Toward a Coherent Virtual Cell Model: Probing Biological World-Model Coherence in Transcriptomic Foundation Models

Noa Moriel
IBM Research
The Hebrew University of Jerusalem
{noa.moriel}@mail.huji.ac.il

Yishai Shimoni
IBM Research
{YISHAIS}@il.ibm.com

Michal Rosen-Zvi
IBM Research
{ROSEN}@il.ibm.com

Michael M. Danziger
IBM Research
{Michael.Danziger}@ibm.com

Abstract

Transcriptomic foundation models (TFMs) promise to act as virtual cell models, but it remains unclear whether they have internalized the biological rules of transcriptomic space. To address this question, we propose assessing the quality of pretrained TFMs by probing the coherence of their internal world model using the pretraining loss on synthetic samples. Our approach combines two complementary tests. First, as a stress test of plausibility, we compare pretraining loss on shuffled cells compared to real samples. Second, to probe the coherence of the internal world model, we evaluate interpolated samples both within and between cell types, quantifying whether the model identifies coherent clusters. Across multiple datasets, TFMs tend to distinguish real and shuffled cells, with entropy of expression value strongly predicting the loss gap. Interpolations reveal "loss barriers" between distant cell types while similar cell types tend not to have barriers. Interestingly, much of the structure of cell embeddings persists despite the shuffling of the values of expressed genes. This approach demonstrates that quantification of an internal world model is possible, even in a "zero resource" setting, without labeled data. We argue that this is a critical step toward identifying whether TFMs can truly function as virtual cell models, rather than stochastic parrots.

1 Introduction

Transcriptomic foundation models (TFMs), large-scale neural networks pretrained on massive single-cell gene expression datasets, are rapidly emerging as powerful tools in single-cell biology (Szałata et al., 2024; Zhang et al., 2025; Guo et al., 2025). By drawing an analogy between sentences in natural language and cells in transcriptomic space, TFMs leverage transformer architectures to capture

Preprint. Under review.

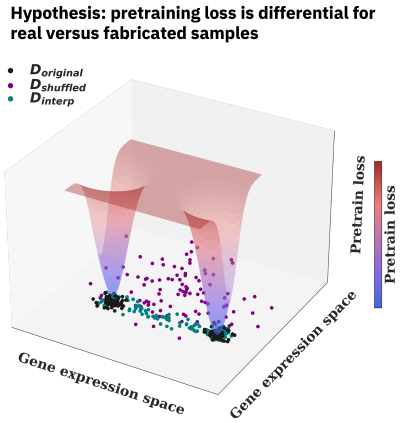

Figure 1: Our hypothesis. A transcriptomic foundation model pretrained on real data should assign lower pretraining error to genuine cell states than to fabricated ones. We compare the pretraining loss of three inputs: **original data**, D_{original}; shuffled data, D_{shuffled}, where gene identities are permuted; and interpolated data, D_{interp}, where synthetic cells are generated between types. Illustration shows our expectation of pretraining loss (z-axis) across projected expression space (xy-plane).

complex gene-gene dependencies across millions of cells. A notable example is scGPT, which learns to generate and embed cellular gene expression profiles in ways that align with biological structure (Cui et al., 2024). These models have already been applied to diverse downstream tasks, including cell type annotation, batch integration, perturbation-response prediction, and gene network inference (Liu et al., 2023; Wenteler et al., 2024; Bendidi et al., 2024; Kedzierska et al., 2025; Boiarsky et al., 2024).

However, despite their promise, evaluating how well TFMs actually understand cellular biology remains a challenge (Szałata et al., 2024). Current evaluations fall into two main categories. First, zero-shot evaluations typically examine the quality of model embeddings, for example in terms of cell-type separability or batch mixing scores such as AvgBio and AvgBatch. While scGPT and related models often produce embeddings consistent with known cell types, simple baselines using highly variable genes can achieve comparable or even superior results, raising questions about what the model truly contributes (Kedzierska et al., 2025). Moreover, naive zero-shot evaluation of the pretraining loss has shown poor calibration, raising further questions about whether the model has learned non-trivial patterns. Second, task-specific fine-tuning assesses TFMs as feature extractors or initializations for downstream models. Here too, results are mixed: in some cases fine-tuned TFMs improve prediction, but in others, trivial heuristics outperform them (Boiarsky et al., 2024; Ahlmann-Eltze et al., 2025). For example, when predicting double perturbation effects, a simple additive rule (summing individual perturbation effects and subtracting the control) surpassed fine-tuned scGPT (Ahlmann-Eltze et al., 2025). Additionally, TFM evaluation currently lacks "zero resource" methods, as each of these evaluations requires labeled sets of samples. It therefore remains unclear the extent to which TFMs are learning a coherent world model rather than behaving like a stochastic parrot (Bender et al., 2021). We define a "world model" for scRNA as an internal loss landscape that captures the statistical and relational organization of single-cell biology, assigning lower loss to biologically plausible profiles and higher loss between more distinct cell types/states.

For language models, the debate about whether the LLMs produce a true internal world model or function as stochastic parrots has raged for several years. Despite the extraordinary leap in abilities across domains that these models show, the absence of a verifiable internal world model makes the question of whether they are truly reasoning unresolved (Shojaee et al., 2025). Most studies exploring this problem for LLMs make use of the fact that the input and output text can be evaluated by humans for its logical consistency, and signs of hallucinations can be flagged by human inspection. For TFMs there is no such option, as the raw scRNA samples are not directly legible.

In this work, we introduce a set of tests to probe the coherence of the internal model of a TFM. We expect a TFM with a coherent world model (Fig. 1) :

1. Lower pretraining loss for a genuine sample compared to its randomly shuffled counterpart.

2. Lower pretraining loss for interpolated samples within a cell type than between cell types.

The first condition represents sample-level coherence, whereas the second condition requires a more global coherence and also requires that the cell type labels in the test data be assigned correctly.

Our contribution is to define the above TFM world model evaluation, and to deploy it on scGPT Cui et al. (2024) with six widely used benchmarking datasets Table 1. Though focus on scGPT for concreteness in this study, because the method does not make any specific assumptions about the details of the TFM, the same approach can be applied to any model. We find that there is evidence for the emergence of a coherent world model but it is far from universal across the samples considered. Specifically, many samples are nearly indistinguishable from random and in general only significantly different cell types are accompanied by a detectable loss barrier. Furthermore, we find that for embedding quality, as long as the gene and expression levels are shuffled among expressed genes only, much of the embedding structure is preserved.

Through these analyses, we go beyond dataset-level benchmarks and provide a direct, unsupervised measure of how well a foundation model has captured the statistical and biological regularities of single-cell data. Our approach complements embedding-based evaluations by focusing on the model's own pretraining objective, offering a "zero-resource" diagnostic of what the model truly knows.

2 Methods and Results

2.1 Shuffled profiles.

By permuting gene identities we generate inputs that preserve some statistics (e.g., the set of expressed genes) while disrupting others (e.g., the mapping between values and gene labels). Our expectation parallels the expectation of a language model to assign lower perplexity (or higher masked-token accuracy) to coherent sentences than to word-shuffled sequences.

We first examine how the shuffling of gene identities affects the pretraining loss and the resulting cell embeddings (Fig. 2). We evaluated scGPT pretrained on human gene expression, focusing on two scenarios: (1) shuffling all gene identities (Fig. 2.A) and (2) shuffling only the identities of expressed genes (Fig. 2.D). We consider the loss:

$$\mathcal{L}(x) = \mathcal{L}_{\text{GEP}}(x) + \mathcal{L}_{\text{GEPC}}(x),$$

where x is the gene expression profile, $\mathcal{L}_{\text{GEP}}$ is the mean squared error (MSE) of masked expression values predicted from the transformer stack, and $\mathcal{L}_{\text{GEPC}}$ is the MSE of masked values predicted from the cell embeddings and gene representations (see Cui et al., 2024 and Appendix A.2). Evaluation datasets were drawn from single batches of the scEval benchmark datasets(Liu et al., 2023) (see Appendix A.3).

Across both shuffling strategies, pretraining loss increased for shuffled profiles relative to their original counterparts (Fig. 2.B,E). However, shuffled cells did not always incur higher error than the real profiles from which they were derived, indicating that the loss function alone is not a perfect discriminator.

For language models, we observe that shorter samples have similar loss whether or not the words are shuffled Fig. 3. scRNA samples do not vary by their length, but they vary widely in their information content. In contrast to text, it is possible in principle to have a sample where all genes were measured in the same quantity. For such samples, shuffling is irrelevant and success at masked value prediction is not informative. To assess the impact of this phenomenon in our measurement, we measured the sample entropy as:

$$H(x) = -\sum_{k=0}^{\infty} p(x=k) \log p(x=k),$$

where $p(x=k)$ is the frequency of expression value k. While expression entropy has been linked to biological properties such as pluripotency (Guo et al., 2017; Teschendorff & Enver, 2017; Liu et al., 2020; Ye et al., 2020), here we focus on its ability to quantify the amount of learnable information in the gene expression profile. Unlike natural language, where the tokens of the input are diverse, scRNA-seq profiles resemble negative binomial distributions dominated by low counts (Luecken & Theis, 2019). Low-entropy cells therefore provide limited signal, making trivial predictions sufficient

to achieve low loss during training. The shuffling procedure is also expected to have a reduced effect in low-entropy cells. To quantify this, we compared shuffle-to-original loss gaps against entropy (Fig. 4). While PBMC cells showed a strong correlation ($r = 0.45$), others such as lung cells showed almost none ($r = 0.05$), suggesting that the shuffle gap captures more than just raw information content.

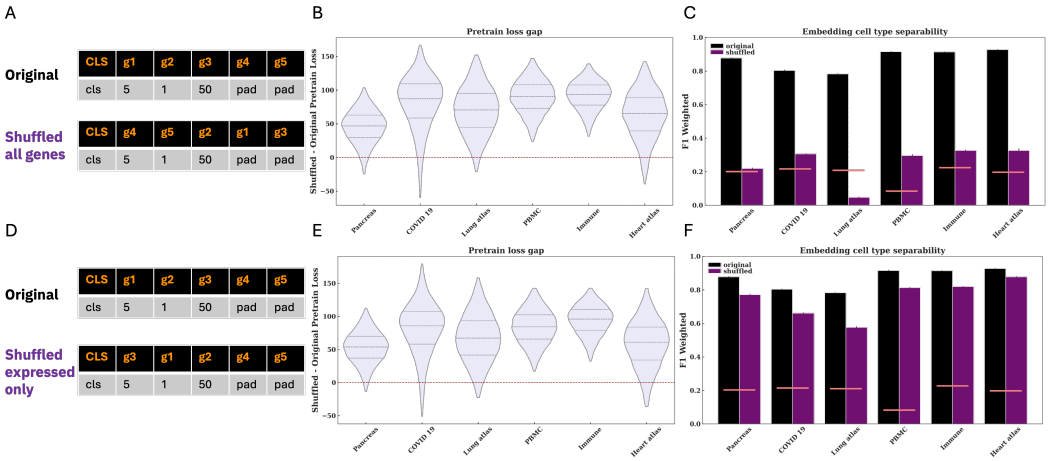

Figure 2: Shuffling expressed genes disrupts pretraining loss but variably affects embeddings. *A,D,* examples of original (top) and shuffled (bottom) inputs for all genes (A) and expressed genes only (D). *B,E,* distribution of pretraining loss gaps (median of shuffled minus median of original over 10 repeats). Dashed black lines mark quartiles, red dashed line marks zero. Median gaps are significantly positive (Bonferroni-corrected Wilcoxon test). *C,F,* separability of cell embeddings, measured as F1 scores of logistic regressors trained on embeddings. Error bars are over 10 repetitions of the analysis. Original (black), shuffled (purple), and stratified baseline (coral dash) are shown. All results are from single batches of scEval datasets (Liu et al., 2023).

Together, these findings demonstrate that with an appropriate setup, pretraining loss can distinguish real from permuted data. While embeddings are commonly used to evaluate TFMs, the loss function itself provides a lightweight probe of model fidelity. Furthermore, it is "zero-resource" in that it does not require cells to be labeled and it is usable on the sample-level. This perspective also points to practical applications: cells with unusually high entropy but low shuffle-to-original loss gap may reflect artifacts or out-of-distribution states and warrant reduced confidence, analogous to quality control in experimental workflows (Luecken & Theis, 2019).

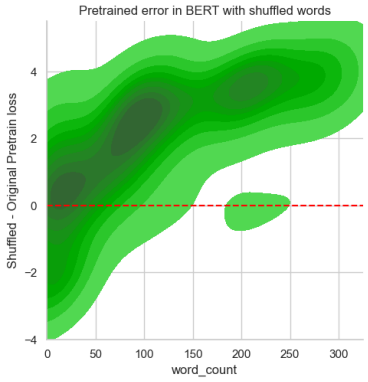

Figure 3: Comparison of genuine and shuffled loss in a pretrained text model. Using the "bert-base-uncased" model and "Salesforce/wikitext" corpus, we examined the measured pretraining loss for random paragraphs of various lengths compared with word-shuffled samples. We see that for shorter samples, the pretrained loss is not significantly different, but for longer samples the difference is significant and persistent. Results shown for 50 samples, shuffled 12 times and masked 10 times at a ratio of 0.3, with a max number of 128 tokens input for all samples.

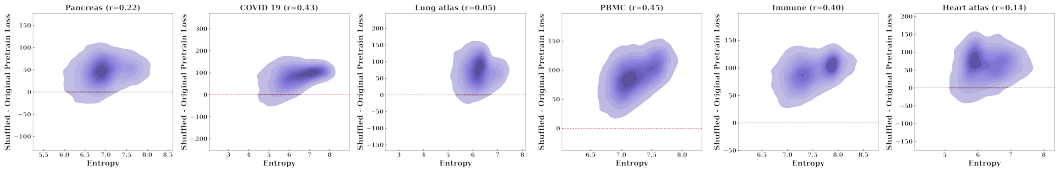

Figure 4: Samples with higher entropy tend to have a larger difference between all-gene-shuffled and original loss. Sample distribution by the difference in pretraining loss (median of shuffled minus median of original over 10 repeats) as a function of sample entropy. Strength of correlation varies across datasets (Pearson r reported in captions). Shown here for all-gene shuffling; expressed-gene shuffling gives similar results (Appendix Fig. 6).

The cell-level embeddings produced by TFMs have been shown to emphasize biological similarity and reduce batch effects and represent one of the most promising uses of this methodology Liu et al. (2023). However, model-free embeddings based on dimensional reduction through limiting to highly variable genes and/or standard methods such as PCA, UMAP or tSNE also produce good cell-level embeddings, again raising questions as to whether the cell representation represents an internal world model or encodes basic correlations. We thus explore the effect of shuffling on the quality of the cell-level embeddings. We do so by examining separability of the embedding vectors into ground truth cell types before and after shuffling. Separability of cell types was assessed by training logistic regression classifiers on embeddings and reporting class-weighted F1 scores, relative to both the original embeddings and a stratified baseline (Fig. 2.C,F). Results differed markedly by strategy: shuffling all gene identities severely disrupted embedding structure, reducing separability close to baseline levels, while shuffling only expressed gene identities largely preserved separability. This reflects similar findings on the significance of binary expression representations **?** and models such as UCE Rosen et al. (2024) which are trained with a strictly binary objective. Further inspection of the immune dataset (Appendix Fig. 7) confirmed that major types such as $CD4^+$ T cells and $CD14^+$ monocytes remain separable when only expressed genes are shuffled, but not when all genes are shuffled. This suggests that in some cases cell representations of scGPT primarily rely on the *set of expressed genes*, analogous to marker genes, rather than the expression values.

In general, we find that scGPT assigns higher loss to shuffled than to original cells, with the magnitude of this gap partially explained by the sample entropy (the entropy of a cell's gene expression values). Low-entropy cells provide little signal, making them harder to learn and to distinguish from their shuffled counterparts. Moreover, comparing loss and embeddings across shuffle strategies reveals a distinction in what the model encodes: pretraining loss, which encapsulates the reconstruction of gene expression values, is sensitive to expression values, while cell embeddings are strongly affected by the set of expressed genes, not always sensitive to their precise values.

2.2 Interpolated profiles.

We also generate synthetic cells by linearly combining expression profiles from different cell types. Here we ask whether scGPT identifies interpolations between clusters as implausible, while treating within-cluster interpolations as realistic. To quantify this, we approximate the convexity of the pretraining loss along interpolation paths: convex increases in loss indicate "barriers" between states. Whereas smooth interpolation between samples is not defined for text, transcriptomic space is continuous, making interpolation a natural probe of the learned manifold. We find that in general, pretraining loss is stable and homogeneous within clusters, but rises across distinct cell-type groups, revealing significant learned barriers in expression space.

We do this by interpolating expression values between distinct cell types and assessing the pretrained loss on the interpolated samples (Fig. 5). Biologically, not all combinations of expression values are plausible. If a TFM has learned the underlying distribution, it should assign lower loss to samples within clusters (corresponding to empirically observed regions of transcriptomic space), and higher loss to interpolations between unrelated clusters. This represents a distinct way of evaluating the pretrained model's knowledge of cell-types, that does not depend on embeddings or clustering metrics. Unlike the shuffling method above, evaluating the pretraining loss within and between cell-types is not "zero-resource"–it relies on accurate cell-type labels.

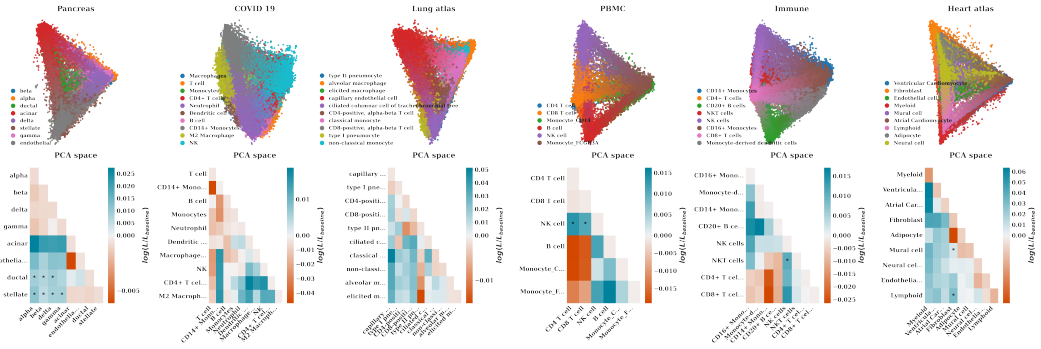

Figure 5: Interpolations reveal barriers between clusters. *A,* schematic of interpolated samples between two cell types, shown in PCA space with blended colors according to interpolation factor α. *B,* average log-ratio of interpolated loss relative to baseline. Values near zero (main diagonal) reflect within-cluster consistency. Positive values across clusters indicate convex loss profiles, i.e. barriers. Significance is determined by Bonferroni-corrected t-tests. For visibility, the 10 most common cell type clusters are shown for each dataset.

For each pair of cell types ct_1, ct_2, we randomly sample pairs of cells $x_i \in ct_1$ and $x_2 \in ct_2$ and draw $\alpha \sim U(0, 1)$ to compute interpolated profiles:

$$x' = \alpha x_i + (1 - \alpha) x_j,$$

producing samples along the chord connecting the pair of cells across cell types in expression space (Fig. 5.A). Repeating this across many pairs and α values produces a dense set of synthetic observations upon which we evaluate the pretraining loss $L(x')$, in comparison to a baseline loss defined as the weighted average of the two cell-type losses:

$$L_{\text{baseline}}(x') = \alpha \cdot L_{\text{ct}_1} + (1 - \alpha) \cdot L_{\text{ct}_2}.$$

The log-ratio

$$r(x') = \log \frac{L(x')}{L_{\text{baseline}}(x')}$$

quantifies whether interpolations incur higher-than-expected loss.

When interpolating between cell-type clusters, we frequently find positive log-ratios, reflecting the existence of a "loss-barrier", an ostensibly forbidden zone of transcriptomic space separating the cell types(Fig. 5.B). However, these barriers were not always observed. For example, in the pancreas dataset (Tran et al., 2020), we observed strong barriers when interpolating between an endocrine (alpha, beta, delta, gamma) and non-endocrine (acinar, endothelial, ductal, stellate) cell type, but did not find barriers when interpolating between different endocrine or non-endocrine cell-types. This is a biological coherent result, reflecting the fact that transcriptomically the endocrine and non-endocrine cells are very different whereas the different endocrine cells are very similar.

We execute the same interpolation *within* cell type clusters. The interpolated cells within a cluster are expected to have loss similar to the real cells within the cluster. In general, we find log-ratios were near zero, indicating stability of the loss function. It is possible that significant differences in loss between interpolated and real cells would arise within a set of samples of a single cell-type. This would reflect a misalignment of the cell-type label and the pretrained model, arising either from a single cell-type label being applied to a hetereogeneous population or from limitations of the pretrained model. In fact, the presence of loss-barriers within a single population could be utilized to identify subtypes without relying on observed sample density, the feature underlying unsupervised clustering methods. In our test datasets Table 1, we did not observe statistically significant loss heterogeneity within cell-type clusters.

3 Discussion

Our work introduces an unsupervised evaluation framework for TFMs, based on comparing pre-training loss across real and synthetic samples. Using scGPT as a case study, we showed that the

pretraining loss reliably distinguishes genuine gene expression profiles from their shuffled counterparts, while intermediate embeddings remain robust when only expression identities rather than values are disrupted. At the regional level, interpolation analysis revealed loss barriers between cell type clusters, consistent with biologically implausible transitions, while maintaining homogeneity within clusters. Together, these results highlight pretraining loss as an interpretable and zero-resource diagnostic of model familiarity and biological plausibility. And because the method makes use only of synthetic data and pretraining loss evaluation, with no assumptions about the details of the TFM, it can be applied across models, consistently.

Our approach touches on broader concepts at the intersection of biology and machine learning. Do TFMs act as world models of gene regulation, or merely stochastic parrots that recapitulate simple correlations? We find that for this TFM (scGPT) and these datasets Table 1 the shuffled sample loss tends be higher than the genuine sample though this is by no means universal–many shuffled samples have lower loss than the genuine samples, especially in the lower entropy regions. This is in contrast to text models, which show a more consistent gap in loss between shuffled and genuine samples Fig. 3. We speculate that this may be due to the "grammar" of scRNA being inherently more flexible than language, due to experimental noise, insufficient training or architectural mismatch. With interpolations, we find loss-barriers in some places but not in others. We hypothesize that a well-trained model would roughly follow the cell-type ontology, with lower barriers between ontologically similar cells and higher barriers between ontologically distant cells. The possibility that a cell-type passes through another cell-types low-loss region on the way to a third cell type must also be considered.

In practical terms, the methods illustrated above have several potential applications. They can define an "autocorrect" for transcriptomic inputs, identifying likely mistaken samples and integrated into quality-control pipelines. Identification of loss barriers could enhance the viability of in-silico gene perturbation studies Theodoris et al. (2023) by identifying the perturbations that are likely to have an effect by examining whether or not they reach high loss regions (implying the need for the other genes to adapt to the perturbation). More ambitiously, incorporating explicit training objectives that enforce smoothness (e.g. (Kim et al., 2024)) or structure in the loss landscape may enable TFMs to better capture the manifold geometry of gene expression space, thereby facilitating downstream applications in generative biology. These directions point toward TFMs not only for improving on known tasks, but evolving into genuine virtual cell models.

A Appendix

A.1 Formulation of our evaluation

Given a pretrained model $\mathcal{M}$ and a dataset $D_{\text{original}} = (x_1, \ldots, x_N)$, where $x_i \in \mathbb{R}_+^g$ represents the gene expression profile for sample i, we evaluate the model's knowledge of D_{original} by comparing its pretraining loss, $L : \mathbb{R}_+^g \to \mathbb{R}$, on three types of input: 1. the original data D_{original}, 2. shuffled data D_{shuffled}, and 3. interpolated data D_{interp}.

A.1.1 Shuffle analysis

To probe sensitivity to biologically implausible inputs, we generate shuffled datasets D_{shuffled} by randomly permuting gene identities (excluding padded indices and special tokens). We consider two settings: shuffling all gene identities or shuffling only the identities of expressed genes.

For each original cell $x \in D_{\text{original}}$, we create a shuffled counterpart $x' \in D_{\text{shuffled}}$ and compare their pretraining losses. Because both the shuffled identities and the random masking of values introduce variability, we repeat this procedure 10 times. In each repetition, a new shuffle is sampled, the pretraining loss is re-evaluated for both x and x', and the difference is summarized as the median gap across repetitions:

$$\Delta L(x) = \text{median}_{r=1,\ldots,10} \left(L^{(r)}(x') - L^{(r)}(x) \right).$$

This repetition scheme stabilizes the estimate of the shuffle-induced loss gap and ensures that results are not driven by a particular permutation or masking pattern.

We also evaluate cell embeddings $\mathcal{M}_{\text{embedding}}$. To assess their biological separability, we train a logistic regression classifier on embeddings with cell-type labels and report class-weighted F1 scores. Data is split to 80% train and 20% test set. A stratified baseline, which samples class labels according to empirical frequencies, is used as control. We run this analysis 10 times.

A.1.2 Interpolation analysis

To examine whether the model encodes barriers between clusters, we generate interpolated datasets D_{interp}. For annotated cell types ct_1 and ct_2, and samples $x_i \in \text{ct}_1$, $x_j \in \text{ct}_2$, we define an interpolated expression profile:

$$x' = \alpha \cdot x_i + (1 - \alpha) \cdot x_j, \quad \alpha \in [0, 1].$$

To normalize for inherent differences in cluster-level loss, we compute a barrier score as the log-ratio:

$$r(x') = \log \left(\frac{L(x')}{L_{\text{baseline}}(x')} \right),$$
$$L_{\text{baseline}}(x') = \alpha \cdot L_{\text{ct}_1} + (1 - \alpha) \cdot L_{\text{ct}_2},$$

where L_{ct} is the average loss over real cells of type ct.

For each pair $(\text{ct}_1, \text{ct}_2)$, we sample $N = 1000$ interpolated cells, drawing $\alpha \sim U[0, 1]$. The average region sensitivity is:

$$R_{\text{ct}_1,\text{ct}_2} = \frac{1}{N} \sum_{x \in S_{\text{ct}_1,\text{ct}_2}} r(x).$$

A value $R_{\text{ct}_1,\text{ct}_2} > 0$ indicates a convex loss profile (a "barrier") between the clusters.

A.2 scGPT model

We refer the reader to the original manuscript (Cui et al., 2024) for details of scGPT and mention here only some of its essentials.

Each cell is represented by three vectors of length M: 1. a gene token vector, 2. a binned gene expression vector ($B = 50$ bins per cell), and 3. an optional condition token vector (unused in our experiments).

A special `<CLS>` token aggregates the cell representation, and padding tokens `<pad>` ensure fixed length.

We adhere to the data preparation pipeline for producing cell embeddings by scGPT [1] where the `<CLS>` is added to the Dataset object and binning, padding and random masking are handled in the DataCollator (for embeddings we disable masking; for loss estimation, we set `mlm_probability=0.5`). To facilitate our two strategies of gene identity shuffling, we provide two optional proceeding data collators: (1) Gene identity shuffling collator which permutes gene identifiers. Parameter `shuffle_zero_gene` toggles between shuffling all genes (True) or expressed genes only (False); (2) Zero-gene exclusion which mimics scGPT's `include_zero_gene` flag, considered in the Dataset object. If `False`, genes with zero expression are removed per cell. As scGPT was pretrained using only expressed values, this setting is crucial for reproducibility.

The input to the transformer layers is:

$$h_0(\mathbf{x}) = \text{emb}_g(\mathbf{t}) + \text{emb}_x(\mathbf{x}).$$

The contextualized embedding is obtained by passing h_0 through n transformer blocks:

$$h_l = \text{transformer\_block}(h_{l-1}), \quad l = 1, \ldots, n.$$

The cell representation is the embedding at the `<CLS>` token:

$$h_c(\mathbf{x}) = h_n(\mathbf{x})[\langle \text{CLS} \rangle].$$

scGPT is trained autoregressively to reconstruct masked or unknown gene expression values (see Cui et al. (2024)). Auxiliary losses are introduced during pretraining to stabilize learning and improve biological interpretability, including:

Gene expression prediction (GEP) loss. Predicts masked expression values directly from the contextualized embedding:

$$f(\mathbf{x}) = \text{MLP}(h_n(\mathbf{x})),$$
$$\mathcal{L}_{\text{GEP}}(\mathbf{x}) = \frac{1}{|M_{\text{mask}}|} \sum_{j \in M_{\text{mask}}} \left(f(\mathbf{x})_j - \mathbf{x}_j\right)^2,$$

where M_{mask} indicates the masked positions and $|M_{\text{mask}}|$ the number of masked values.

Gene expression prediction for cell modeling (GEPC) loss. Links gene token representations to the global cell embedding:

$$q_j = \text{MLP}(\text{emb}_g(\mathbf{t}_g))_j, \quad g(\mathbf{x})_j = q_j \cdot W h_c(\mathbf{x}),$$
$$\mathcal{L}_{\text{GEPC}}(\mathbf{x}) = \frac{1}{|M_{\text{mask}}|} \sum_{j \in M_{\text{mask}}} \left(g(\mathbf{x})_j - \mathbf{x}_j\right)^2,$$

where W is a learned projection matrix.

Here, we approximate the pretraining loss L using the sum of the gene expression prediction loss and the gene expression prediction for cell modeling loss:

$$L(x) = L_{\text{GEP}}(x) + L_{\text{GEPC}}(x).$$

A.3 Datasets

We use datasets from the scEval benchmark (Liu et al., 2023). For each dataset, we evaluate the largest batch to minimize batch effects. Preprocessing follows scGPT's pipeline: filtering low-count genes, normalization, log-transformation, and selection of 1,200 highly variable genes. Only genes within scGPT's vocabulary are retained.

A.4 Extended analysis of results

[1] See code in: https://github.com/bowang-lab/scGPT/blob/main/scgpt/tasks/cell_emb.py

Dataset	Batch	Genes (K)	Cells (K)	Cell Types	Input Genes (K)
Pancreas (Tran et al., 2020)	1	15.56	8.57	13	1.17
Immune (Luecken et al., 2022)	10X	12.30	10.73	12	1.18
Heart atlas (Litviňuková et al., 2020)	AH1_Nuclei_Multiome-v1	32.73	2.33	11	1.01
PBMC (Zheng et al., 2017)	0	33.69	8.10	9	1.09
COVID-19 (Stephenson et al., 2021)	1	1.20	12.74	31	1.15
Lung atlas (Luecken et al., 2022)	Banovich_Kropski_2020	27.96	6.12	42	1.10

Table 1: Datasets used in evaluation. For each dataset, we report the largest batch, total number of genes and cells, annotated cell types, and the number of genes retained after preprocessing and restriction to scGPT's vocabulary.

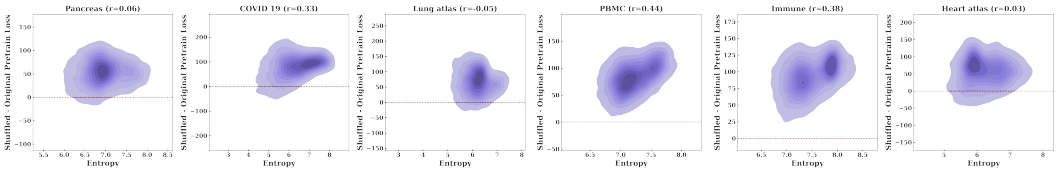

Figure 6: Entropy versus shuffle-induced loss gap where only expressed genes are shuffled. Like Fig. 4, sample distribution by the difference in pretraining loss (median of shuffled minus median of original over 10 repeats) as a function of sample entropy. Strength of correlation varies across datasets (Pearson r reported in captions).

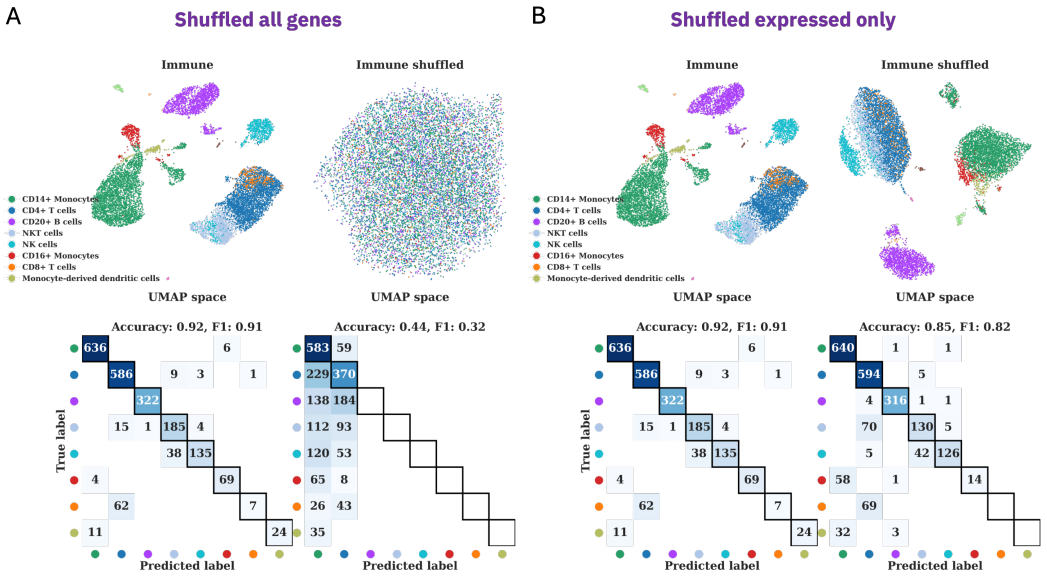

Figure 7: Effect of shuffling on immune cell embeddings. Top row: embeddings of original (left) and shuffled (right) cells colored by cell-type. Bottom row: corresponding confusion matrices of logistic regression classifiers trained on embeddings. Shuffling all genes (A) disrupts separability, while shuffling only expressed genes (B) preserves major type boundaries (e.g. $CD4^+$ T cells vs. $CD14^+$ monocytes).

References

Constantin Ahlmann-Eltze, Wolfgang Huber, and Simon Anders. Deep-learning-based gene perturbation effect prediction does not yet outperform simple linear baselines. *Nature Methods*, pp. 1–5, 2025.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pp. 610–623, 2021.

Ihab Bendidi, Shawn Whitfield, Kian Kenyon-Dean, Hanene Ben Yedder, Yassir El Mesbahi, Emmanuel Noutahi, and Alisandra K Denton. Benchmarking transcriptomics foundation models for

perturbation analysis: one pca still rules them all. *arXiv preprint arXiv:2410.13956*, 2024.

Rebecca Boiarsky, Nalini M Singh, Alejandro Buendia, Ava P Amini, Gad Getz, and David Sontag. Deeper evaluation of a single-cell foundation model. *Nature Machine Intelligence*, 6(12):1443–1446, 2024.

Haotian Cui, Chloe Wang, Hassaan Maan, Kuan Pang, Fengning Luo, Nan Duan, and Bo Wang. scgpt: toward building a foundation model for single-cell multi-omics using generative ai. *Nature Methods*, pp. 1–11, 2024.

Fei Guo, Renchu Guan, Yaohang Li, Qi Liu, Xiaowo Wang, Can Yang, and Jianxin Wang. Foundation models in bioinformatics. *National science review*, 12(4):nwaf028, 2025.

Minzhe Guo, Erik L Bao, Michael Wagner, Jeffrey A Whitsett, and Yan Xu. Slice: determining cell differentiation and lineage based on single cell entropy. *Nucleic acids research*, 45(7):e54–e54, 2017.

Kasia Z Kedzierska, Lorin Crawford, Ava P Amini, and Alex X Lu. Zero-shot evaluation reveals limitations of single-cell foundation models. *Genome Biology*, 26(1):101, 2025.

Jason Z Kim, Nicolas Perrin-Gilbert, Erkan Narmanli, Paul Klein, Christopher R Myers, Itai Cohen, Joshua J Waterfall, and James P Sethna.
Gamma-vae: Curvature regularized variational autoencoders for uncovering emergent low dimensional geometric structure in high dimensional data. *arXiv preprint arXiv:2403.01078*, 2024.

Monika Litviňuková, Carlos Talavera-López, Henrike Maatz, Daniel Reichart, Catherine L Worth, Eric L Lindberg, Masatoshi Kanda, Krzysztof Polanski, Matthias Heinig, Michael Lee, et al. Cells of the adult human heart. *Nature*, 588(7838):466–472, 2020.

Jingxin Liu, You Song, and Jinzhi Lei. Single-cell entropy to quantify the cellular order parameter from single-cell rna-seq data. *Biophysical Reviews and Letters*, 15(01):35–49, 2020.

Tianyu Liu, Kexing Li, Yuge Wang, Hongyu Li, and Hongyu Zhao. Evaluating the utilities of foundation models in single-cell data analysis. *bioRxiv*, pp. 2023–09, 2023.

Malte D Luecken and Fabian J Theis. Current best practices in single-cell rna-seq analysis: a tutorial. *Molecular systems biology*, 15(6):e8746, 2019.

Malte D Luecken, Maren Büttner, Kridsadakorn Chaichoompu, Anna Danese, Marta Interlandi, Michaela F Müller, Daniel C Strobl, Luke Zappia, Martin Dugas, Maria Colomé-Tatché, et al. Benchmarking atlas-level data integration in single-cell genomics. *Nature methods*, 19(1):41–50, 2022.

Yanay Rosen, Yusuf Roohani, Ayush Agrawal, Leon Samotorčan, Tabula Sapiens Consortium, Stephen R. Quake, and Jure Leskovec. Universal cell embeddings: A foundation model for cell biology. *bioRxiv*, 2024. doi: 10.1101/2023.11.28.568918. URL https://www.biorxiv.org/content/early/2024/10/06/2023.11.28.568918.

Parshin Shojaee, Iman Mirzadeh, Keivan Alizadeh, Maxwell Horton, Samy Bengio, and Mehrdad Farajtabar. The illusion of thinking: Understanding the strengths and limitations of reasoning models via the lens of problem complexity. *arXiv preprint arXiv:2506.06941*, 2025.

Emily Stephenson, Gary Reynolds, Rachel A Botting, Fernando J Calero-Nieto, Michael D Morgan, Zewen Kelvin Tuong, Karsten Bach, Waradon Sungnak, Kaylee B Worlock, Masahiro Yoshida, et al. Single-cell multi-omics analysis of the immune response in covid-19. *Nature medicine*, 27 (5):904–916, 2021.

Artur Szałata, Karin Hrovatin, Sören Becker, Alejandro Tejada-Lapuerta, Haotian Cui, Bo Wang, and Fabian J Theis. Transformers in single-cell omics: a review and new perspectives. *Nature Methods*, 21(8):1430–1443, 2024.

Andrew E Teschendorff and Tariq Enver. Single-cell entropy for accurate estimation of differentiation potency from a cell's transcriptome. *Nature communications*, 8(1):15599, 2017.

Christina V Theodoris, Ling Xiao, Anant Chopra, Mark D Chaffin, Zeina R Al Sayed, Matthew C Hill, Helene Mantineo, Elizabeth M Brydon, Zexian Zeng, X Shirley Liu, et al. Transfer learning enables predictions in network biology. *Nature*, 618(7965):616–624, 2023.

Hoa Thi Nhu Tran, Kok Siong Ang, Marion Chevrier, Xiaomeng Zhang, Nicole Yee Shin Lee, Michelle Goh, and Jinmiao Chen. A benchmark of batch-effect correction methods for single-cell rna sequencing data. *Genome biology*, 21:1–32, 2020.

Aaron Wenteler, Martina Occhetta, Nikhil Branson, Magdalena Huebner, Victor Curean, WT Dee, WT Connell, Alex Hawkins-Hooker, Siu Pui Chung, Yasha Ektefaie, et al. Perteval-scfm: benchmarking single-cell foundation models for perturbation effect prediction. *bioRxiv*, pp. 2024–10, 2024.

Yusong Ye, Zhuoqin Yang, Meixia Zhu, and Jinzhi Lei. Using single-cell entropy to describe the dynamics of reprogramming and differentiation of induced pluripotent stem cells. *International Journal of Modern Physics B*, 34(30):2050288, 2020.

Fan Zhang, Hao Chen, Zhihong Zhu, Ziheng Zhang, Zhenxi Lin, Ziyue Qiao, Yefeng Zheng, and Xian Wu. A survey on foundation language models for single-cell biology. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 528–549, 2025.

Grace XY Zheng, Jessica M Terry, Phillip Belgrader, Paul Ryvkin, Zachary W Bent, Ryan Wilson, Solongo B Ziraldo, Tobias D Wheeler, Geoff P McDermott, Junjie Zhu, et al. Massively parallel digital transcriptional profiling of single cells. *Nature communications*, 8(1):14049, 2017.