Scaling Laws for Masked-Reconstruction Transformers on Single-Cell Transcriptomics

Anonymous authors
Paper under double-blind review

Abstract

Neural scaling laws—power-law relationships between loss, model size, and data—have been extensively documented for language and vision transformers, yet their existence in single-cell genomics remains largely unexplored. We present the first systematic study of scaling behaviour for masked-reconstruction transformers trained on single-cell RNA sequencing (scRNA-seq) data. Using expression profiles from the CELLxGENE Census, we construct two experimental regimes: a *data-rich* regime (512 highly variable genes, 200,000 cells) and a *data-limited* regime (1,024 genes, 10,000 cells). Across seven model sizes spanning three orders of magnitude in parameter count (533 to 3.4×10^8 parameters), we fit the parametric scaling law $L = aP^{-\alpha} + c$ to validation mean squared error (MSE). The data-rich regime exhibits clear power-law scaling ($\alpha = 0.234$, $R^2 = 0.823$) with an irreducible loss floor of $c \approx 1.44$, while the data-limited regime shows negligible scaling ($\alpha = 0.009$, $R^2 = 0.017$), indicating that model capacity is not the binding constraint when data are scarce. These results establish that scaling laws analogous to those observed in natural language processing do emerge in single-cell transcriptomics when sufficient data are available, and they identify the data-to-parameter ratio as a critical determinant of scaling behaviour. A preliminary conversion of the data-rich asymptotic floor to information-theoretic units yields an estimate of approximately 2.30 bits of entropy per masked gene position. We discuss implications for the design of single-cell foundation models and outline the additional measurements needed to refine this entropy estimate.

1 Introduction

A central empirical finding in deep learning is that, under appropriate conditions, the loss of a neural network follows a power law as model size, dataset size, or training compute increases (Kaplan et al., 2020; Hestness et al., 2017). These *scaling laws* have proven remarkably consistent across modalities: they hold for autoregressive language models (Kaplan et al., 2020; Hoffmann et al., 2022), vision transformers (Zhai et al., 2022), and multimodal systems (Cherti et al., 2023). Beyond their theoretical interest, scaling laws have immediate practical value: they enable researchers to predict the performance of expensive large-scale training runs from cheap small-scale experiments, thereby guiding resource allocation and architectural choices.

Single-cell RNA sequencing (scRNA-seq) has become a cornerstone of modern biology, generating expression profiles for millions of individual cells across diverse tissues, species, and disease states (Regev et al., 2017). In response, the community has developed increasingly large *foundation models* for single-cell data, including scVI (Lopez et al., 2018), scGPT (Cui et al., 2024), Geneformer (Theodoris et al., 2023), scBERT (Yang et al., 2022), and scFoundation (Hao et al., 2024). These models are typically evaluated on downstream tasks such as cell-type annotation, batch correction, or perturbation prediction, but the fundamental question of how pretraining loss scales with model size and data volume has received almost no attention.

This gap matters for several reasons. First, without scaling laws, practitioners cannot determine whether a larger model will yield meaningful improvements or whether the current bottleneck is data, compute, or intrinsic noise. Second, the asymptotic loss floor c in a scaling law of the form $L = aP^{-\alpha} + c$ provides an empirical upper bound on the *reducible* error, which can be related to the intrinsic entropy of the data-

generating process (Kaplan et al., 2020; Hoffmann et al., 2022). For transcriptomics, such an estimate would quantify the fundamental information content of gene expression measurements, a quantity of independent biological interest.

In this work, we take a first step towards establishing scaling laws for single-cell transcriptomics. We train masked-reconstruction transformers—a pretraining paradigm analogous to masked language modelling (Devlin et al., 2019)—on scRNA-seq data drawn from the CELLxGENE Census (CZI Single-Cell Biology, 2023), varying the model size across seven presets that span from 533 to 3.4×10^8 trainable parameters. We study two regimes that differ in both gene vocabulary size and dataset size, fit power-law models to the resulting loss curves, and analyse the conditions under which scaling behaviour emerges or breaks down.

Contributions.

1. We present, to our knowledge, the first systematic investigation of neural scaling laws for masked-reconstruction pretraining on scRNA-seq data.
2. We demonstrate that power-law scaling ($\alpha \approx 0.23$, $R^2 \approx 0.82$) emerges in a data-rich regime (200,000 cells, 512 genes), with an identifiable irreducible loss floor.
3. We show that scaling breaks down ($\alpha \approx 0.009$, $R^2 \approx 0.02$) in a data-limited regime (10,000 cells, 1,024 genes), providing direct evidence that data scarcity—not model capacity—is the binding constraint.
4. We identify the key experimental ingredients—information-theoretic loss functions, matched datasets, and comprehensive token accounting—needed to progress from loss-vs-parameters curves to transcriptomic entropy estimates.

2 Related Work

Neural scaling laws. Hestness et al. (2017) first documented power-law scaling of test error with dataset size across several domains. Kaplan et al. (2020) established that language model cross-entropy loss follows $L(P) = (P_c/P)^{\alpha_P}$ over several orders of magnitude in parameter count P, with $\alpha_P \approx 0.076$ for autoregressive transformers. Importantly, they found that scaling with model size, data size, and compute each follows its own power law, and that training can be either model-limited or data-limited depending on the regime. Hoffmann et al. (2022) refined these findings, demonstrating that Kaplan *et al.*'s protocol over-allocates parameters relative to data and proposing compute-optimal (*Chinchilla*) scaling rules. Subsequent work has extended scaling laws to vision transformers (Zhai et al., 2022), contrastive multimodal models (Cherti et al., 2023), and various other architectures and tasks (Clark et al., 2022; Muennighoff et al., 2024).

Single-cell foundation models. The single-cell community has developed several large pretrained models. scVI (Lopez et al., 2018) introduced variational autoencoders for scRNA-seq, and subsequent work extended this framework to multi-omic data (Gayoso et al., 2022). More recently, transformer-based architectures have gained prominence: scGPT (Cui et al., 2024) adopts a generative pretraining approach with gene tokens; Geneformer (Theodoris et al., 2023) pretrains on rank-ordered gene expression using a BERT-like masked objective; scBERT (Yang et al., 2022) similarly uses masked gene expression prediction; and scFoundation (Hao et al., 2024) scales to 100 million parameters trained on 50 million cells. While these works report downstream benchmark results, none systematically characterizes how pretraining loss scales with model size, which is the question we address here.

Masked autoencoders for genomics. Masked reconstruction is a well-established self-supervised pretraining strategy. In natural language processing, BERT (Devlin et al., 2019) popularized masked token prediction. In computer vision, masked autoencoders (MAE) (He et al., 2022) demonstrated that masking high fractions of image patches leads to effective visual representations. For single-cell data, the masked reconstruction objective is a natural fit: by randomly masking a subset of gene expression values and training the model to reconstruct them from the remaining genes, the model learns co-expression structure and gene regulatory relationships. Our architecture follows this paradigm, using a permutation-invariant transformer encoder with a per-gene reconstruction head.

3 Data

3.1 Source and Preprocessing

All data are drawn from the CELLxGENE Census (CZI Single-Cell Biology, 2023), a curated aggregation of publicly available single-cell RNA-seq datasets encompassing diverse human tissues and cell types. We construct two experimental regimes, each defined by a gene vocabulary size V (the number of highly variable genes retained) and a dataset size D (the number of cells):

- **Regime A** ($V = 512$, $D = 200{,}000$): 200,000 cells with 512 highly variable genes (HVGs).
- **Regime B** ($V = 1{,}024$, $D = 10{,}000$): 10,000 cells with 1,024 HVGs.

The preprocessing pipeline proceeds as follows. First, we load the expression matrix from the Census and select the expression layer with a valid library size (preferring raw counts where available). Zero-library cells are removed. Second, we select highly variable genes using the Seurat v3 method (Stuart et al., 2019), which ranks genes by their variance-to-mean ratio after variance-stabilizing transformation. Third, we apply library-size normalization to a target sum of 10^4 followed by the $\log(1 + x)$ transform. The resulting matrix is stored in float32. Finally, we split cells into training (90%), validation (5%), and test (5%) partitions using stratified sampling based on available cell-type annotations where possible, with random splitting as fallback.

3.2 Regime Comparison

The two regimes differ along two axes simultaneously: gene vocabulary size and dataset size. This confounding is a limitation of the current study (see Section 9), but it reflects the practical constraints of the available data and allows us to probe whether scaling behaviour is robust or fragile across substantially different settings. Regime A provides a data-rich environment ($D/V \approx 390$ cells per gene) while Regime B is comparatively data-starved ($D/V \approx 10$ cells per gene).

4 Model Architecture

4.1 Overview

We employ a permutation-invariant transformer encoder designed for set-structured inputs, where each element of the set corresponds to a gene. The architecture follows the standard transformer encoder (Vaswani et al., 2017) but omits positional encodings entirely, since the ordering of genes carries no inherent biological meaning. The model receives three inputs: gene identifiers (integer indices into a learned embedding table), observed expression values, and a binary mask indicating which values have been hidden.

4.2 Input Embedding

Each gene token is formed by summing two components:

1. A **gene identity embedding** $\mathbf{e}_g \in \mathbb{R}^d$, obtained by looking up the gene index in a learned embedding table $\mathbf{E} \in \mathbb{R}^{V \times d}$.
2. A **value projection** $\mathbf{v}_g = \mathbf{W}_v x_g + \mathbf{b}_v$, where $\mathbf{W}_v \in \mathbb{R}^{d \times 1}$ and x_g is the (possibly masked) expression value for gene g.

At masked positions, the value projection is replaced by a learned **mask token** $\mathbf{m} \in \mathbb{R}^d$, initialized to zero and updated during training. The resulting token representation for gene g is:

$$\mathbf{h}_g^{(0)} = \mathbf{e}_g + \begin{cases} \mathbf{v}_g & \text{if gene } g \text{ is observed} \\ \mathbf{m} & \text{if gene } g \text{ is masked.} \end{cases} \tag{1}$$

Table 1: Model size presets used in scaling experiments. Parameter counts are computed for gene vocabularies $V = 512$ and $V = 1{,}024$. d: model dimension; L: number of encoder layers; H: number of attention heads; FFN mult: feed-forward expansion factor.

Size	d	L	H	**FFN mult**	**Params** ($V = 512$)	**Params** ($V = 1{,}024$)
XXS	1	1	1	1	534	1,046
TINY	16	1	1	1	9,937	18,129
XS	64	2	4	4	132,993	165,761
S	128	4	8	4	859,137	924,673
M	512	6	8	4	19,178,497	19,440,641
L	1020	8	12	4	101,033,041	101,555,281
XL	1536	12	16	4	341,295,105	341,557,249

4.3 Transformer Encoder

The token representations $\{\mathbf{h}_g^{(0)}\}_{g=1}^{V}$ are processed by a stack of L standard transformer encoder layers (Vaswani et al., 2017), each comprising multi-head self-attention and a position-wise feed-forward network with GELU activation (Hendrycks & Gimpel, 2016). We use pre-layer normalization (Pre-LN) (Xiong et al., 2020), which places layer normalization before (rather than after) the attention and feed-forward sub-layers, yielding more stable training dynamics at scale. The feed-forward hidden dimension is $4d$ by default.

4.4 Prediction Head and Pooling

The output of the encoder at each gene position is passed through a linear prediction head that maps $\mathbb{R}^d \to \mathbb{R}^1$, producing a scalar expression estimate $\hat{x}_g$ for each gene. The loss is computed only at masked positions (Section 5). For downstream applications requiring a cell-level representation, the model produces a pooled embedding $\mathbf{z} \in \mathbb{R}^d$ via mean-pooling over all gene tokens.

4.5 Model Sizes

To study scaling behaviour, we define seven size presets that span approximately five orders of magnitude in parameter count. Table 1 summarizes their configurations. Parameter counts are reported for the two vocabulary sizes used in our experiments; the difference arises from the gene embedding matrix $\mathbf{E} \in \mathbb{R}^{V \times d}$.

5 Training

5.1 Masked Reconstruction Objective

Following the masked language modelling paradigm (Devlin et al., 2019), we randomly mask 15% of gene expression values in each training sample. The training objective is to minimize the mean squared error (MSE) between the model's predictions and the true expression values at masked positions only:

$$\mathcal{L} = \frac{1}{|\mathcal{M}|} \sum_{g \in \mathcal{M}} \left(\hat{x}_g - x_g\right)^2, \tag{2}$$

where $\mathcal{M}$ denotes the set of masked gene indices and $|\mathcal{M}|$ its cardinality. We also monitor masked mean absolute error (MAE) as a secondary metric.

5.2 Optimization

We use AdamW (Loshchilov & Hutter, 2019) with weight decay $\lambda = 0.01$ and gradient clipping at norm 1.0. For Regime A runs on GPU, the base learning rate is 3.125×10^{-5}, computed as $\mathrm{lr}_{\mathrm{base}} \cdot \mathrm{eff_batch}/256$ where $\mathrm{lr}_{\mathrm{base}} = 2.5 \times 10^{-4}$ and the effective batch size is 32 (physical batch of 4 with 8 gradient accumulation steps).

Table 2: Best validation MSE ranges per model size and regime. Ranges indicate the minimum and maximum across random seeds. A dash indicates that the configuration was not run.

Size	**Regime A** ($V = 512$, 200k cells)	**Regime B** ($V = 1{,}024$, 10k cells)
XXS	2.07–2.14	1.04–1.42
TINY	1.59–1.61	–
XS	1.51–1.58	1.08–1.87
S	1.47–1.52	1.13–1.91
M	1.46–1.51	1.10–1.21
L	1.47–1.54	1.21
XL	–	1.26

All GPU runs use mixed-precision training (float16 autocast) for efficiency. Models in Regime A are trained for 60,000 optimization steps; Regime B runs use 20,000–30,000 steps depending on the sweep configuration.

5.3 Evaluation Protocol

We evaluate on the held-out validation split every 1,000 steps (or every 500 steps for shorter runs), saving the checkpoint with the lowest validation MSE (`ckpt_best.pt`). The final scaling analysis uses the best validation MSE achieved during training for each run, which we denote L^*. Each model size in Regime A is trained with three random seeds (7, 8, 9) to characterize run-to-run variance; Regime B coverage varies by model size.

6 Scaling Law Formulation

Following Kaplan et al. (2020), we model the relationship between validation MSE and parameter count P as a power law with an additive offset:

$$L(P) = a \cdot P^{-\alpha} + c, \tag{3}$$

where $\alpha > 0$ is the *scaling exponent* governing the rate of improvement, $a > 0$ is a multiplicative constant, and $c \geq 0$ represents the *irreducible loss floor*—the error that cannot be reduced by increasing model capacity, attributable to intrinsic data noise, biological stochasticity, or information lost during preprocessing.

We fit Equation equation 3 by grid search over candidate values of c in $[0, 0.99 \cdot \min_i L_i]$, and for each candidate perform ordinary least squares on the linearized form $\log(L - c) = \log a - \alpha \log P$. The fit with the highest R^2 on the log-scale residuals is selected. We fit the two regimes independently, as their differing data sizes and vocabularies make a pooled fit uninformative.

7 Results

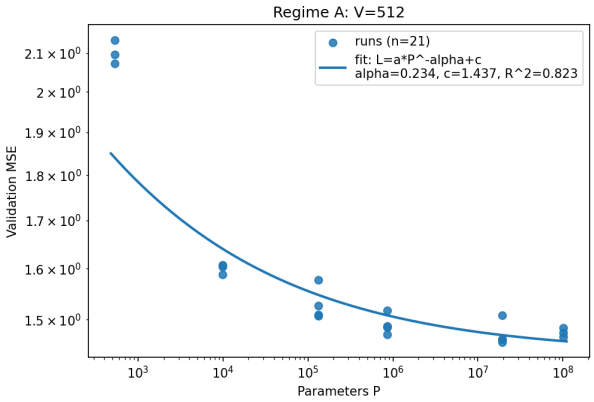
7.1 Overview of Training Runs

Table 2 summarizes the best validation MSE achieved for each model size and regime, aggregated across random seeds. In total, we conducted 48 training runs: 21 runs for Regime A (seven sizes $\times$ three seeds, minus sizes not run at all seeds) and 27 runs for Regime B across various configurations.

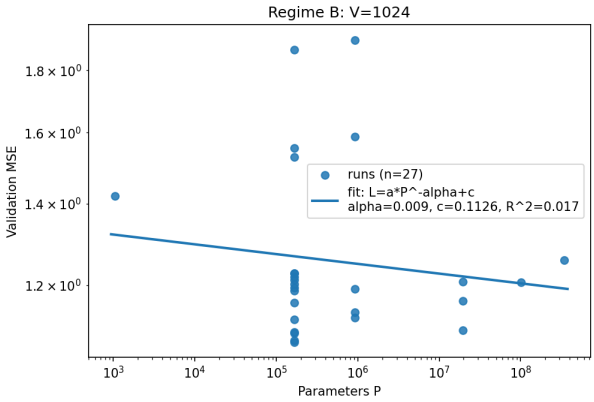
Several patterns are immediately apparent. In Regime A, MSE decreases monotonically from XXS ($\sim$2.1) through M ($\sim$1.46), after which the L model provides no further improvement and in fact shows slightly higher MSE than M, suggesting that the irreducible floor has been reached. In Regime B, the MSE values are lower overall (consistent with the larger gene vocabulary providing a richer reconstruction target on normalized data), but the variation across model sizes is small and inconsistent, with no clear monotonic trend.

Table 3: Fitted parameters of the scaling law $L = aP^{-\alpha} + c$ for each regime. n: number of training runs included in the fit. The "canonical" Regime A row uses the 18 standardized runs (6 sizes $\times$ 3 seeds, all 60k steps); the "all runs" row includes additional 30k-step runs.

Regime	n	α	a	c	R^2
A: $V = 512$ (all runs)	21	0.234	1.750	1.437	0.823
A: $V = 512$ (canonical)	18	0.266	2.153	1.444	0.858
B: $V = 1{,}024$, 10k cells	27	0.009	1.283	0.113	0.017



(a) Regime A ($V = 512$, 200k cells).



(b) Regime B ($V = 1{,}024$, 10k cells).

Figure 1: Validation MSE versus parameter count P on log-log axes, with fitted scaling curves (Equation 3). Each point represents one training run (best checkpoint). Regime A (left) shows clear power-law decay with $R^2 = 0.82$; Regime B (right) is effectively flat with $R^2 = 0.02$.

7.2 Power-Law Fits

Table 3 reports the fitted scaling-law parameters for each regime, and Figures 1a and 1b show the corresponding log-log plots.

7.2.1 Regime A: Data-Rich Scaling

The $V = 512$ regime with 200,000 cells exhibits a clear scaling relationship. Using the full set of 21 runs (which includes some 30k-step runs), the fitted exponent is $\alpha = 0.234$ with $R^2 = 0.823$. Restricting to the 18 canonical runs (6 sizes $\times$ 3 seeds, all trained for 60,000 steps with identical hyperparameters) yields a tighter fit: $\alpha = 0.266$ with $R^2 = 0.858$. In the canonical fit, each tenfold increase in parameter count reduces the excess loss $(L - c)$ by a factor of approximately $10^{0.266} \approx 1.84$. The R^2 of 0.86 confirms a strong fit, though it is lower than the $R^2 > 0.95$ typically reported for language model scaling (Kaplan et al., 2020), likely because our model size range is more limited and our MSE metric introduces more noise than cross-entropy.

The irreducible floor $c \approx 1.44$ is consistent across both fit variants. It suggests that even an infinitely large model trained on this dataset and preprocessing pipeline would achieve a validation MSE no lower than approximately 1.44. Under a Gaussian noise assumption, this floor corresponds to approximately 2.30 bits of entropy per masked gene position (Section 7.4). The floor encompasses several sources of irreducible error: biological stochasticity (intrinsic gene expression noise), technical measurement noise from the sequencing process, information lost during HVG selection (the remaining 512 genes cannot capture all co-expression patterns in the transcriptome), and the inherent information-theoretic limit of the reconstruction task under the chosen masking strategy.

The observation that L-size models (∼101M parameters) do not improve over M-size models (∼19M parameters) is consistent with the fitted curve, which has largely flattened by $P \approx 10^7$. This provides practical guidance: for this dataset and task, models beyond ∼20M parameters offer diminishing returns.

7.2.2 Regime B: Data-Limited Plateau

In stark contrast, the $V = 1{,}024$ regime shows essentially no scaling. The fitted exponent $\alpha = 0.009$ is near zero, and $R^2 = 0.017$ indicates that parameter count explains less than 2% of the variance in validation MSE. The scatter in Figure 1b reveals no discernible trend: small and large models achieve comparable loss values.

This result admits a straightforward interpretation. With only 10,000 cells and 1,024 genes, the dataset is too small to benefit from increased model capacity. Larger models memorize the training data without improving generalization, a classic symptom of the data-limited regime described by Kaplan et al. (2020). The cell-to-gene ratio $D/V \approx 10$ in Regime B contrasts sharply with $D/V \approx 390$ in Regime A, suggesting that data richness relative to the input dimensionality is a key factor in whether scaling laws emerge.

Additional heterogeneity in Regime B's results comes from inconsistent training durations across runs (20,000–60,000 steps) and the inclusion of short benchmark runs (≤400 steps), which may not have converged. This mixing of converged and unconverged runs further obscures any underlying scaling trend.

7.3 Comparison with Language Model Scaling

The scaling exponent $\alpha = 0.23$–0.27 observed in Regime A (depending on the run subset) is substantially larger than the $\alpha_P \approx 0.076$ reported by Kaplan et al. (2020) for autoregressive language models, or the $\alpha_P \approx 0.34$ reported by Hoffmann et al. (2022) after correcting for compute-optimal training. Direct comparison is complicated by several factors: we use MSE rather than cross-entropy (Appendix A), our masking rate (15%) is much lower than typical language model sequence lengths, and the structure of gene expression data differs fundamentally from natural language. Nevertheless, the qualitative finding—that loss follows a power law with diminishing returns at large scale—is consistent across domains.

The dual-regime behaviour we observe echoes the "contradiction" discussed by Kaplan et al. (2020), who noted that scaling with parameters appears to follow a different exponent than scaling with data. In our setting, the contradiction manifests more starkly: the same architecture and training procedure produces clear scaling in one regime and no scaling in the other, purely as a function of the data-to-parameter ratio.

7.4 Preliminary Entropy Estimate for Regime A

A central motivation for fitting scaling laws with an additive offset c is that the irreducible floor can, under appropriate assumptions, be converted to an estimate of the *intrinsic entropy* of the data-generating process—the information content per prediction that no model can recover. While a full entropy estimate requires a native information-theoretic loss (Section 9), we present a preliminary estimate for Regime A using two complementary derivations from the fitted floor values.

7.4.1 Gaussian NLL Derivation from MSE

Our training runs record MSE as the primary loss. To obtain a Gaussian negative log-likelihood (NLL), we derive it post hoc under the assumption that the residual errors at masked positions are approximately Gaussian with variance equal to the MSE. Specifically, for a Gaussian with variance σ^2, the NLL per position is:

$$\mathcal{L}_{\text{NLL}} = \tfrac{1}{2}\ln(2\pi\sigma^2) + \tfrac{1}{2}, \tag{4}$$

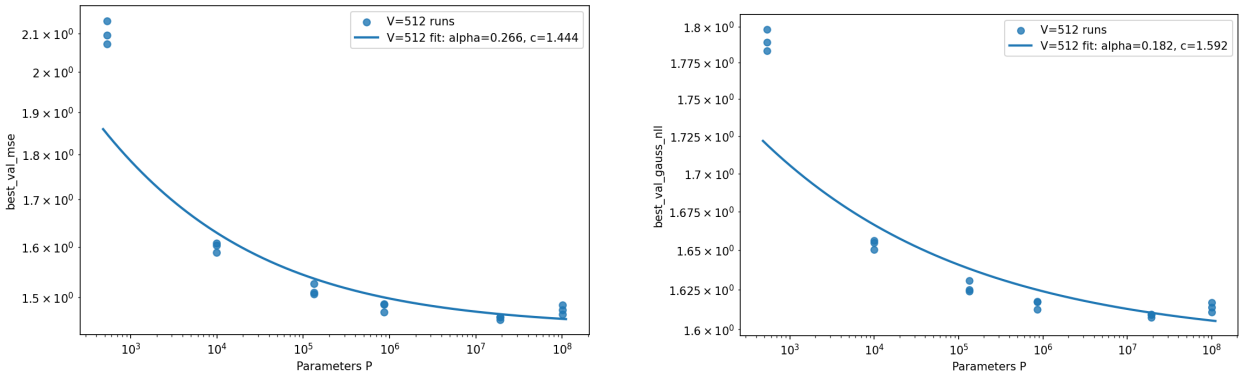
and the differential entropy (in bits) is:

$$h = \tfrac{1}{2}\log_2(2\pi e\,\sigma^2). \tag{5}$$

Using a standardized set of 18 canonical Regime A runs (6 sizes × 3 seeds, all trained for 60,000 steps with identical hyperparameters), we refit the scaling law for both the MSE and the derived Gaussian NLL metrics. Table 4 reports the results.

Table 4: Scaling-law fits for Regime A ($V = 512$, 200k cells) using the 18 canonical runs, with entropy floor estimates. The NLL values are derived from MSE under a Gaussian assumption (Equation 4), so the two estimates are not fully independent.

Metric	n	α	a	c	R^2	**Entropy floor (bits/masked pos.)**
MSE	18	0.266	2.153	1.444	0.858	2.312
Gauss. NLL	18	0.182	0.401	1.592	0.838	2.296



(a) MSE vs. parameters ($\alpha = 0.266$, $c = 1.444$).

(b) Gaussian NLL vs. parameters ($\alpha = 0.182$, $c = 1.592$).

Figure 2: Canonical Regime A scaling fits (18 runs, 6 sizes $\times$ 3 seeds, all 60k steps). Left: MSE metric. Right: derived Gaussian NLL metric. Both converge on an entropy floor of $\sim$2.30 bits per masked position.

Figure 2 shows the scaling curves for both metrics on the canonical run set.

7.4.2 Entropy Estimate

The two derivations yield closely agreeing estimates:

- **From MSE floor** ($c_{\text{MSE}} = 1.444$): interpreting c as the irreducible variance σ^2 and applying Equation equation 5 gives $h = \frac{1}{2}\log_2(2\pi e \cdot 1.444) = 2.312$ bits per masked position.
- **From NLL floor** ($c_{\text{NLL}} = 1.592$ nats): converting to bits via $h = c_{\text{NLL}}/\ln 2 = 2.296$ bits per masked position.

Both approaches converge on an estimate of approximately **2.30 bits per masked gene position** for the $V = 512$ regime.

7.4.3 Caveats

This estimate is preliminary and subject to several important caveats. First, the Gaussian NLL values are currently *derived from* the MSE values (not independently logged during training), so the two estimates in Table 4 are not fully independent; their agreement primarily reflects the self-consistency of the Gaussian assumption rather than an independent validation. Future runs should log the Gaussian NLL natively (predicting both μ and σ^2) to obtain a genuinely independent estimate.

Second, the entropy floor depends on the preprocessing pipeline. The $\log(1 + x)$ transform and library-size normalization change the scale and distribution of gene expression values; the 2.30-bit figure applies specifically to masked positions in $\log(1 + x)$-transformed, library-size-normalized expression over 512 HVGs. Changing the vocabulary size, normalization target, or transformation would alter the floor.

Third, a complete entropy estimate requires not only the $L(P)$ scaling law (loss vs. parameters) but also the $L(D)$ law (loss vs. dataset size) and ideally the intersection $L(C_{\min})$ of the compute-optimal frontier, following the methodology of Hoffmann et al. (2022). Our current experiments do not yet include a systematic $L(D)$ sweep for Regime A.

Despite these limitations, the convergence of the two derivations on ∼2.30 bits provides a useful first benchmark: it suggests that, for each masked gene among the top 512 HVGs, approximately 2.3 bits of expression information can be predicted from the remaining genes, and the rest is irreducible noise given this data and preprocessing.

8 Discussion

8.1 Implications for Single-Cell Foundation Models

Our results carry several implications for the design and training of single-cell foundation models.

Data is the bottleneck. The sharp contrast between Regimes A and B underscores that, in the current single-cell setting, data quantity (relative to model capacity) is the primary determinant of whether larger models are beneficial. This finding aligns with the compute-optimal perspective of Hoffmann et al. (2022), who argue that language models have historically been over-parameterized relative to their training data. For single-cell models, the implication is that scaling up model size without proportionally increasing the training corpus is unlikely to yield gains.

Diminishing returns beyond ∼20M parameters. Within Regime A, the loss curve flattens around 10^7 parameters. While this specific threshold depends on the dataset, gene vocabulary, and preprocessing choices, it suggests that for datasets of ∼200,000 cells, models in the tens of millions of parameters are sufficient to capture the learnable structure. This is a useful calibration point for practitioners deciding whether to invest in training larger models.

The irreducible floor as a biological signal. The estimated $c \approx 1.44$ (MSE) for Regime A, corresponding to approximately 2.30 bits of entropy per masked gene position (Section 7.4), reflects the combined effect of biological noise, technical noise, and information loss from dimensionality reduction. Disentangling these contributions is an important direction for future work: if the floor is dominated by technical noise, it may be reducible through improved experimental protocols; if it reflects intrinsic biological stochasticity, it represents a fundamental limit on expression prediction. The 2.30-bit figure provides a concrete benchmark against which future models and datasets can be compared.

8.2 The Data-vs-Parameter Ratio

A unifying interpretation of our two regimes is that scaling behaviour emerges when the data-to-parameter ratio exceeds some threshold. In Regime A, even the L model (∼101M parameters) is trained on 200,000 cells $\times$ 512 genes $\approx 10^8$ values, giving a ratio of ∼1 datum per parameter. In Regime B, the XL model (∼341M parameters) sees only 10,000 $\times$ 1,024 $\approx 10^7$ values, a ratio of ∼0.03 data per parameter. The transition from effective to ineffective scaling appears to occur somewhere in between.

We note that the effective data budget is further multiplied by the number of training steps and the mask rate: each cell is seen multiple times with different random masks, effectively augmenting the dataset. A full accounting of *tokens seen* (steps $\times$ effective batch size $\times V$) would enable a cleaner analysis of data-vs-compute scaling, which we leave to future work.

8.3 Comparison to Downstream Evaluations

Prior single-cell foundation models are typically evaluated on downstream tasks (cell-type classification, batch correction, perturbation response prediction) rather than pretraining loss. Our work is complementary: we focus on the pretraining loss itself, which is the quantity most directly comparable across model sizes and

most amenable to scaling-law analysis. A natural extension would be to study whether downstream task performance also follows predictable scaling curves, as has been observed in some NLP settings (Wei et al., 2022; Isik et al., 2024).

9 Limitations and Future Work

Confounded regime comparison. The two regimes differ in both gene vocabulary size (V) and dataset size (D) simultaneously. This confounding prevents us from isolating the effect of each factor. Future work should construct a matrix of experiments varying V and D independently, *e.g.* by building $V = 512$ and $V = 1{,}024$ datasets at matched cell counts (10k, 50k, 200k each).

MSE as loss metric and derived entropy. We use MSE as both the training loss and the scaling-law target. While we provide a preliminary entropy estimate of $\sim$2.30 bits per masked position for Regime A (Section 7.4), this estimate relies on a Gaussian assumption applied post hoc to MSE values, rather than a natively trained information-theoretic loss. The Gaussian NLL values in our analysis are derived from MSE, not independently logged, so the two entropy derivations are not fully independent. To obtain a rigorous transcriptomic entropy estimate, future work should train with a parametric Gaussian NLL (predicting both μ and σ^2) or a discretized cross-entropy over binned expression levels. Additionally, the full $L(D)$ and $L(C_{\min})$ intersection experiment described by Hoffmann et al. (2022) is needed to disentangle data-limited from model-limited floors.

Limited model size range. While our seven presets span five orders of magnitude, the "useful" range for Regime A is narrower: the XXS model (533 parameters) is too small to be informative about neural scaling behaviour, and models beyond L show no improvement due to the data constraint. A denser sweep in the 10^5–10^7 range, with more seeds, would yield tighter confidence intervals on α and c.

Training compute as a scaling axis. We have studied scaling with respect to parameters only. A complete picture requires also fitting $L(C)$ as a function of compute C (measured in FLOPs or tokens processed), and $L(D)$ as a function of dataset size. Our runs do record tokens processed (steps $\times$ effective batch $\times V$), but the current analysis does not fit compute scaling laws, which is deferred to future work.

Heterogeneous training configurations. The Regime B runs used varying training durations and batch sizes, introducing noise that obscures potential scaling trends. Future sweeps should use strictly standardized hyperparameters across all model sizes within a regime.

Generalization to other tissues, species, and modalities. Our data are drawn from a single Census subset. It remains to be seen whether the observed scaling exponents and floors generalize across tissue types, developmental stages, species, or multi-omic modalities (ATAC-seq, spatial transcriptomics).

10 Conclusion

We have presented the first systematic study of neural scaling laws for masked-reconstruction transformers on single-cell RNA-seq data. Our principal finding is that scaling behaviour analogous to that observed in language modelling does emerge in single-cell transcriptomics, but only when sufficient data are available relative to model capacity. In a data-rich regime (200,000 cells, 512 genes), validation MSE follows a power law in parameter count with exponent $\alpha \approx 0.27$ and an identifiable irreducible floor of $c \approx 1.44$, achieving $R^2 = 0.86$. In a data-limited regime (10,000 cells, 1,024 genes), scaling is absent ($\alpha \approx 0.009$, $R^2 = 0.02$), confirming that data scarcity is the binding constraint. A preliminary conversion of the asymptotic loss floor to information-theoretic units yields an estimate of approximately 2.30 bits of entropy per masked gene position—a first quantitative benchmark for the predictability of gene expression from co-expression context.

These results establish a quantitative framework for reasoning about the scaling behaviour of single-cell foundation models. They suggest that the field's current emphasis on increasing model size should be balanced by equivalent investment in curating larger, more diverse training corpora. With the addition of natively

trained information-theoretic loss functions, matched experimental conditions, and comprehensive compute accounting, this framework can be refined to produce rigorous estimates of the intrinsic entropy of the transcriptome—a quantity of fundamental biological and information-theoretic interest.

Reproducibility Statement

All code, data-building scripts, training loops, and analysis pipelines are publicly available at `https://github.com/[REDACTED]/scaling` (link anonymized for review; will be deanonymized upon acceptance). Datasets are constructed deterministically from the CELLxGENE Census with fixed random seeds (seed 42 for data splits, seeds 7/8/9 for model training). The full set of training run outputs, including checkpoints, history logs, and metadata, is preserved in the repository. Detailed hyperparameter configurations for all runs are provided in Appendix B. All experiments were conducted on consumer GPU hardware, making them accessible for independent replication.

References

Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Aidan Clark, Diego de Las Casas, Adria Guy, Arthur Sherburn, Daniel Sherburn, and Simon Sherburn. Unified scaling laws for routed language models. In *International Conference on Machine Learning (ICML)*, 2022.

Haotian Cui, Chloe Wang, Hassaan Maan, Kuan Pang, Fengning Luo, Nidhi Doshi, and Bo Wang. scGPT: Toward building a foundation model for single-cell multi-omics using generative AI. *Nature Methods*, 21(8): 1470–1480, 2024.

CZI Single-Cell Biology. CZ CELLxGENE Discover: A single-cell data platform for scalable exploration, analysis, and modeling of aggregated data. *bioRxiv*, 2023. doi: 10.1101/2023.10.30.563174.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, 2019.

Adam Gayoso, Romain Lopez, Galen Xing, Pierre Boyeau, Valeh Valiber Pine, Mohammad Lotfollahi, Geoffrey Schiebinger, and Nir Yosef. A Python library for probabilistic analysis of single-cell omics data. *Nature Biotechnology*, 40(2):163–166, 2022.

Minsheng Hao, Jing Gong, Xin Zeng, Chiming Liu, Yucheng Guo, Xingyi Cheng, Taifeng Wang, Jianzhu Ma, Xuegong Zhang, and Le Song. Large-scale foundation model on single-cell transcriptomics. *Nature Methods*, 21(8):1481–1491, 2024.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (GELUs). *arXiv preprint arXiv:1606.08415*, 2016.

Joel Hestness, Sharan Narang, Newsha Ardalani, Gregory Diamos, Heewoo Jun, Hassan Kianinejad, Md Mostofa Ali Patwary, Yang Yang, and Yanqi Zhou. Deep learning scaling is predictable, empirically. *arXiv preprint arXiv:1712.00409*, 2017.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Adria Guy, Simon Osindero, Karen Simonyan,

Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. Training compute-optimal large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Berivan Isik, Natalia Ponomareva, Hussein Hazimeh, Dimitris Paparas, Sergei Vassilvitskii, and Sanmi Kale. Scaling laws for downstream task performance of large language models. In *Findings of ACL*, 2024.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Romain Lopez, Jeffrey Regier, Michael B. Cole, Michael I. Jordan, and Nir Yosef. Deep generative modeling for single-cell transcriptomics. *Nature Methods*, 15(12):1053–1058, 2018.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019.

Niklas Muennighoff, Alexander Rush, Boaz Barak, Teven Le Scao, Aleksandra Piktus, Nouamane Tazi, Sampo Pyysalo, Thomas Wolf, and Colin Raffel. Scaling data-constrained language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Aviv Regev, Sarah A. Teichmann, Eric S. Lander, Ido Amit, Christophe Benoist, Ewan Birney, Bernd Bodenmiller, Peter Campbell, Piero Carninci, Menna Clatworthy, et al. The Human Cell Atlas. *eLife*, 6: e27041, 2017.

Tim Stuart, Andrew Butler, Paul Hoffman, Christoph Hafemeister, Efthymia Papalexi, William M. Mauck, Yuhan Hao, Marlon Stoeckius, Peter Smibert, and Rahul Satija. Comprehensive integration of single-cell data. *Cell*, 177(7):1888–1902, 2019.

Christina V. Theodoris, Ling Xiao, Anant Chopra, Mark D. Chaffin, Zeina R. Al Sayed, Matthew C. Hill, Heloise Manberola, Esther Sayed, Jamie Karber, Ellinor Ellinor, and Patrick T. Ellinor. Transfer learning enables predictions in network biology. *Nature*, 618:616–624, 2023.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. Emergent abilities of large language models. *Transactions on Machine Learning Research*, 2022.

Ruibin Xiong, Yunchang Yang, Di He, Kai Zheng, Shuxin Zheng, Chen Xing, Huishuai Zhang, Yanyan Lan, Liwei Wang, and Tie-Yan Liu. On layer normalization in the transformer architecture. In *International Conference on Machine Learning (ICML)*, 2020.

Fan Yang, Wenjin Wang, Fang Wang, Yuan Fang, Duyu Tang, Junzhou Huang, Hui Lu, and Jianhua Yao. scBERT as a large-scale pretrained deep language model for cell type annotation of single-cell RNA-seq data. *Nature Machine Intelligence*, 4:852–866, 2022.

Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

A On the Choice of Loss Metric

Mean squared error is the most common regression loss for continuous expression values, but it is not directly comparable to the cross-entropy losses used in language model scaling laws. Ideally, one would train with a

Gaussian negative log-likelihood where the model predicts both a mean μ_g and a variance σ_g^2:

$$\mathcal{L}_{\text{NLL}} = \frac{1}{|\mathcal{M}|} \sum_{g \in \mathcal{M}} \left[\frac{(x_g - \mu_g)^2}{2\sigma_g^2} + \log \sigma_g + \frac{1}{2} \log(2\pi) \right]. \tag{6}$$

This formulation measures loss in nats (convertible to bits by dividing by $\ln 2$) and would allow the asymptotic floor c to be directly interpreted as an entropy estimate.

In the current study, we derive Gaussian NLL values post hoc from MSE by assuming homoscedastic Gaussian residuals with variance $\sigma^2 = \text{MSE}$:

$$\hat{\mathcal{L}}_{\text{NLL}} = \tfrac{1}{2} \ln(2\pi \cdot \text{MSE}) + \tfrac{1}{2}. \tag{7}$$

This derivation is exact if the prediction errors are truly $\mathcal{N}(0, \text{MSE})$, but it conflates aleatoric and epistemic uncertainty and cannot capture heteroscedastic noise across genes. In our 18 canonical Regime A runs, all Gaussian NLL values carry the provenance label `derived_from_best_mse` in the analysis pipeline, confirming that they are not independently measured.

The entropy conversion from each floor then proceeds as follows:

$$h_{\text{NLL}} = c_{\text{NLL}} / \ln 2 = 1.592/0.693 = 2.296 \text{ bits}, \tag{8}$$

$$h_{\text{MSE}} = \tfrac{1}{2} \log_2(2\pi e \cdot c_{\text{MSE}}) = \tfrac{1}{2} \log_2(2\pi e \cdot 1.444) = 2.312 \text{ bits}. \tag{9}$$

The close agreement ($\Delta = 0.016$ bits) reflects the self-consistency of the Gaussian assumption rather than independent validation. A natively trained NLL model—where the network outputs both μ_g and $\log \sigma_g^2$—would provide a genuinely independent estimate and is the natural next step.

B Detailed Run Configuration

All runs use the following shared configuration unless otherwise noted:

- **Mask rate**: 15%
- **Optimizer**: AdamW with $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$
- **Weight decay**: 0.01
- **Gradient clipping**: max norm 1.0
- **Mixed precision**: float16 autocast on GPU
- **Activation**: GELU
- **Normalization**: Pre-LN (layer norm before attention and FFN)
- **Pooling**: mean pooling over gene tokens
- **Prediction head**: single linear layer
- **Initialization**: Xavier uniform for linear weights; normal ($\sigma = 0.02$) for embeddings

Regime A (GPU) runs use physical batch size 4, gradient accumulation 8 (effective batch 32), learning rate 3.125×10^{-5}, and 60,000 steps. Regime B runs use the same effective batch size but vary in total steps (20,000–60,000).