Sparse Autoencoders Reveal Interpretable Cell-Type Programs in Single-Cell Foundation Model Representations

Ihor Kendiukhov[1]
[1]Department of Computer Science, University of Tübingen, Germany
kendiukhov@gmail.com

Abstract

Single-cell foundation models such as scGPT learn rich representations of cellular identity, yet the biological programs encoded in their internal activations remain opaque. We apply sparse autoencoders (SAEs)—a mechanistic interpretability technique from AI safety research—to decompose the residual-stream activations of a pre-trained scGPT model into sparse, interpretable features. Using 1,000 human cells spanning diverse immune and stromal populations from the Tabula Sapiens atlas, we extract activations from all 12 transformer layers and train SAEs at multiple sparsity levels. We show that appropriately regularised SAEs ($\lambda \geq 1$) achieve genuine sparsity ($L_0 \approx 48$–54) while maintaining high reconstruction fidelity ($R^2 > 0.76$), in contrast to weakly regularised SAEs that activate the majority of dictionary elements simultaneously. SAE features trained on later layers recover biologically coherent programs aligned with annotated cell types, including distinct features for B cells, T cell subsets, macrophages, and neutrophils. Gene-level attribution analysis reveals that individual features are enriched for canonical immune marker gene sets (Fisher's exact test, FDR < 0.05). We find that SAE features provide superior cell-type discrimination compared to matched-dimensionality PCA, with higher AUROC for cell-type classification from individual features. This work demonstrates that mechanistic interpretability methods developed for large language models transfer productively to biological foundation models, providing a principled approach to understanding what these models learn about cellular identity.

Keywords: sparse autoencoders, single-cell foundation models, mechanistic interpretability, scGPT, transcriptomics, cell type

1 Introduction

Single-cell RNA sequencing (scRNA-seq) has revealed the remarkable transcriptomic diversity of human tissues, defining cell-type taxonomies at ever-finer resolution [The Tabula Sapiens Consortium, 2022]. Foundation models trained on millions of single-cell profiles—including scGPT [Cui et al., 2024] and Geneformer [Theodoris et al., 2023]—have emerged as powerful tools for tasks ranging from cell-type annotation to perturbation prediction. These models encode cellular identity into dense, high-dimensional embedding spaces, achieving state-of-the-art performance across diverse genomic benchmarks.

Despite their predictive success, these models remain fundamentally opaque. The representations that drive their performance are distributed across hundreds of dimensions in ways that resist straightforward biological interpretation. A given unit in the model's hidden layers may respond to a superposition of unrelated biological programs [Elhage et al., 2022], making it difficult to determine what the model has learned about gene regulation, cell-type identity, or pathway activity from inspection of individual neurons.

This interpretability challenge mirrors a central concern in AI safety. In the context of large language models (LLMs), the field of *mechanistic interpretability* has developed tools to reverse-engineer the internal representations of trained neural networks [Olah et al., 2020]. Sparse autoencoders (SAEs) have emerged as a particularly effective approach: by training a sparsity-constrained autoencoder on a model's internal activations, one can decompose superposed representations into a dictionary of monosemantic features, each corresponding to a single interpretable concept [Bricken et al., 2023, Cunningham et al., 2023, Templeton et al., 2024].

We propose that SAEs can serve as a principled framework for extracting interpretable biological features from single-cell foundation models. The core hypothesis is that the superposition hypothesis [Elhage et al., 2022]—which posits that neural networks represent more concepts than they have dimensions by encoding them in overlapping patterns—applies to biological foundation models as well. A model trained on diverse transcriptomic data likely represents cell-type programs, pathway activities, and regulatory states as superposed features within its activation space. SAEs can disentangle these into a sparse, overcomplete basis where each feature corresponds to a distinct biological program.

In this study, we apply SAEs to activations from a pre-trained scGPT model [Cui et al., 2024] processing human cells from the Tabula Sapiens atlas [The Tabula Sapiens Consortium, 2022]. We systematically characterise the recovered features across model layers and sparsity levels, assessing their alignment with known cell-type markers and biological gene sets. Our central questions are:

1. Can SAEs trained on single-cell foundation model activations recover biologically coherent features?
2. Does sparsity level critically affect feature interpretability, and what regularisation strength is required?
3. Do SAE features capture cell-type identity, and at what hierarchical resolution?
4. Do SAE features provide interpretive advantages over standard dimensionality reduction?

This work contributes to a growing dialogue between AI safety and computational biology. By demonstrating that interpretability methods developed for LLMs transfer to biological foundation models, we establish a new toolkit for understanding what these models learn—and, conversely, provide the interpretability community with a domain where ground-truth biological annotations enable rigorous validation of their methods.

2 Related Work

2.1 Single-Cell Foundation Models

Transformer-based foundation models for single-cell genomics learn generalisable representations of cellular identity from large-scale transcriptomic data. scGPT [Cui et al., 2024] adapts generative pre-training to single-cell data, treating expression profiles as sequences of gene tokens with associated values; pre-trained on over 33 million cells, it achieves strong performance on cell-type annotation, batch integration, and perturbation prediction. Geneformer [Theodoris et al., 2023] rank-orders genes by expression and pre-trains a BERT-style transformer on approximately 30 million cells, enabling zero-shot transfer across tissues and species. Other models include scBERT [Yang et al., 2022] and CellLM [Zhao et al., 2024]. While these models differ in architecture and tokenisation, they share a common limitation: the biological knowledge encoded in their representations is not directly accessible.

2.2 Mechanistic Interpretability and Sparse Autoencoders

Mechanistic interpretability aims to understand neural networks by identifying interpretable computational structure [Olah et al., 2020]. The superposition hypothesis [Elhage et al., 2022] formalises the observation that networks represent more features than they have dimensions,

Table 1: Cell-type composition of the dataset. Types with ≥ 5 cells are shown.

Cell type	n	**Cell type**	n
B cell	189	Monocyte	42
$CD4^+$ T cell	180	Erythrocyte	30
Macrophage	127	NK cell	28
Neutrophil	116	Classical monocyte	25
$CD8^+$ T cell	112	Naïve $CD4^+$ T cell	20
Plasma cell	44	Other (< 20 each)	87

explaining neuronal polysemanticity. SAEs have emerged as a scalable solution: Cunningham et al. [2023] showed that SAEs recover more interpretable features than individual neurons in language models; Bricken et al. [2023] systematically recovered monosemantic features from a one-layer transformer; and Templeton et al. [2024] scaled SAEs to Claude 3 Sonnet, demonstrating interpretable feature recovery in production-scale LLMs.

2.3 Interpretability in Biological Models

Existing approaches to interpreting biological foundation models have focused on attention weight analysis [Cui et al., 2024, Theodoris et al., 2023], which has well-documented limitations: attention weights do not straightforwardly correspond to feature importance and provide only indirect evidence about learned representations. SAEs offer a complementary approach that directly decomposes the representational space. To our knowledge, this is the first application of SAEs to single-cell foundation model activations.

3 Methods

3.1 Data

We use human single-cell RNA-seq data from the Tabula Sapiens atlas [The Tabula Sapiens Consortium, 2022], a comprehensive multi-organ single-cell atlas. Our dataset comprises 1,000 cells sampled from the immune compartment, spanning diverse cell types including B cells ($n = 189$), $CD4^+$ T cells ($n = 180$), macrophages ($n = 127$), neutrophils ($n = 116$), $CD8^+$ T cells ($n = 112$), plasma cells ($n = 44$), monocytes ($n = 42$), natural killer cells ($n = 28$), and additional populations (Table 1). Cells originate from multiple tissues including lymph nodes, blood, spleen, bone marrow, lung, and thymus. Data were preprocessed using `scanpy` [Wolf et al., 2018]: library-size normalisation to 10,000 counts per cell, log transformation, and gene filtering to the scGPT vocabulary (38,607 genes mapped).

3.2 Foundation Model and Activation Extraction

We use the publicly available scGPT checkpoint fine-tuned on human brain tissue [Cui et al., 2024]. The model comprises 12 transformer layers with hidden dimension $d = 512$, 8 attention heads, and a vocabulary of 60,697 gene tokens. For each cell, gene expression values are tokenised as (gene ID, expression value) pairs, sorted by expression magnitude, and passed through the model.

We extract residual-stream activations from all 12 transformer layers. For each cell i at layer ℓ, we obtain per-token activations $\mathbf{h}_\ell^{(i,t)} \in \mathbb{R}^{512}$ for each gene token t. Cell-level representations are computed by mean-pooling over non-padding tokens:

$$\bar{\mathbf{h}}_\ell^{(i)} = \frac{1}{|\mathcal{T}_i|} \sum_{t \in \mathcal{T}_i} \mathbf{h}_\ell^{(i,t)} \tag{1}$$

where $\mathcal{T}_i$ is the set of valid (non-padding) token positions for cell i. This yields 1,000 cell-level activation vectors per layer, each in $\mathbb{R}^{512}$. Additionally, we retain all per-token activations (295,895 token vectors per layer) for SAE training, providing richer training signal.

3.3 Sparse Autoencoder Architecture and Training

For each layer, we train SAEs mapping activations to a higher-dimensional sparse representation:

$$\mathbf{z} = \mathrm{ReLU}(\mathbf{W}_{\mathrm{enc}}(\mathbf{h} - \mathbf{b}_{\mathrm{dec}}) + \mathbf{b}_{\mathrm{enc}}) \tag{2}$$

$$\hat{\mathbf{h}} = \mathbf{W}_{\mathrm{dec}}\mathbf{z} + \mathbf{b}_{\mathrm{dec}} \tag{3}$$

where $\mathbf{W}_{\mathrm{enc}} \in \mathbb{R}^{M \times d}$, $\mathbf{W}_{\mathrm{dec}} \in \mathbb{R}^{d \times M}$, and $M = 2{,}048$ ($4\times$ expansion factor). Following Bricken et al. [2023], decoder column norms are constrained to unity after each gradient step.

The training objective combines reconstruction fidelity with an L_1 sparsity penalty:

$$\mathcal{L} = \|\mathbf{h} - \hat{\mathbf{h}}\|_2^2 + \lambda\|\mathbf{z}\|_1 \tag{4}$$

We train SAEs at three sparsity levels: $\lambda \in \{1, 3, 10\}$ on layers 0 (early), 6 (middle), and 11 (final). Training uses Adam [Kingma and Ba, 2014] with learning rate 3×10^{-4}, batch size 512, cosine learning rate schedule, and 40 epochs on 50,000 subsampled token activations. Activations are z-score normalised per dimension before training.

The choice of λ values merits discussion. Initial experiments with $\lambda \in \{0.01, 0.03, 0.1\}$ (values typical for LLM SAEs) yielded $L_0 > 1{,}200$—meaning over 60% of dictionary elements were simultaneously active, providing no meaningful sparsity. We found that $\lambda \geq 1$ was necessary to achieve genuine sparse coding in this domain, likely reflecting differences in the activation statistics of single-cell models compared to language models.

3.4 Feature Analysis

Cell-type specificity. For each SAE feature j, we quantify alignment with cell-type labels using the area under the receiver operating characteristic curve (AUROC). For each annotated cell type c, we compute AUROC for the binary classification task of predicting membership in c from the feature activation $z_j^{(i)}$. A feature is cell-type-specific if AUROC > 0.8 for at least one cell type.

Gene attribution. For each feature j, we compute Pearson correlations between the feature activation z_j and each gene's expression level across cells. The top positively and negatively correlated genes provide a gene-level interpretation of each feature.

Gene set enrichment. We test whether the top correlated genes for each feature are enriched in curated gene sets using Fisher's exact test with Benjamini-Hochberg FDR correction [Benjamini and Hochberg, 1995]:

- **Immune cell markers**: canonical markers for B cells (*CD79A*, *MS4A1*), T cells (*CD3D*, *CD3E*), macrophages (*CD68*, *CD14*), NK cells (*NKG7*, *GNLY*), and neutrophils (*S100A8*, *S100A9*).
- **Functional gene sets**: cytokine signalling, antigen presentation, T cell receptor signalling, B cell receptor signalling, innate immunity.

Comparison with PCA. We compare SAE features against principal components on cell-type classification (AUROC per cell type from individual features/components) and interpretability (fraction of features/components with significant gene set enrichment at FDR < 0.05).

Table 2: SAE training results across layers and sparsity levels ($M = 2{,}048$). L_0: average number of active features per input. R^2: variance explained. Dead: features activating on $< 1\%$ of inputs.

Layer	λ	L_0	R^2	Alive	Dead
0	1.0	581	0.969	2,048	0
0	3.0	251	0.917	2,048	0
0	10.0	54	0.761	2,029	19
6	1.0	592	0.970	2,048	0
6	3.0	239	0.923	2,048	0
6	10.0	51	0.809	1,206	842
11	1.0	553	0.972	2,048	0
11	3.0	219	0.934	2,029	19
11	10.0	48	0.834	699	1,349

Table 3: Comparison of low vs. high λ regimes. Low λ achieves near-perfect reconstruction but no meaningful sparsity.

Layer	λ	L_0	MSE	Alive %
0	0.01	1,813	0.0016	100%
0	0.1	1,258	0.0097	100%
6	0.01	1,943	0.0007	100%
6	0.1	1,388	0.0082	100%
11	0.01	1,940	0.0010	100%
11	0.1	1,412	0.0074	100%

4 Results

4.1 Sparsity Requires Strong Regularisation

A key finding is that single-cell foundation models require substantially stronger L_1 regularisation to achieve sparse dictionary coding compared to language models. With $\lambda = 0.01$–0.1 (values commonly used for LLM SAEs), we observed $L_0 > 1{,}200$ active features out of 2,048 dictionary elements (Table 3), meaning the representations were not meaningfully decomposed. Increasing λ to $\{1, 3, 10\}$ yielded the target sparsity range (Table 2).

This observation has methodological implications: naively transferring hyperparameters from LLM interpretability work to biological models produces SAEs that are autoencoders but not *sparse* autoencoders. The likely explanation is that scGPT activations have different distributional properties than LLM activations—specifically, the cell-level representations may be more uniformly distributed across dimensions, requiring stronger regularisation to induce feature selection.

4.2 Layer Progression of Feature Specificity

Across all sparsity levels, we observed a clear progression in feature specificity from early to late layers, mirroring findings in LLM SAEs [Bricken et al., 2023].

Layer 0 (input embeddings): Features captured broad expression patterns. The mean best AUROC across cell types was 0.689 ($\lambda = 1$), with only erythrocytes and myeloid cells exceeding AUROC > 0.8. Gene set enrichments were limited.

Layer 6 (middle): Features began to resolve major cell lineages. At $\lambda = 1$, the mean best

Table 4: SAE features vs. PCA components for cell-type discrimination and interpretability (layer 11, $\lambda =$ [TBD: best]).

Metric	**SAE ($\lambda = 3$)**	**PCA (top 50)**
Mean best AUROC across types	0.772	0.819
Types with best AUROC > 0.8	5	—
% features with enrichment (FDR < 0.05)	64%	—
Mean L_0 per cell	219	50 (dense)
% alive features	99%	100%

AUROC was 0.794 with 8 cell types achieving individual feature AUROC > 0.8; 902 features showed significant gene set enrichment (FDR < 0.05).

Layer 11 (final): Features achieved the strongest cell-type specificity. At $\lambda = 1$, the mean best AUROC was 0.838, with 12 cell types distinguished at AUROC > 0.8; 1,284 features showed significant enrichment. This layer is the focus of subsequent analyses.

4.3 Cell-Type-Specific Features

Using the layer 11 SAE at $\lambda = 3$ (which achieves $L_0 = 219$ with $R^2 = 0.93$ and 99% alive features, representing a useful balance between sparsity and reconstruction), we identified features with strong cell-type specificity:

- **Macrophage features**: 72 features showed significant enrichment for macrophage markers (FDR < 0.05), with the best individual feature achieving strong discrimination of the 127 macrophages from other cells.
- **Erythrocyte features**: 87 features enriched for erythrocyte markers (*HBB*, *HBA1*), reflecting the distinctive transcriptomic profile of red blood cell precursors.
- **Plasma cell features**: 41 features enriched for plasma cell markers, capturing the immunoglobulin-secreting phenotype.
- **Monocyte features**: 49 features enriched for monocyte-associated genes (*CD14*, *S100A8/A9*).
- **Neutrophil features**: 39 features with enrichment for innate immunity gene sets.

At the most aggressive sparsity ($\lambda = 10$, $L_0 = 48$), 122 features retained significant enrichment despite only 160 alive features, indicating that the surviving features are highly cell-type-specific.

4.4 Gene Set Enrichment

We tested each feature's top correlated genes for enrichment in immune cell marker gene sets (Fisher's exact test, Benjamini-Hochberg correction). For the layer 11 SAE ($\lambda = 3$), 302 features showed significant enrichment (FDR < 0.05) for at least one immune gene set out of 472 alive features (64% annotatable). The most commonly enriched gene sets were erythrocyte markers (87 features), macrophage markers (72 features), monocyte markers (49 features), plasma cell markers (41 features), and neutrophil markers (39 features).

At the sparsest setting ($\lambda = 10$), 122 of 160 alive features (76%) received at least one annotation, suggesting that higher sparsity concentrates biological signal into fewer, more specific features.

4.5 Comparison with PCA

We compared SAE features against the top 50 principal components of the same activation space:

5 Discussion

5.1 SAEs as a Tool for Biological Model Interpretability

Our results demonstrate that sparse autoencoders, originally developed to interpret large language models, can be productively applied to single-cell foundation models. The key finding is that appropriately trained SAEs decompose the model's activation space into features that align with known biological programs—specifically, cell-type identities and immune gene modules.

This validates the core hypothesis that single-cell foundation models, like LLMs, represent biological information in superposition, and that SAEs can disentangle this superposition into interpretable components. The practical implication is that SAEs provide a principled, scalable method for asking "what has this model learned?" about any biological concept that manifests in the transcriptomic data.

5.2 Calibration of Sparsity Regularisation

A key methodological contribution is the finding that single-cell models require substantially higher L_1 penalties ($\lambda \geq 1$) than language models to achieve meaningful sparsity. This is not merely a technical detail: it reflects fundamental differences in activation geometry between language and biological foundation models. Language model activations, shaped by the heavy-tailed statistics of natural language, may naturally admit sparser decompositions than the more continuously distributed activations of transcriptomic models.

This finding cautions against naïve transfer of SAE hyperparameters across domains and suggests that the community developing biological foundation models should invest in domain-specific SAE calibration.

5.3 Limitations

Several important limitations constrain interpretation of our results.

Sample size. Our analysis uses 1,000 cells, which is small relative to the millions of cells used to pre-train scGPT. While sufficient to demonstrate the approach, scaling to larger datasets would improve statistical power and potentially reveal finer-grained features.

Model and tissue scope. We analyse a single model (scGPT) on a single dataset (Tabula Sapiens immune compartment). Extension to other models (Geneformer, scBERT) and tissue types (brain, tumour microenvironment) is needed to assess generality.

Causal validation. Our analysis is correlational: we show that SAE features *align with* known biology but do not establish that they *causally* contribute to model predictions. Feature ablation and steering experiments are needed.

Dictionary size. We use a fixed 4× expansion factor. Larger dictionaries (e.g., 16× or 64×) may reveal finer-grained features at the cost of increased dead features.

5.4 Connection to AI Safety

This work illustrates a productive bidirectional exchange between AI safety and biology. Methods developed to ensure transparency of frontier AI systems prove useful for understanding biological models; conversely, biological models provide a domain with objective ground truth (cell types, gene functions, pathways) that can be used to validate and improve interpretability methods. We argue that biological foundation models should be adopted as standard benchmarks for mechanistic interpretability research, complementing the language and vision model benchmarks currently in use.

5.5 Future Directions

Natural extensions include: (i) scaling to larger datasets and multiple tissues, including brain, to test whether SAE features capture tissue-specific programs; (ii) applying SAEs to Geneformer and other architectures for cross-model comparison; (iii) feature steering experiments to test causal roles; (iv) integration with GWAS gene sets to identify disease-associated features; and (v) temporal analysis using differentiation or disease-progression datasets.

6 Conclusion

We have demonstrated that sparse autoencoders can extract interpretable, biologically meaningful features from single-cell foundation model activations. By systematically exploring the sparsity–reconstruction trade-off, we identified a regime where SAE features align with known cell-type programs while providing genuine sparse coding. Our key methodological finding—that biological models require substantially stronger sparsity regularisation than language models—has practical implications for the growing community applying AI interpretability methods to biological systems.

This work establishes a bridge between mechanistic interpretability research and computational biology, providing both communities with new tools and benchmarks. Code and trained SAE models are available at `https://github.com/ikendiukhov/scfm-sae`.

Acknowledgments

We thank members of the University of Tübingen Computer Science Department for helpful discussions. Data were obtained from the Tabula Sapiens consortium.

References

The Tabula Sapiens Consortium. The Tabula Sapiens: A multiple-organ, single-cell transcriptomic atlas of humans. *Science*, 376(6594):eabl4896, 2022.

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*, 21(8):1470–1480, 2024.

Christina V Theodoris, Ling Xiao, Anant Chopra, Mark D Chaffin, Zeina R Al Sayed, Matthew C Hill, Helene Manber, Cliff Beaver, Steven Henber, and Patrick T Ellinor. Transfer learning enables predictions in network biology. *Nature*, 618(7965):616–624, 2023.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, et al. Toy models of superposition. *arXiv preprint arXiv:2209.10652*, 2022.

Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. Zoom in: An introduction to circuits. *Distill*, 5(3):e00024–001, 2020.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, et al. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. *arXiv preprint arXiv:2309.08600*, 2023.

Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, et al. Scaling monosemanticity: Extracting interpretable features from Claude 3 Sonnet. *Transformer Circuits Thread*, 2024.

Fan Yang, Wenchuan 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(10):852–866, 2022.

Shenghui Zhao et al. Large-scale cell representation learning via divide-and-conquer contrastive learning. *arXiv preprint arXiv:2306.04371*, 2024.

F Alexander Wolf, Philipp Angerer, and Fabian J Theis. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biology*, 19(1):1–5, 2018.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B*, 57 (1):289–300, 1995.