

Measurement noise scaling laws for cellular representation learning

Gokul Gowri^{1, 2, 3*}†, Igor Sadalski^{4†}, Dan Raviv⁴, Peng Yin^{1, 2},
Jonathan Rosenfeld², Allon Klein^{1*}

¹Department of Systems Biology, Harvard Medical School, Boston, MA, USA.

²Wyss Institute for Biologically Inspired Engineering, Harvard Medical School, Boston, MA, USA.

³Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, Cambridge, MA, USA.

⁴Somite Therapeutics, 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 fit models that learn representations of cellular systems, extracting informative structure from data. In other domains, model performance improves predictably with dataset size, providing a basis for allocating data and computation. In biological data, however, performance is also limited by measurement noise arising from technical factors such as molecular undersampling or imaging variability. By learning representations of gene expression, sequence, and image data, we show that noise defines a distinct axis along which performance improves predictably across tasks. This scaling follows a simple logarithmic law that is consistent across model types, tasks, and datasets, and can be derived quantitatively from a model of noise propagation. We identify robustness to noise and saturating performance as properties that vary across models and tasks. Applied to a 12-million-cell mouse embryogenesis dataset, a large Transformer-based model shows greater noise robustness but lower saturating performance than a variational autoencoder-based model.

Keywords: scaling laws, single-cell analysis

¹ Introduction

- ² 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

7 technical nuisance factors [5]. Several deep learning approaches underlie such models to date, including
8 transformer-based architectures, autoencoder-based architectures, and contrastive losses [6–9].

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

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

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

30 Here, we recapitulate sample-size scaling in the quality of learned representations of scRNA-Seq,
31 spatial transcriptomics and image data, and we show evidence for a general and quantitative scal-
32 ing relationship between measurement noise and model performance. To show this law, we introduce
33 an information-theoretic framework for studying the scalability of representation learning models with
34 respect to changes in measurement noise and dataset size. We show that the noise-scaling law can be
35 derived by analogy to additive Gaussian noise channels, and that the resulting theoretical framework can
36 be used to guide experimental design. When applied to a 12-million-cell mouse embryogenesis dataset
37 [21], our framework suggests that a Transformer-based model is more robust to noise, but has lower
38 saturating performance than a variational autoencoder-based model.

39 **Results**

40 **A metric for representation-learning model performance**

41 In neural scaling analyses, it is typical to evaluate the quality of models directly by evaluating their loss
42 in reconstructing test data [10–12, 14]. However, model loss is not comparable between model families,
43 or even for a single model applied to data with different statistical properties [22] such as different noise
44 properties. Therefore, to study the effect of noise on representation learning model performance, we
45 introduced an alternative approach to measuring representation quality, by estimating the mutual infor-
46 mation (MI) between the representations learnt by a model and some information about each sample
47 that remains hidden until after learning is completed (Fig. 1). Formally, this approach is a generaliza-
48 tion of linear probing [23, 24], which estimates MI between a representation and a classification label.
49 Our generalization uses a neural network-based estimator of MI that accommodates high-dimensional
50 and continuous auxiliary signals [25]. This approach provides a performance metric that is comparable
51 between model types and noise levels in a given data set.

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

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

62 For each of these, we evaluated two linear baselines: random projection and dimensionality reduction
63 by principal component analysis (PCA), and we compared these to two modern generative models: single
64 cell variational inference (SCVI) [29] and Geneformer [6]. SCVI is a variational autoencoder designed to
65 compress high-dimensional gene expression information into a low-dimensional latent space. Geneformer
66 instead uses a Transformer-based language model that maps gene expression vectors to sequences by
67 ordering gene-specific tokens based on expression level. Our implementations of SCVI and Geneformer
68 have 7 million and 13 million parameters respectively, although we note that parameter counts across
69 different model families are not directly comparable. We train Geneformer at scale across multiple GPUs
70 using Deepspeed [30]. Model implementation details can be found in Appendix E and data preprocessing
71 details can be found in Appendix D .

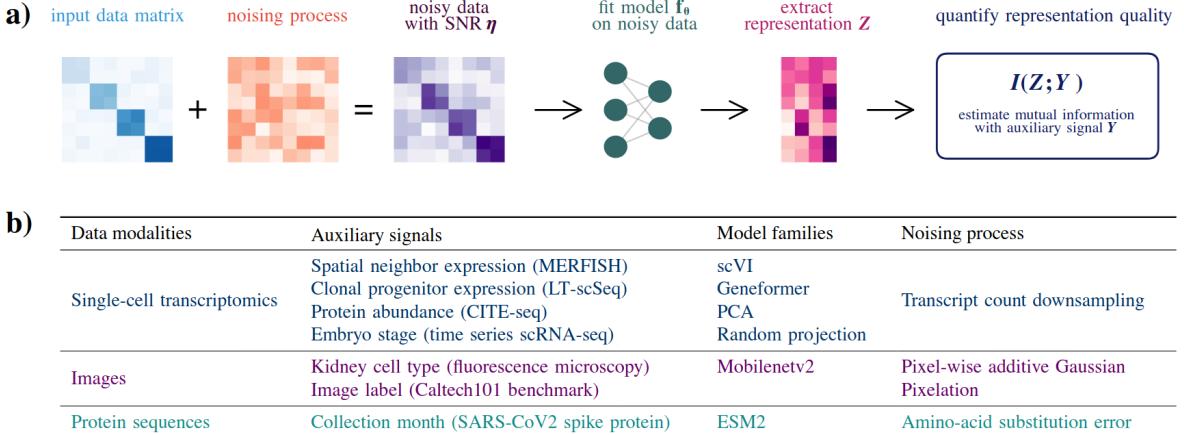


Fig. 1: Framework for understanding the role of noise in biological representation learning. (a) Workflow of a single measurement in a noise scaling curve. This process is repeated across many noise levels (resulting in various signal-to-noise ratios η). (b) Overview of settings in which we empirically demonstrate noise scaling in this work.

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

76 Cell number scaling for cellular representations

77 As a baseline for understanding the impact of noise on model learning, we first tested whether auxiliary-
 78 MI performance, I , shows expected scaling behavior with the number of samples N – here, single cells
 79 – used in training. In large language models, performance scales as a power of sample number [14]. We
 80 indeed found that I is well-described by a saturating power-law across all neural network-based models
 81 and auxiliary tasks, $I(N) = I_\infty - (N/N_{\text{sat}})^{-s}$, where the parameters $I_\infty, N_{\text{sat}}, s$ characterize how each
 82 model learns from new data (fit residual sum of squares $R^2 = 0.918 \pm 0.009$ across $n = 36$ model and task
 83 combinations). The fits are shown collectively across models and datasets in **Fig. 2ab**, with parameter
 84 values and model comparisons in **Appendix Fig. B1** and Appendix B.

85 Noise scaling for cellular representations

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

92 The dependence of model performance I on the degree of downsampling noise is shown in **Fig. 2c**.
 93 As expected, reducing the depth per cell degrades the performance of all models, across all datasets.
 94 Of note, no model or dataset exhibits large regimes of noise robustness. Instead, many of the measured
 95 performance curves are sigmoidal, indicating only limited robustness at the transcript levels present in
 96 the original datasets before performance steadily deteriorates (**Fig. 2c**). A subset of the curves show a
 97 ‘hockey-stick’ shape, indicating negligible robustness to downsampling noise, even at full transcript levels.

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

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

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

111 where \mathcal{I}_{\max} is the maximal information that can be extracted from a noiseless measurement at a
 112 fixed sample size, and $\bar{\eta}$ serves as a measure of noise robustness (specifically, the signal-to-noise ratio
 113 at which a model can gain at most 1/2 a bit of information by increasing measurement sensitivity).
 114 These parameters are annotated on an empirical curve for Geneformer and temporal information in the
 115 mouse embryogenesis dataset in **Fig. 2d**. To connect this general relationship to cellular measurements,
 116 we note that the signal-to-noise ratio introduced by molecular undersampling follows Poisson statistics,
 117 $\eta^2 = \text{CV}^{-2} \propto \text{UMI}$ [32], and $\bar{\eta}$ then takes on units of UMIs per cell.

118 In **Fig. 2c**, the theoretical curves defined by Eq. (1) (dotted lines) closely match the empirical
 119 performance curves (scatter points) across models and datasets. Strikingly, the noise-scaling relationship
 120 holds across model architectures and across single-cell datasets spanning nearly five orders of magnitude
 121 in sample size. When rescaled by their fitted \mathcal{I}_{\max} and $\bar{\eta}$ values, 160 empirically measured curves from
 122 all four model families collapse onto a single universal relationship (**Fig. 2e**), indicating that a shared
 123 principle governs how measurement noise in transcriptomic data limits representation learning.

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

128 In **Fig. 2f**, we compare inferred $\bar{\eta}$ and \mathcal{I}_{\max} values across models. Geneformer consistently shows the
129 greatest robustness to noise: across all tasks, it approaches within 0.5 bits of its asymptotic performance
130 at fewer than 1,000 UMI per cell. scVI displays similarly low noise sensitivity for three of the four tasks,
131 but in the protein-abundance task it becomes noise-sensitized at $\sim 4,000$ UMI per cell. PCA, by con-
132 trast, shows far greater sensitivity to noise, with $\bar{\eta}$ values 2.5–12.8-fold larger than those of Geneformer,
133 consistent with the limited denoising capacity of linear methods.

134 Despite its robustness to noise, Geneformer is not a strong model in terms of its capacity. Across all
135 tasks, its capacity, \mathcal{I}_{\max} , is lower than those of scVI by 0.4–1.4 bits – corresponding to approximately
136 halving the complexity of the captured signal. This difference in performance is not only in its asymptotic
137 capacity, but also at the noise level present in the datasets (**Fig. 2c**). Thus, scVI ultimately extracts
138 more auxiliary information in the limit of low noise. It is possible that other models may simultaneously
139 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 :

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

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

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

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

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

Two characteristic quantities govern this scaling:

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

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.

140

141 **Generalization of noise scaling**

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

147 For image representation, we used MobileNetV2, a lightweight convolutional architecture designed for
 148 image classification [34]. We trained and evaluated this model on two different image datasets (1) a 5-class
 149 subset of the Caltech101 dataset [35], consisting of 240×240 pixel images with 2,707 total images and
 150 (2) a fluorescence microscopy dataset of 236,386 human kidney cortex cells annotated with one of eight
 151 cell type labels [36]. Images were perturbed with two distinct forms of degradation: additive Gaussian
 152 noise and reduced spatial resolution. Pixel-wise Gaussian noise is common in imaging measurements [37].
 153 We then used auxiliary-MI to evaluate model performance under both forms of image noise. As with

154 the transcriptomic models, we quantified performance using auxiliary-MI, here measuring the mutual
155 information between the learned representations and the true image labels. We trained the MobileNetV2
156 models and computed the auxiliary-MI between predicted and true labels on held-out images, assessing
157 performance for two tasks: classification of all class labels, as well as multiple one-vs-all problems. We
158 introduced Gaussian noise with $\eta = 1/\sigma_N^2$, where σ_N is the noise standard deviation, while resolution
159 degradation was introduced by averaging local pixel neighborhoods, with $\eta = 1/f$ for downsampling
160 factor f . Downsampling introduces a noise $\text{SNR} \propto f$. For both types of noise, we found that Eq. 1
161 accurately reproduced the observed noise–performance curves for all classification tasks (Fig. 2g, $R^2 =$
162 0.984 ± 0.004 for Caltech101 and $R^2 = 0.979 \pm 0.005$ for kidney cortex models).

163 A similar pattern emerged in representation learning models of protein sequence. We finetuned ESM2
164 models (8M, 35M, and 150M parameter variants) [38] on a set of $\sim 63,000$ SARS-CoV-2 spike protein
165 sequences spanning the course of the pandemic [39], after introducing controlled levels of amino-acid
166 substitution to simulate increasingly corrupted measurements. We then quantified auxiliary-MI between
167 the learned representations and the collection date of each sequence, measured as number of months
168 since the pandemic outbreak in January 2020 (with sequences up to April 2025). Despite the discrete
169 and highly structured nature of protein sequences, the resulting noise–performance curves again closely
170 match the form predicted by Eq. 1, with increasing substitution rates driving systematic and predictable
171 declines in mutual information (Fig. 2g).

172 The shared behavior of these models is demonstrated by collapsing the 33 additional image and
173 sequence curves by appropriate rescaling in Fig. 2e. Together, this analysis adds to the evidence that
174 noise in training data is a systematic determinant of representation quality – one that can be modeled
175 alongside sample size when characterizing learning behavior.

176 Noise scaling and experimental design

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

183 For transcriptomic data, η is proportional to the total UMIs per cell, enabling an estimate of the
184 sequencing depth needed for a representation to reach, for example, 90% of its maximum informativity.
185 These depth requirements (UMI90) are reported for all model–task pairs in Table C1. Several clear
186 patterns emerge. Geneformer consistently operates below its UMI90 on all datasets examined, indicating
187 that its performance is already near its asymptotic limit under typical sequencing depths. In contrast,

188 UMI90 for scVI exceeds the observed UMI counts for protein abundance, spatial information, and clonal
189 information tasks—suggesting that these tasks remain sensitivity-limited and would benefit from deeper
190 sequencing. These examples illustrate how noise-scaling relationships can guide the choice of models and
191 allocation of sequencing depth across tasks with different intrinsic difficulty.

192 To assess whether these experimental-design conclusions extend beyond transcriptomics, we applied
193 the same analysis to image and protein-sequence representation learning. For image classification
194 (**Fig. 2g**), the fitted $\bar{\eta}$ values provided interpretable design guidance. In the 8-way kidney cell type
195 annotation task under pixelation noise, $\eta_{90} \approx 3 \cdot 10^{-3}$ corresponding to an effective resolution thresh-
196 old of $\sim 15 \times 15$ pixels: images downsampled beyond this point lose more than ~ 0.15 bits of label
197 information, corresponding to 10% information loss. Certain classes (e.g., Podocyte) exhibited a steeper
198 performance decay (larger $\bar{\eta}$), indicating that their recognition relies on higher-resolution features. For
199 protein sequence, auxiliary-MI remained stable up to substitution rates of approximately 1 in 1,000
200 amino acids—a noise level well above what is typical in modern sequencing. This is consistent with DNA
201 and protein sequence models to date being able to largely ignore measurement noise. These results par-
202 allel the transcriptomic findings and show that noise tolerance thresholds can be used to rationally plan
203 training of image models as well.

204 Discussion

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

214 This work provides practical guidance for designing and evaluating biological representation models.
215 The fitted scaling parameters $\bar{\eta}$ and \mathcal{I}_{\max} jointly characterize model behavior: $\bar{\eta}$ reflects noise robustness,
216 while \mathcal{I}_{\max} represents the maximal task-relevant information that a model can encode. Nonlinear models
217 such as Geneformer and scVI exhibit substantially greater robustness to measurement noise than PCA,
218 consistent with the expectation that nonlinear architectures more effectively denoise sparse molecular
219 measurements. However, robustness alone is insufficient. Geneformer, despite its stability under noise,
220 often attains a relatively low \mathcal{I}_{\max} , capturing less auxiliary information than scVI and, for certain tasks,

even linear baselines. These results emphasize that noise robustness and representational capacity must be jointly optimized in model design.

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

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

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

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

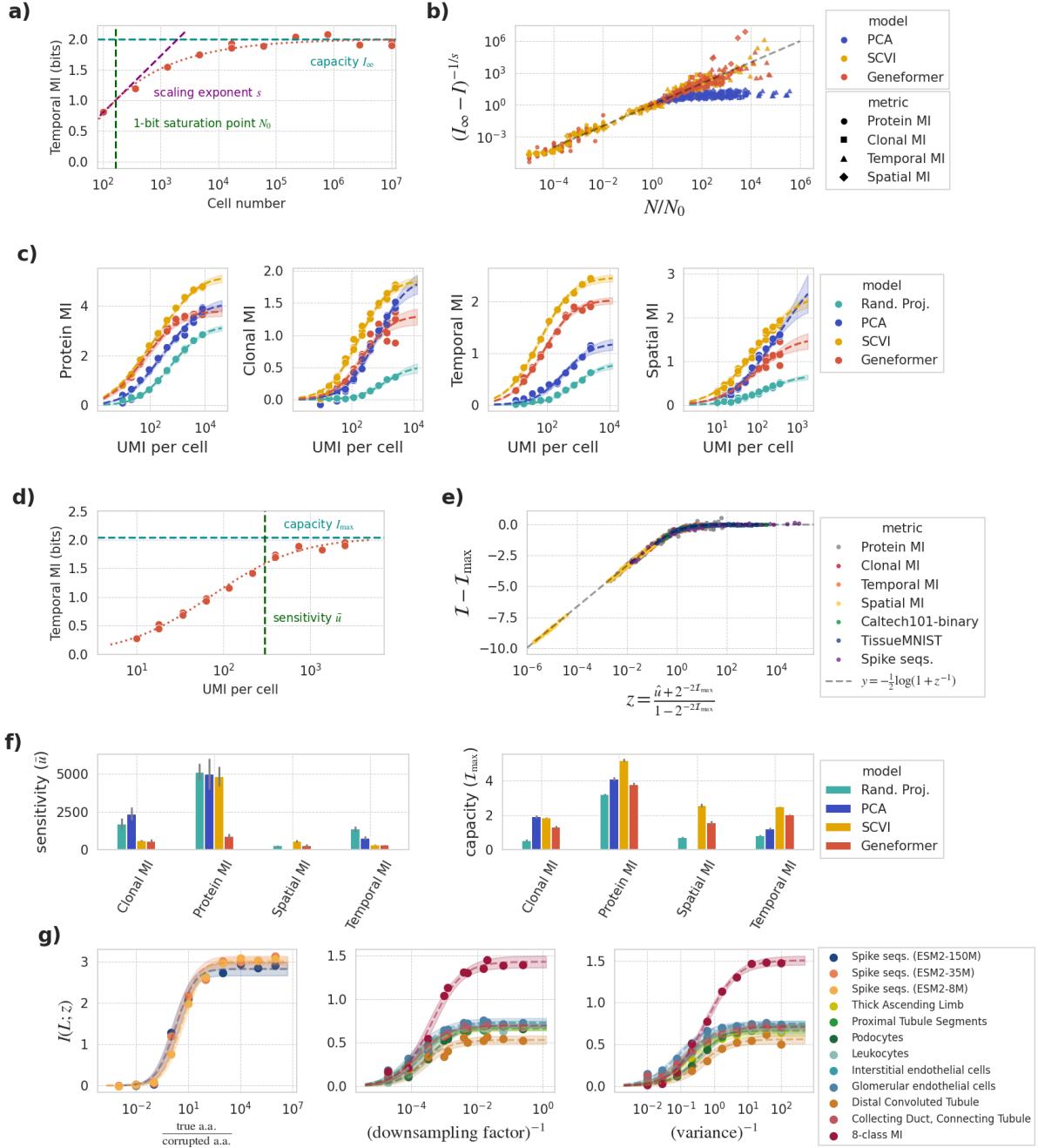


Fig. 2 (previous page): **Scaling laws for cellular representation learning.** (a) Geneformer representation quality, measured by information about developmental time, as a function of number of training data points drawn from a mouse embryo development atlas [21]. Theory curve is shown with a dashed line. Cell number scaling parameters I_∞ , s , N_0 are annotated on the theory curve. (b) Scaling collapse of 54 different cell number scaling curves across three model families and four datasets. Datasets with transcript counts downsampled by more than one order of magnitude are omitted. (c) Observations and noise scaling law fits for representation quality as a function of molecules detected per cell. Confidence bands show 2σ interval. (d) Geneformer representation quality, measured by information about developmental time, as a function of number of transcripts captured per cell. Theory curve is shown with a dashed line. Noise scaling parameters I_{\max} , \bar{u} are annotated on the theory curve. (e) Scaling collapse of 193 different noise scaling curves derived from measurements of 1670 distinct trained models spanning transcriptomic data (1600 models), image data (40 models), and sequence data (30 models). Theory curevs fit with average $R^2 = 0.979 \pm 0.003$. (f) Comparison of noise scaling parameters across model families with fixed-size (no cell subsampling) datasets. Parameters for PCA on the spatial metric are underconstrained and omitted from the plot. (g) Observations and noise scaling law fits for kidney cortex cell type information of Mobilenetv2 models [34] and temporal information of ESM2 models finetuned on SARS-CoV2 spike protein sequences. Confidence bands show 2σ interval.

251 **Acknowledgements.** This work is supported by funding from NIH Pioneer Award DP1GM133052,
252 R01HG012926 to P.Y., and Molecular Robotics Initiative at the Wyss Institute. A.M.K. acknowledges
253 support of an Edward Mallinckrodt Jr. Scholar Award. G.G. acknowledges support from the Tayebati
254 Postdoctoral Fellowship Program.

255 **Appendix A Analytical results for a toy model**

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

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

$$X = Y + U$$

256 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 :

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

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

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

$$I(Y; Z) = \frac{1}{2} \log \frac{\det(\Sigma_Y + \Sigma_U + \alpha^{-2} I_n)}{\det(\Sigma_U + \alpha^{-2} I_n)} \quad (\text{A1})$$

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

$$I(Y; Z) = \frac{1}{2} \log \frac{\alpha^2(\sigma_Y^2 + \sigma_U^2) + 1}{1 + \sigma_U^2 \alpha^2} \quad (\text{A2})$$

Proof We will build on a basic result for Gaussian vector noise channels [31, 33] which states that for independent Gaussian random vectors X, N ,

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

263 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

$$\begin{aligned} Z &= \alpha X + \mathcal{N}(0, I_n) \\ &= \alpha(Y + U) + \mathcal{N}(0, I_n) \\ &= \alpha(Y + \mathcal{N}(0, \Sigma_U)) + \mathcal{N}(0, I_n) \end{aligned}$$

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 [40]

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

Now we can directly apply the Gaussian vector channel result:

$$\begin{aligned} I(Y; Z) &= I(Y; Y + \mathcal{N}(0, \Sigma_U + \alpha^{-2}I_n)) \\ &= \frac{1}{2} \log \frac{\det(\Sigma_Y + \Sigma_U + \alpha^{-2}I_n)}{\det(\Sigma_U + \alpha^{-2}I_n)} \end{aligned}$$

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

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

264

□

265 Appendix B Additional details on cell number scaling behavior

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

$$I(N) = I_\infty - (N/N_{\text{sat}})^{-s} \quad (\text{B3})$$

268 where $I_\infty, N_{\text{sat}}, s$ are parameters fit to observations. The saturating performance, I_∞ , measures the
 269 capacity of a model to capture auxiliary information in the limit of infinite data. The saturation scale
 270 parameter N_{sat} quantifies the number of cells required to approach saturation, specifically, to be within
 271 1 bit of I_∞ . The scaling exponent s describes the model's sensitivity to new data (when $N \lesssim N_{\text{sat}}$).
 272 Finally, baseline performance, I_0 , represents the performance of an untrained model and does not directly
 273 influence the scalability of the model. A representative fit identifying these parameters is shown in Fig.
 274 **2a**, for the example of Geneformer capturing developmental time information in a mouse embryo atlas
 275 [21]. The generality of this law can be appreciated from plotting performance of all 54 distinct scaling
 276 experiments on a single rescaled plot, **Fig. 2b** (across models, metrics, and dataset noise levels) after
 277 appropriate rescaling.

278 The scaling of model performance with cell number reveals important differences between the models
 279 in their saturating performance, need for data, and sensitivity. The estimated parameters for each model
 280 and task are shown in **Fig. B1**

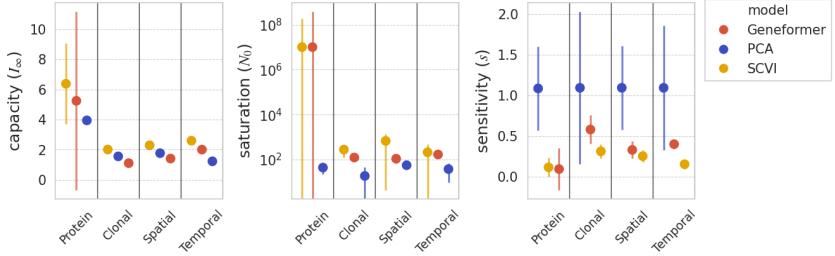


Fig. B1: Comparison of cell number scaling parameters across model families and representation quality metrics. Error bars denote 2σ confidence interval. Random projection is omitted due to lack of scaling behavior. All parameters are estimated from datasets without artificial noise.

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

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

289 The models also differ considerably in their saturation scale. PCA learns with very little data –
 290 with N_0 in the tens of cells for all metrics. This indicates that PCA representation quality saturates
 291 almost immediately. This rapid convergence is expected of linear models [41] and suggests that PCA
 292 representations do not benefit from large dataset sizes. While Geneformer and scVI saturate much less
 293 rapidly, most models and metrics still have an $N_0 < 10^4$ cells. As this is lower than the actual dataset
 294 sizes, it suggests model performance on the evaluated metrics is largely saturated with respect to cell
 295 number.

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

299 Appendix C Inverting the noise scaling form

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

304 information content with respect to a specified external signal.

$$\text{UMI}(\mathcal{I}) = \bar{u} \frac{2^{2\mathcal{I}} - 1}{2^{2\mathcal{I}_{\max}} - 2^{2\mathcal{I}}} \quad (\text{C4})$$

305 Using Eqn. C4, we compute the UMI₉₀ required to saturate auxiliary-MI across all the model-task
306 pairs studied here, and provide these in Table C1.

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 C1: η_{90} values by metric and model family, with $\pm 2\sigma$.

307 Appendix D Data preprocessing methods

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

313 D.1 MERFISH mouse brain dataset for spatial information probing

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

320 D.2 LARRY hematopiesis dataset for clonal information probing

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

327 **D.3 CITE-seq PBMC dataset for protein and cell type probing**

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

329 **D.4 Caltech101**

330 We use the Caltech101 [35] as distributed by Pytorch [43]. We rescale pixel intensity values to $[1, 1]$,
331 and crop images to 240×240 pixels. We then select the 5 classes with the largest number of images and
332 subset only images from those 5 classes. This leaves a total of 2707 images. To downsample resolution by
333 factor f , we tile the image in $240/\sqrt{f} \times 240/\sqrt{f}$ and each pixel is reassigned with the mean pixel value
334 within its respective tile, in effect pixelating the image. To add Gaussian noise, we sample a 240×240
335 matrix i.i.d Gaussians with 0 mean and specified variance for each image and add it to the pixel values.

336 **D.5 Kidney Cortex**

337 Kidney tissue nuclear stain images were obtained from the MedMNIST dataset [36] at size 224×224 pixels
338 and were preprocessed through a standardized transformation pipeline. First, pixel values normalized
339 to the range $[0, 1]$. As the images contain a single channel (DAPI), channels were replicated to create
340 3-channel RGB-format images by repeating the single channel three times. Then, the noising process –
341 either pixel-wise additive Gaussian noise or patch-based pooling (as in the Caltech101 experiments) to
342 simulate pixelation – was applied.

343 **D.6 SARS-CoV2 sequences**

344 We obtain SARS-CoV-2 spike protein sequences from the GISAID database [39] until the month of
345 04/2025. For practicality, we use a subsample of sequences. As sequences from certain months (e.g., early
346 2021) are highly overrepresented, we chose not to uniformly randomly subsample the sequences. Instead,
347 we sampled with a cap of 1000 sequences per collection month. This results in a total of 63,374 sequences
348 across 71 months. We then randomly split this data into 75% training sequences and 25% test sequences.
349 Then for each noise level, we randomly replace amino acids with a new amino acid uniformly sampled
350 from the alphabet with a rate according to the noise level. We then tokenize the noised sequence using
351 the ESM2 tokenizer distributed by huggingface [44] (`esm2_t6_8M_UR50D`).

352 **Appendix E Model implementation details**

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

356 **E.1 Random projection implementation**

357 We first further preprocess the count matrix by rescaling counts to $1e4$ per cell, then log transforming,
358 and unit-variance zero-mean standardizing each gene, and subsetting to the 750 highly variable genes.
359 We then randomly project our preprocessed count matrix of 750 genes to 16 dimensions with a 750×16
360 matrix populated with i.i.d standard unit variance Gaussian elements.

361 **E.2 PCA implementation**

362 We follow the same preprocessing pipeline as used for random projection, then compute principal
363 components using the randomized SVD method implemented in `sklearn` [45].

364 **E.3 VAE implementation**

365 We use the `scvi` software package [42]. For each encoder and decoder, we use a single hidden layer with
366 512 units. We use a 16-dimensional latent space. We train with an early stopping patience of 5 epochs.
367 Further training and implementation details are given in Table E2.

Table E2: Hyperparameters and Implementation Details for SCVI

Parameter	Value
<i>Model Architecture</i>	
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	Isotropic normal
<i>Training Hyperparameters</i>	
Learning rate	1×10^{-3}
Optimizer	Adam
Weight decay	1×10^{-6}
Batch size	512
<i>KL Annealing</i>	
KL warmup epochs	1
Max KL weight	1.0
Min KL weight	0.0
<i>Training configuration</i>	
Train/Val split	80% / 20%
Shuffle split	True
Max epochs	depends on dataset size
<i>Early Stopping</i>	
Early stopping	Based on val. loss
Patience	5 epochs
Min delta	0.01

368 **E.4 MobileNetv3 for image classification**

369 We finetune the ImageNet pretrained MobileNetv3 architecture distributed with Pytorch [43]. To adapt
370 it to our 5-class subset of Caltech101, the final classification layer is replaced with a fully connected
371 layer with a 5 dimensional output. To adapt it to the 8-class kidney cortex dataset, we similarly replace
372 the classification layer with a 8 dimensional output. For Caltech101, we use a 1 : 1 train-test split,
373 and optimize a cross-entropy loss for 5 epochs. For the kidney cortex dataset, we use the train-test
374 split provided with MedMNIST [36] (165,466 and 47,280 images respectively), and train for 30 epochs.
375 Training and implementation details for the kidney cortex dataset experiments are provided in Table E3.

Table E3: Hyperparameters and Implementation Details for Kidney Cortex Cell Type Annotation with MobileNetV3

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

376 **E.5 ESM2 for SARS-CoV2 spike protein sequence representations**

377 We obtained pretrained ESM2 models of three different sizes (8M, 35M, 150M) as distributed with the
378 huggingface `transformers` library [44]. We initialize models at the pretrained weights, and train them for
379 a single epoch on the collection of SARS-CoV2 protein sequences curated from GISAID [39] (preprocessed
380 as described in Appendix D. Training details including hyperparameters are summarized in Table E4.

Table E4: Hyperparameters and Implementation Details for Sequence Experiments with ESM

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

381 **References**

- 382 [1] Yao, Z., Velthoven, C.T.J., Kunst, M., Zhang, M., McMillen, D., Lee, C., Jung, W., Goldy, J., Abdel-
383 hak, A., Aitken, M., Baker, K., Baker, P., Barkan, E., Bertagnolli, D., Bhandiwad, A., Bielstein, C.,
384 Bishwakarma, P., Campos, J., Carey, D., Casper, T., Chakka, A.B., Chakrabarty, R., Chavan, S.,
385 Chen, M., Clark, M., Close, J., Crichton, K., Daniel, S., DiValentin, P., Dolbeare, T., Ellingwood,
386 L., Fiabane, E., Fliss, T., Gee, J., Gerstenberger, J., Glandon, A., Gloe, J., Gould, J., Gray, J., Guil-
387 ford, N., Guzman, J., Hirschstein, D., Ho, W., Hooper, M., Huang, M., Hupp, M., Jin, K., Kroll,
388 M., Lathia, K., Leon, A., Li, S., Long, B., Madigan, Z., Malloy, J., Malone, J., Maltzer, Z., Martin,
389 N., McCue, R., McGinty, R., Mei, N., Melchor, J., Meyerdierks, E., Mollenkopf, T., Moonsman, S.,
390 Nguyen, T.N., Otto, S., Pham, T., Rimorin, C., Ruiz, A., Sanchez, R., Sawyer, L., Shapovalova,
391 N., Shepard, N., Slaughterbeck, C., Sulc, J., Tieu, M., Torkelson, A., Tung, H., Valera Cuevas, N.,
392 Vance, S., Wadhwan, K., Ward, K., Levi, B., Farrell, C., Young, R., Staats, B., Wang, M.-Q.M.,
393 Thompson, C.L., Mufti, S., Pagan, C.M., Kruse, L., Dee, N., Sunkin, S.M., Esposito, L., Hawrylycz,
394 M.J., Waters, J., Ng, L., Smith, K., Tasic, B., Zhuang, X., Zeng, H.: A high-resolution transcrip-
395 tomic and spatial atlas of cell types in the whole mouse brain. *Nature* **624**(7991), 317–332 (2023)
396 <https://doi.org/10.1038/s41586-023-06812-z>
- 397 [2] Zhang, J., Ubas, A.A., Borja, R., Svensson, V., Thomas, N., Thakar, N., Lai, I., Winters, A., Khan,
398 U., Jones, M.G., Tran, V., Pangallo, J., Papalex, E., Sapre, A., Nguyen, H., Sanderson, O., Nigos,
399 M., Kaplan, O., Schroeder, S., Hariadi, B., Marrujo, S., Salvino, C.C.A., Gallareta Olivares, G.,
400 Koehler, R., Geiss, G., Rosenberg, A., Roco, C., Merico, D., Alidoust, N., Goodarzi, H., Yu, J.:
401 *Tahoe-100M*: A Giga-scale single-cell perturbation atlas for context-dependent gene function and
402 cellular modeling. *bioRxiv*, 2025–0220639398 (2025) <https://doi.org/10.1101/2025.02.20.639398>
- 403 [3] Bunne, C., Roohani, Y., Rosen, Y., Gupta, A., Zhang, X., Roed, M., Alexandrov, T., AlQuraishi,
404 M., Brennan, P., Burkhardt, D.B., Califano, A., Cool, J., Dernburg, A.F., Ewing, K., Fox, E.B.,
405 Haury, M., Herr, A.E., Horvitz, E., Hsu, P.D., Jain, V., Johnson, G.R., Kalil, T., Kelley, D.R., Kelley,
406 S.O., Kreshuk, A., Mitchison, T., Otte, S., Shendure, J., Sofroniew, N.J., Theis, F., Theodoris, C.V.,
407 Upadhyayula, S., Valer, M., Wang, B., Xing, E., Yeung-Levy, S., Zitnik, M., Karaletsos, T., Regev,
408 A., Lundberg, E., Leskovec, J., Quake, S.R.: How to build the virtual cell with artificial intelligence:
409 Priorities and opportunities. *arXiv* [q-bio.QM] (2024) [arXiv:2409.11654](https://arxiv.org/abs/2409.11654) [q-bio.QM]
- 410 [4] Wang, H., Leskovec, J., Regev, A.: Limitations of cell embedding metrics assessed using drifting
411 islands. *Nature biotechnology*, 1–4 (2025) <https://doi.org/10.1038/s41587-025-02702-z>
- 412 [5] Gunawan, I., Vafaei, F., Meijering, E., Lock, J.G.: An introduction to representation learning for

- 413 single-cell data analysis. *Cell reports methods* **3**(8), 100547 (2023) <https://doi.org/10.1016/j.crmeth.2023.100547>
- 415 [6] Theodoris, C.V., Xiao, L., Chopra, A., Chaffin, M.D., Al Sayed, Z.R., Hill, M.C., Mantineo, H.,
416 Brydon, E.M., Zeng, Z., Liu, X.S., Ellinor, P.T.: Transfer learning enables predictions in network
417 biology. *Nature* **618**(7965), 616–624 (2023) <https://doi.org/10.1038/s41586-023-06139-9>
- 418 [7] Cui, H., Wang, C., Maan, H., Pang, K., Luo, F., Duan, N., Wang, B.: scGPT: toward building a
419 foundation model for single-cell multi-omics using generative AI. *Nature methods* **21**(8), 1470–1480
420 (2024) <https://doi.org/10.1038/s41592-024-02201-0>
- 421 [8] Heimberg, G., Kuo, T., DePianto, D.J., Salem, O., Heigl, T., Diamant, N., Scalia, G., Biancalani,
422 T., Turley, S.J., Rock, J.R., Corrada Bravo, H., Kaminker, J., Vander Heiden, J.A., Regev, A.: A
423 cell atlas foundation model for scalable search of similar human cells. *Nature*, 1–3 (2024) <https://doi.org/10.1038/s41586-024-08411-y>
- 425 [9] Richter, T., Bahrami, M., Xia, Y., Fischer, D.S., Theis, F.J.: Delineating the effective use of self-
426 supervised learning in single-cell genomics. *Nature machine intelligence*, 1–11 (2024) <https://doi.org/10.1038/s42256-024-00934-3>
- 428 [10] Hestness, J., Narang, S., Ardalani, N., Diamos, G., Jun, H., Kianinejad, H., Patwary, M.M.A.,
429 Yang, Y., Zhou, Y.: Deep Learning Scaling is Predictable, Empirically. *arXiv* [cs.LG] (2017)
430 [arXiv:1712.00409](https://arxiv.org/abs/1712.00409) [cs.LG]
- 431 [11] Rosenfeld, J.S., Rosenfeld, A., Belinkov, Y., Shavit, N.: A constructive prediction of the generaliza-
432 tion error across scales. *arXiv* [cs.LG] (2019) [arXiv:1909.12673](https://arxiv.org/abs/1909.12673) [cs.LG]
- 433 [12] Kaplan, J., McCandlish, S., Henighan, T., Brown, T.B., Chess, B., Child, R., Gray, S., Radford, A.,
434 Wu, J., Amodei, D.: Scaling laws for neural language models. *arXiv* [cs.LG] (2020) [arXiv:2001.08361](https://arxiv.org/abs/2001.08361)
435 [cs.LG]
- 436 [13] Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., Casas, D.d.L.,
437 Hendricks, L.A., Welbl, J., Clark, A., Hennigan, T., Noland, E., Millican, K., Driessche, G., Damoc,
438 B., Guy, A., Osindero, S., Simonyan, K., Elsen, E., Rae, J.W., Vinyals, O., Sifre, L.: Training
439 Compute-Optimal Large Language Models. *arXiv* [cs.CL] (2022) [arXiv:2203.15556](https://arxiv.org/abs/2203.15556) [cs.CL]
- 440 [14] Bahri, Y., Dyer, E., Kaplan, J., Lee, J., Sharma, U.: Explaining neural scaling laws. *Proceedings*
441 of the National Academy of Sciences of the United States of America
- 442 **121**(27), 2311878121 (2024) <https://doi.org/10.1073/pnas.2311878121>

- 443 [15] Chen, D., Zhu, Y., Zhang, J., Du, Y., Li, Z., Liu, Q., Wu, S., Wang, L.: Uncovering neural scaling laws
444 in molecular Representation Learning. Neural Information Processing Systems **abs/2309.15123**,
445 1452–1475 (2023) <https://doi.org/10.48550/arXiv.2309.15123>
- 446 [16] Stoler, N., Nekrutenko, A.: Sequencing error profiles of Illumina sequencing instruments. NAR
447 genomics and bioinformatics **3**(1), 019 (2021) <https://doi.org/10.1093/nargab/lqab019>
- 448 [17] Hagemann-Jensen, M., Ziegenhain, C., Chen, P., Ramsköld, D., Hendriks, G.-J., Larsson, A.J.M.,
449 Faridani, O.R., Sandberg, R.: Single-cell RNA counting at allele and isoform resolution using Smart-
450 seq3. Nature biotechnology **38**(6), 708–714 (2020) <https://doi.org/10.1038/s41587-020-0497-0>
- 451 [18] Svensson, V., Natarajan, K.N., Ly, L.-H., Miragaia, R.J., Labalette, C., Macaulay, I.C., Cvejic, A.,
452 Teichmann, S.A.: Power analysis of single-cell RNA-sequencing experiments. Nature methods **14**(4),
453 381–387 (2017) <https://doi.org/10.1038/nmeth.4220>
- 454 [19] Lichtman, J.W., Conchello, J.-A.: Fluorescence microscopy. Nature methods **2**(12), 910–919 (2005)
455 <https://doi.org/10.1038/nmeth817>
- 456 [20] Bansal, Y., Ghorbani, B., Garg, A., Zhang, B., Krikun, M., Cherry, C., Neyshabur, B., Firat,
457 O.: Data scaling laws in NMT: The effect of noise and architecture. arXiv [cs.LG] (2022)
458 [arXiv:2202.01994](https://arxiv.org/abs/2202.01994) [cs.LG]
- 459 [21] Qiu, C., Martin, B.K., Welsh, I.C., Daza, R.M., Le, T.-M., Huang, X., Nichols, E.K., Taylor, M.L.,
460 Fulton, O., O'Day, D.R., Gomes, A.R., Ilcisin, S., Srivatsan, S., Deng, X., Disteche, C.M., Noble,
461 W.S., Hamazaki, N., Moens, C.B., Kimelman, D., Cao, J., Schier, A.F., Spielmann, M., Murray, S.A.,
462 Trapnell, C., Shendure, J.: A single-cell time-lapse of mouse prenatal development from gastrula to
463 birth. Nature **626**(8001), 1084–1093 (2024) <https://doi.org/10.1038/s41586-024-07069-w>
- 464 [22] Brandfonbrener, D., Anand, N., Vyas, N., Malach, E., Kakade, S.: Loss-to-loss prediction: Scaling
465 laws for all datasets. arXiv [cs.LG] (2024) [arXiv:2411.12925](https://arxiv.org/abs/2411.12925) [cs.LG]
- 466 [23] Belinkov, Y.: Probing classifiers: Promises, shortcomings, and advances. Computational linguistics
467 (Association for Computational Linguistics) **48**(1), 207–219 (2022) https://doi.org/10.1162/coli_a_00422
- 468 [24] Pimentel, T., Valvoda, J., Maudslay, R.H., Zmigrod, R., Williams, A., Cotterell, R.: Information-
469 theoretic probing for linguistic structure. arXiv [cs.CL] (2020) [arXiv:2004.03061](https://arxiv.org/abs/2004.03061) [cs.CL]
- 470 [25] Gowri, G., Lun, X.-K., Klein, A.M., Yin, P.: Approximating mutual information of high-dimensional
471 variables using learned representations. In: Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet,

- 473 U., Tomczak, J., Zhang, C. (eds.) Advances in Neural Information Processing Systems, vol. 37, pp.
474 132843–132875. Curran Associates, Inc., ??? (2024)
- 475 [26] Hao, Y., Hao, S., Andersen-Nissen, E., Mauck, W.M. 3rd, Zheng, S., Butler, A., Lee, M.J., Wilk, A.J.,
476 Darby, C., Zager, M., Hoffman, P., Stoeckius, M., Papalex, E., Mimitou, E.P., Jain, J., Srivastava,
477 A., Stuart, T., Fleming, L.M., Yeung, B., Rogers, A.J., McElrath, J.M., Blish, C.A., Gottardo, R.,
478 Smibert, P., Satija, R.: Integrated analysis of multimodal single-cell data. *Cell* **184**(13), 3573–358729
479 (2021) <https://doi.org/10.1016/j.cell.2021.04.048>
- 480 [27] Weinreb, C., Rodriguez-Fraticelli, A., Camargo, F.D., Klein, A.M.: Lineage tracing on transcrip-
481 tional landscapes links state to fate during differentiation. *Science* **367**(6479) (2020) <https://doi.org/10.1126/science.aaw3381>
- 482 [28] Vizgen: Vizgen Data Release V1.0. Title of the publication associated with this dataset: Mouse
483 Brain Receptor Map (2021)
- 484 [29] Lopez, R., Regier, J., Cole, M.B., Jordan, M.I., Yosef, N.: Deep generative modeling for
485 single-cell transcriptomics. *Nature methods* **15**(12), 1053–1058 (2018) <https://doi.org/10.1038/s41592-018-0229-2>
- 486 [30] Rasley, J., Rajbhandari, S., Ruwase, O., He, Y.: DeepSpeed: System optimizations enable training
487 deep learning models with over 100 billion parameters. In: Proceedings of the 26th ACM SIGKDD
488 International Conference on Knowledge Discovery & Data Mining. ACM, New York, NY, USA
489 (2020). <https://doi.org/10.1145/3394486.3406703>
- 490 [31] Guo, D.: Gaussian channels: Information, estimation and multiuser detection. PhD thesis (2004)
- 491 [32] Klein, A.M., Mazutis, L., Akartuna, I., Tallapragada, N., Veres, A., Li, V., Peshkin, L., Weitz, D.A.,
492 Kirschner, M.W.: Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells.
493 *Cell* **161**(5), 1187–1201 (2015) <https://doi.org/10.1016/j.cell.2015.04.044>
- 494 [33] Polyanskiy, Y., Wu, Y.: Information Theory: From Coding to Learning. Cambridge university press,
495 ??? (2024)
- 496 [34] Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.-C.: MobileNetV2: Inverted residuals
497 and linear bottlenecks. *arXiv* [cs.CV] (2018) [arXiv:1801.04381](https://arxiv.org/abs/1801.04381) [cs.CV]
- 498 [35] Fei-Fei, L., Fergus, R., Perona, P.: Learning generative visual models from few training examples:
499 An incremental bayesian approach tested on 101 object categories, 178–178 (2004)

- 502 [36] Yang, J., Shi, R., Wei, D., Liu, Z., Zhao, L., Ke, B., Pfister, H., Ni, B.: MedMNIST v2 - A large-
503 scale lightweight benchmark for 2D and 3D biomedical image classification. *Scientific Data* **10**(1),
504 41 (2023) <https://doi.org/10.1038/s41597-022-01721-8>
- 505 [37] Costa, G.B.P., Contato, W.A., Nazare, T.S., Neto, J.a.E.S.B., Ponti, M.: An empirical study on the
506 effects of different types of noise in image classification tasks. *arXiv [cs.CV]* (2016) [arXiv:1609.02781](https://arxiv.org/abs/1609.02781)
507 [cs.CV]
- 508 [38] Rives, A., Meier, J., Sercu, T., Goyal, S., Lin, Z., Liu, J., Guo, D., Ott, M., Zitnick, C.L., Ma,
509 J., Fergus, R.: Biological structure and function emerge from scaling unsupervised learning to 250
510 million protein sequences. *Proceedings of the National Academy of Sciences of the United States of
511 America* **118**(15) (2021) <https://doi.org/10.1073/pnas.2016239118>
- 512 [39] Shu, Y., McCauley, J.: GISAID: Global initiative on sharing all influenza data - from vision to
513 reality. *Euro surveillance : bulletin European sur les maladies transmissibles [Euro surveillance
514 : European communicable disease bulletin]* **22**(13) (2017) [https://doi.org/10.2807/1560-7917.ES.
2017.22.13.30494](https://doi.org/10.2807/1560-7917.ES.
515 2017.22.13.30494)
- 516 [40] Cover, T.M., Thomas, J.A.: Elements of Information Theory. Wiley & Sons, Incorporated, John,
517 ??? (2006)
- 518 [41] Cai, T.T., Ma, Z., Wu, Y.: Sparse PCA: Optimal rates and adaptive estimation. *arXiv [math.ST]*
519 (2012) <https://doi.org/10.48550/arXiv.1211.1309> [arXiv:1211.1309](https://arxiv.org/abs/1211.1309) [math.ST]
- 520 [42] Virshup, I., Bredikhin, D., Heumos, L., Palla, G., Sturm, G., Gayoso, A., Kats, I., Koutrouli,
521 M., Scverse Community, Berger, B., Pe'er, D., Regev, A., Teichmann, S.A., Finotello, F., Wolf,
522 F.A., Yosef, N., Stegle, O., Theis, F.J.: The scverse project provides a computational ecosystem
523 for single-cell omics data analysis. *Nature biotechnology* **41**(5), 604–606 (2023) [https://doi.org/10.1038/s41587-023-01733-8](https://doi.org/10.
524 1038/s41587-023-01733-8)
- 525 [43] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein,
526 N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy,
527 S., Steiner, B., Fang, L., Bai, J., Chintala, S.: PyTorch: An Imperative Style, High-Performance
528 Deep Learning Library. *arXiv [cs.LG]* (2019) [arXiv:1912.01703](https://arxiv.org/abs/1912.01703) [cs.LG]
- 529 [44] Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R.,
530 Funtowicz, M., Davison, J., Shleifer, S., Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T.,
531 Gugger, S., Drame, M., Lhoest, Q., Rush, A.: Transformers: State-of-the-art natural language pro-
532 cessing. In: Liu, Q., Schlangen, D. (eds.) *Proceedings of the 2020 Conference on Empirical Methods*

533 in Natural Language Processing: System Demonstrations, pp. 38–45. Association for Computational
534 Linguistics, Stroudsburg, PA, USA (2020). <https://doi.org/10.18653/v1/2020.emnlp-demos.6>

535 [45] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Pret-
536 tenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M.,
537 Perrot, M., Duchesnay, E.: Scikit-learn: Machine Learning in Python. Journal of Machine Learning
538 Research **12**(85), 2825–2830 (2011)