48550/arXiv.1712.00409 (2017).
11. Rosenfeld, J. S., Rosenfeld, A., Belinkov, Y. & Shavit, N. A constructive prediction of the generalization error across scales. In *Proc. 8th International Conference on Learning Representations* 9075–9083 (ICLR, 2020).
12. Kaplan, J. et al. Scaling laws for neural language models. Preprint at https://doi.org/10.48550/arXiv.2001.08361 (2020).
13. Hoffmann, J. et al. An empirical analysis of compute-optimal large language model training. In *Proc. Advances in Neural Information Processing Systems* 35 (eds Koyejo, S. et al.) 30016–30030 https://doi.org/10.52202/068431-2176 (Curran Associates, 2022).
14. Bahri, Y., Dyer, E., Kaplan, J., Lee, J. & Sharma, U. Explaining neural scaling laws. *Proc. Natl Acad. Sci. USA* **121**, e2311878121 https://doi.org/10.1073/pnas.2311878121 (2024).
15. Chen, D. et al. Uncovering neural scaling laws in molecular representation learning. In Proc. *Advances in Neural Information Processing Systems 36* (eds Oh, A. et al.) 1452–1475 https://doi.org/10.52202/075280-0072 (Curran Associates, 2023).
16. Stoler, N. & Nekrutenko, A. Sequencing error profiles of Illumina sequencing instruments. *NAR Genom. Bioinform.* **3**, lqab019 https://doi.org/10.1093/nargab/lqab019 (2021).
17. Hagemann-Jensen, M. et al. Single-cell RNA counting at allele and isoform resolution using Smart-seq3. *Nat. Biotechnol.* **38**, 708–714 https://doi.org/10.1038/s41587-020-0497-0 (2020).
18. Svensson, V. et al. Power analysis of single-cell RNA-sequencing experiments. *Nat. Methods* **14**, 381–387 https://doi.org/10.1038/nmeth.4220 (2017).
19. Lichtman, J. W. & Conchello, J.-A. Fluorescence microscopy. *Nat. Methods* **2**, 910–919 https://doi.org/10.1038/nmeth817 (2005).
20. Bansal, Y. et al. Data scaling laws in NMT: the effect of noise and architecture. In *Proc. 39th International Conference on Machine Learning* (eds Chaudhuri, K. et al.) 1466–1482 (PMLR, 2022).
21. Li, Y. et al. Textbooks are all you need II: phi-1.5 technical report. Preprint at https://doi.org/10.48550/arXiv.2309.05463 (2023).
22. Brandfonbrener, D., Anand, N., Vyas, N., Malach, E. & Kakade, S. Loss-to-loss prediction: scaling laws for all datasets. *Trans. Mach. Learn. Res.* https://openreview.net/forum?id=1Avb4jYjLb (2025).
23. Chang, J. et al. Reading tea leaves: how humans interpret topic models. In *Proc. Advances in Neural Information Processing Systems 22* (eds Bengio, Y. et al.) 288–296 (Curran Associates, 2009).
24. Theis, L., van den Oord, A. & Bethge, M. A note on the evaluation of generative models. In *Proc. 4th International Conference on Learning Representations* (eds Bengio, Y. & LeCun, Y.) https://doi.org/10.48550/arXiv.1511.01844 (ICLR, 2016).
25. Gowri, G., Lun, X.-K., Klein, A. M. & Yin, P. Approximating mutual information of high-dimensional variables using learned representations. In *Proc. Advances in Neural Information Processing Systems 37* (eds Globerson, A. et al.) 132843–132875 https://doi.org/10.52202/079017-4223 (Curran Associates, 2024).
26. Belinkov, Y. Probing classifiers: promises, shortcomings, and advances. *Comput. Linguist.* **48**, 207–219 https://doi.org/10.1162/coli_a_00422 (2022).
27. Pimentel, T. et al. Information-theoretic probing for linguistic structure. In *Proc. 58th Annual Meeting of the Association for Computational Linguistics* (eds Jurafsky, D. et al.) 4609–4622 https://doi.org/10.18653/v1/2020.acl-main.420 (Association for Computational Linguistics, 2020).
28. Qiu, C. et al. A single-cell time-lapse of mouse prenatal development from gastrula to birth. *Nature* **626**, 1084–1093 https://doi.org/10.1038/s41586-024-07069-w (2024).
29. Hao, Y. et al. Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573–3587.e29 https://doi.org/10.1016/j.cell.2021.04.048 (2021).
30. Weinreb, C. et al. Lineage tracing on transcriptional landscapes links state to fate during differentiation. *Science* **367**, eaaw3381 https://doi.org/10.1126/science.aaw3381 (2020).
31. Vizgen. MERFISH Mouse Brain Receptor Map, Vizgen Data Release V1.0 https://info.vizgen.com/mouse-brain-map (2021).
32. Vershynin, R. *High-Dimensional Probability: An Introduction with Applications in Data Science* (Cambridge Univ. Press, 2018).
33. Lopez, R., Regier, J., Cole, M. B., Jordan, M. I. & Yosef, N. Deep generative modeling for single-cell transcriptomics. *Nat. Methods* **15**, 1053–1058 https://doi.org/10.1038/s41592-018-0229-2 (2018).
34. Adduri, A. K. et al. Predicting cellular responses to perturbation across diverse contexts with State. *Cell* **189**, 5914–5931.e20 (2026).
35. Rasley, J., Rajbhandari, S., Ruwase, O. & He, Y. DeepSpeed: system optimizations enable training deep learning models with over 100 billion parameters. In *Proc. 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (eds Gupta, R. et al.) 3505–3506 (ACM, 2020).

36. DenAdel, A. et al. Evaluating the role of pretraining dataset size and diversity on single-cell foundation model performance. *Nat. Methods* **23**, 1447–1457 (2026).
37. Bengio, Y., Courville, A. & Vincent, P. Representation learning: a review and new perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.* **35**, 1798–1828 (2013).
38. Guo, D., Shamai, S. & Verdú, S. Mutual information and minimum mean-square error in Gaussian channels. *IEEE Trans. Inf. Theory* **51**, 1261–1282 https://doi.org/10.1109/TIT.2005.844072 (2005).
39. Klein, A. M. et al. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell* **161**, 1187–1201 https://doi.org/10.1016/j.cell.2015.04.044 (2015).
40. Stuart, T. et al. Comprehensive integration of single-cell data. *Cell* **177**, 1888–1902.e21 https://doi.org/10.1016/j.cell.2019.05.031 (2019).
41. Howard, A. et al. Searching for MobileNetV3. In *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)* 1314–1324 (IEEE, 2019).
42. Fei-Fei, L., Fergus, R. & Perona, P. Learning generative visual models from few training examples: an incremental Bayesian approach tested on 101 object categories. In *Proc. 2004 Conference on Computer Vision and Pattern Recognition Workshop* 178 (IEEE, 2004).
43. Yang, J. et al. MedMNIST v2: a large-scale lightweight benchmark for 2D and 3D biomedical image classification. *Sci. Data* **10**, 41 https://doi.org/10.1038/s41597-022-01721-8 (2023).
44. Costa, G. B. P., Contato, W. A., Nazare, T. S., Neto, J. E. S. B. & Ponti, M. An empirical study on the effects of different types of noise in image classification tasks. Preprint at https://doi.org/10.48550/arXiv.1609.02781 (2016).
45. Lin, Z. et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123–1130 (2023).
46. Shu, Y. & McCauley, J. GISAID: global initiative on sharing all influenza data—from vision to reality. *Euro Surveill.* **22**, 30494 https://doi.org/10.2807/1560-7917.ES.2017.22.13.30494 (2017).
47. Svensson, V., da Veiga Beltrame, E. & Pachter, L. Quantifying the tradeoff between sequencing depth and cell number in single-cell RNA-seq. Preprint at *bioRxiv* https://doi.org/10.1101/762773 (2019).
48. Daras, G. et al. Ambient diffusion: Learning clean distributions from corrupted data. In *Proc. Advances in Neural Information Processing Systems 36* (eds Oh, A. et al.) 288–313 (Curran Associates, 2023).
49. Lin, F. et al. Data scaling laws in imitation learning for robotic manipulation. In *Proc. 13th International Conference on Learning Representations* (eds Yue, Y. et al.) 54877–54910 (ICLR, 2025).
50. Virshup, I. et al. The scverse project provides a computational ecosystem for single-cell omics data analysis. *Nat. Biotechnol.* **41**, 604–606 https://doi.org/10.1038/s41587-023-01733-8 (2023).
51. Paszke, A. et al. PyTorch: an imperative style, high-performance deep learning library. In *Proc. Advances in Neural Information Processing Systems 32* (eds Wallach, H. et al.) 8026–8037 (Curran Associates, 2019).
52. Wolf, T. et al. Transformers: state-of-the-art natural language processing. In *Proc. 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* 38–45 https://doi.org/10.18653/v1/2020.emnlp-demos.6 (Association for Computational Linguistics, 2020).
53. Pedregosa, F. et al. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
54. Pearce, J. D. et al. TranscriptFormer: a generative cell atlas across 1.5 billion years of evolution. *Science* **393**, aec8514 https://doi.org/10.1126/science.aec8514 (2026).
55. Cover, T. M. & Thomas, J. A. *Elements of Information Theory* 2nd edn https://doi.org/10.1002/047174882X (John Wiley & Sons, 2006).
56. Gowri, G. g-kl/noise-scaling: journal version. *Zenodo* https://doi.org/10.5281/zenodo.22943757 (2026).
57. Polyanskiy, Y. & Wu, Y. *Information Theory: From Coding to Learning* https://doi.org/10.1017/9781108966351 (Cambridge Univ. Press, 2025).

Acknowledgements

We thank the members of the Klein lab for their input throughout the development of the project.

Author contributions

Conceptualization: G.G., J.R. and A.K. Methodology: G.G. Software: G.G. and I.S. Investigation: G.G. and I.S. Formal analysis: G.G. and A.K. Data curation: G.G. Visualization: G.G. Writing and editing: G.G. and A.K. Supervision: A.K., P.Y., D.R. and G.G. Funding acquisition: A.K. and P.Y.

Funding

This work is supported by funding from NIH grants R01GM153805 and R01EY036381 to A.K.; NIH Pioneer Award DP1GM133052 and R01HG012926, the Molecular Robotics Initiative at the Wyss Institute to P.Y.; and by Cellular Intelligence. A.K. further acknowledges support from an Edward Mallinckrodt Jr Scholar Award. G.G. acknowledges support from the Tayebati Postdoctoral Fellowship Program.

Competing interests

A.K. and J.R. are co-founders and consultants to Cellular Intelligence (CI). D.R. is employed by CI, and I.S. was employed by CI at the time that this study was carried out. The other authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s42256-026-01332-7.

Correspondence and requests for materials should be addressed to Gokul Gowri or Allon Klein.

Peer review information *Nature Machine Intelligence* thanks Alan DenAdel, Lorin Crawford, Romain Lopez, Zhiyuan Yuan and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2026

nature portfolio

Corresponding author(s): Allon Klein, Gokul Gowri

Last updated by author(s): 09/24/2026

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☒	☐	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	N/A
Data analysis	Python 3 and its open-source ecosystem (Pytorch, numpy, pandas and others). The full list of packages and version numbers can be found in the associated Github repository.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

The MERFISH dataset analyzed in this study was obtained from Vizgen's mouse brain dataset (https://info.vizgen.com/mouse-brain-data). Lineage tracing data were obtained from the Allon Klein Lab paper-data repository (https://github.com/AllonKleinLab/paper-data/tree/master/Lineage_tracing_on_transcriptional_landscapes_links_state_to_fate_during_differentiation). The kidney cortex image dataset was obtained from Zenodo (https://

zenodo.org/records/5208230). The mouse developmental atlas data were obtained from the University of Washington Organogenesis/Mouse Gene Expression resource (https://omg.gs.washington.edu/). All other datasets used in this study are distributed with the relevant software packages and were downloaded programmatically by running the scripts provided in the code repository.

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	N/A
Reporting on race, ethnicity, or other socially relevant groupings	N/A
Population characteristics	N/A
Recruitment	N/A
Ethics oversight	N/A

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	N/A
Data exclusions	N/A
Replication	N/A
Randomization	N/A
Blinding	N/A

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Plants

Seed stocks	N/A
Novel plant genotypes	N/A
Authentication	N/A