Transcoder-based Circuit Analysis for Interpretable Single-Cell Foundation Models

Sosuke Hosokawa[1], **Toshiharu Kawakami**[2], **Satoshi Kodera**[2], **Masamichi Ito**[2], **Norihiko Takeda**[2]

[1]The University of Tokyo, Tokyo, Japan
[2]The University of Tokyo Hospital, Tokyo, Japan
hoso-sosuke0421@g.ecc.u-tokyo.ac.jp, kawakami-toshinaru555@g.ecc.u-tokyo.ac.jp,
koderasatoshi@gmail.com, mitou.tky@gmail.com, ntakeda-tky@g.ecc.u-tokyo.ac.jp

Abstract

Single-cell foundation models (scFMs) have demonstrated state-of-the-art performance on various tasks, such as cell-type annotation and perturbation response prediction, by learning gene regulatory networks from large-scale transcriptome data. However, a significant challenge remains: the decision-making processes of these models are less interpretable compared to traditional methods like differential gene expression analysis. Recently, transcoders have emerged as a promising approach for extracting interpretable decision circuits from large language models (LLMs). In this work, we train transcoders on all 24 layers of the cell2sentence (C2S) model, a state-of-the-art scFM, and develop systematic pipelines for biological interpretation. Our analysis reveals that over 80% of transcoder features across most layers are biologically interpretable through Gene Set Enrichment Analysis (GSEA). Through a case study on endothelial cell classification, we demonstrate that extracted circuits correctly identify cell-type-specific genes and significantly enrich for relevant pathways (FDR = 0.0013), confirming that transcoders can identify internal features aligned with biological knowledge within complex single-cell models.

Introduction

In recent years, single-cell foundation models (scFMs) such as cell2sentence (C2S) (Levine et al. 2024) and Geneformer (Theodoris et al. 2023) have garnered significant attention in the field of computational biology. These models adapt techniques from large language models (LLMs) in natural language processing, combining pre-training on large-scale transcriptome data corpora to learn general gene-gene relationships with task-specific fine-tuning on smaller datasets (Theodoris et al. 2023; Cui et al. 2024). While these models have achieved state-of-the-art performance on various single-cell analysis tasks including cell-type annotation and cellular response prediction, their low interpretability, stemming from the inherent nature of neural network algorithms, remains a significant challenge. This is particularly crucial in single-cell analysis models, where biological interpretation of model predictions is essential, making improved interpretability in scFMs an urgent priority.

A major goal in efforts to elucidate the internal mechanisms of large-scale models like LLMs and scFMs includes identifying internal circuits: the combinations of features that determine model behavior (circuit tracing) (Dunefsky, Chlenski, and Nanda 2025; Elhage et al. 2021). In single-cell analysis models, discovered internal circuits could potentially lead to new discoveries when connected with biological insights. This pursuit of understanding model internal mechanisms falls under the research field of mechanistic interpretability, which has recently attracted considerable attention.

In the domain of natural language processing, mechanistic interpretability of LLMs has emerged as a major research topic, with various methods being proposed. Among these, sparse autoencoders (SAEs) (Huben et al. 2024) and their variant, transcoders (Dunefsky, Chlenski, and Nanda 2025), have gained attention as methods that can resolve the "polysemanticity" within LLMs and transform internal representations into interpretable features. Transcoders, in particular, have been shown to extract input-invariant and highly interpretable features by training neural networks with wide, sparsely activated intermediate layers that replace MLP layers, enabling the extraction of model internal circuits.

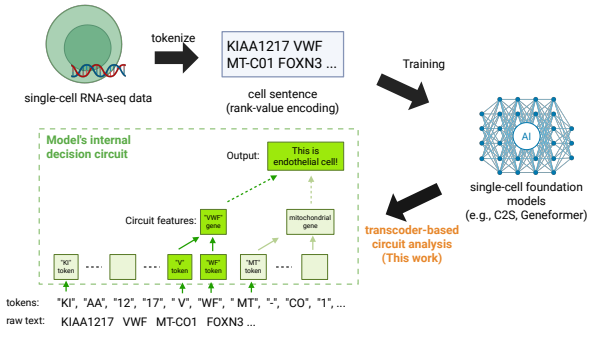

Figure 1: Pipeline for transcoder-based circuit tracing in scFMs. We train a transcoder on each MLP and attribute across features and attention to recover a sparse computational subgraph (circuit) that underlies cell-type predictions. The recovered features align with known endothelial biology (e.g., VWF, PTPRB, SPARCL1).

In this work, we apply transcoders to the C2S model, a state-of-the-art scFM, to extract its internal circuits and bi-

ologically interpret the circuit components (Figure 1). Our contributions are summarized as follows:

- We present the first application of transcoders to scFMs, successfully training transcoders on all 24 layers of the C2S model and demonstrating their effectiveness for mechanistic interpretability.
- We develop systematic biological interpretation pipelines: (1) an automated pipeline using GSEA to evaluate individual transcoder features, and (2) a pipeline to biologically interpret extracted circuits through gene identification and pathway analysis.
- We provide comprehensive empirical evidence showing that over 80% of transcoder features across most layers are biologically interpretable, with layer-wise analysis revealing interpretability patterns consistent with transformer architectural principles.
- We demonstrate practical circuit extraction through a case study on cell type classification, where the extracted circuit correctly identified endothelial-specific genes (e.g., VWF, PTPRB) and significantly enriched for "Endothelial cell: heart" pathways (FDR = 0.0013), demonstrating the biological relevance of our approach.

The remainder of this paper is organized as follows. We first explain the background methods of scFMs and transcoders along with their use for circuit analysis. We then discuss case studies of experiments applying transcoders to the C2S model. Subsequently, we summarize related work on interpretability in single-cell analysis models, and finally present conclusions and future perspectives.

Single-cell Foundation Models

Single-cell foundation models (scFMs) are transformer-based models pre-trained on large-scale transcriptome data. These models process input by ranking genes within each cell based on their expression levels and other factors, then arranging genes in rank order to form gene sequences. Prominent examples include Geneformer, scGPT, and cell2sentence (C2S).

ScFMs exhibit several architectural variations:

- **Architecture type**: Models can be either encoder-based or decoder-based.
- **Tokenization method**: Some models tokenize at the gene level, while others leverage natural language tokenizers to process gene sequences represented as natural language strings.
- **Gene ranking methods**: Different approaches exist for ranking genes (Theodoris et al. 2023; Levine et al. 2024), which represents a unique challenge specific to scFMs.

In this work, we focus on the C2S model, which employs a decoder-based architecture and utilizes natural language tokenization. C2S leverages the Pythia (Biderman et al. 2023) architecture and tokenizer, pre-trained on 57 million human and mouse cells from scRNA-seq data, along with biological literature abstracts (Levine et al. 2024). This approach enables the model to capture both gene expression patterns and broader biological knowledge from scientific texts.

Transcoders and Circuit Tracing

The Need for Transcoders: Resolving Polysemanticity

Understanding the internal representations of LLMs faces a fundamental challenge known as *polysemanticity*: the phenomenon where individual neurons or weights simultaneously encode multiple distinct concepts or functions (Elhage et al. 2022). For instance, a single neuron might strongly respond to both Japanese city names and gene names. This polysemanticity makes it difficult to disentangle and understand which parts of the weight matrices represent specific concepts through direct observation.

To resolve or mitigate polysemanticity and reconstruct the internals of large models into human-interpretable units, methods such as sparse autoencoders (SAEs) and their variant, transcoders, have proven effective (Huben et al. 2024; Dunefsky, Chlenski, and Nanda 2025).

Sparse Autoencoders and Transcoders

Sparse Autoencoders (SAE) SAEs consist of an encoder that transforms an input vector $\mathbf{x} \in \mathbb{R}^d$ into a higher-dimensional activation vector $\mathbf{z} \in \mathbb{R}^l_{>0}$ (where $l > d$), and a decoder that reconstructs the original dimension:

$$\mathbf{z} = \mathrm{ReLU}(\mathbf{W}_{\mathrm{enc}}\mathbf{x} + \mathbf{b}_{\mathrm{enc}}) \tag{1}$$

$$\hat{\mathbf{x}} = \mathbf{W}_{\mathrm{dec}}\mathbf{z} + \mathbf{b}_{\mathrm{dec}} \tag{2}$$

Typical SAEs use the same LLM hidden state for both encoder input and decoder output, and are trained to minimize the following loss function:

$$\mathcal{L} = ||\hat{\mathbf{x}} - \mathbf{x}||_2^2 + \lambda||\mathbf{z}||_1 \tag{3}$$

where the first term represents reconstruction error, the second term is a sparsity penalty that encourages sparse activations, and λ is a hyperparameter controlling the L1 weight.

Transcoders Transcoders are a variant of SAEs that learn on the input and output of each transformer layer's MLP rather than on the same hidden state:

$$\mathcal{L} = ||\hat{\mathbf{x}} - \mathrm{MLP}(\mathbf{x})||_2^2 + \lambda||\mathbf{z}||_1 \tag{4}$$

This formulation enables transcoders to approximate the transformer's MLP, decomposing MLP neurons into interpretable components.

Key Differences between SAEs and Transcoders While standard SAEs are trained to reproduce hidden states and extract input-dependent features, transcoders approximate specific modules within transformers (the MLPs) and thus extract input-invariant features. For explaining general model behavior, input-invariant features are preferable; therefore, transcoders are better suited for extracting circuits within transformers.

Circuit Tracing with Transcoders

Recent work has proposed methods for tracing circuits within LLMs using transcoders. We outline the key components below.

Attribution Between Transcoder Feature Pairs The contribution of transcoder feature i in layer l to feature j in layer l' is computed as:

$$z^{(l,i)}(x) \times \left(f_{\text{dec}}^{(l,i)} \cdot f_{\text{enc}}^{(l',j)}\right) \tag{5}$$

where $z^{(l,i)}(x)$ represents the input-dependent activation level, and the dot product $(f_{\text{dec}} \cdot f_{\text{enc}})$ is an input-independent fixed value. Here, $f_{\text{enc/dec}}^{(l,i)}$ denotes feature vector i of the transcoder encoder/decoder in layer l, corresponding to row vectors of $\mathbf{W}_{\text{enc}}$ and column vectors of $\mathbf{W}_{\text{dec}}$ respectively. This decomposition allows separate treatment of "input-independent general connections" and "input-specific importance."

Attribution Through Attention Heads Inter-feature relationships propagate not only through MLPs within the same token but also from different tokens via attention heads. Through the OV matrices of attention heads (Kamath et al. 2025), we can track which token's information contributes to specific features. Mathematically, by combining attention scores with OV matrices, we can compute how representations from source tokens contribute to downstream transcoder features.

Finding Computational Subgraphs By iteratively applying the attribution calculations above, we can identify the primary computational paths that activate specific features (Dunefsky, Chlenski, and Nanda 2025). The process involves:

1. Search for upstream features that strongly contribute to the target feature
2. Retain only top contributors and extend the paths
3. Iterate to obtain a set of important computational paths

Integrating these paths yields a sparse *computational subgraph* (circuit) that represents the model's internal decision-making process.

Extraction of Biologically Interpretable Features and Circuit Analysis

To evaluate whether the learned transcoders are useful for biological research, we constructed (1) an automated pipeline to assess the biological interpretability of individual transcoder features, and (2) a pipeline to perform biological interpretation of circuits extracted using transcoders.

Biological Interpretation Pipeline for Individual Transcoder Features

The interpretation pipeline for transcoder features consists of two stages: 1. Identification of tokens that specifically activate each feature and their corresponding genes. 2. Execution of Gene Set Enrichment Analysis (GSEA) (Subramanian et al. 2005) using the identified gene lists.

1. Identification of Tokens and Corresponding Genes that Specifically Activate Features

For each transcoder feature f, we calculate the frequency distribution of tokens that activate feature f. Specifically, using a corpus of gene sequences created from the Heart Cell Atlas v2 (Kanemaru et al. 2023), we compute the activation value $E(f,t)$ for each token t. In the cell2sentence model we employed, gene names can be split into multiple tokens (sub-words). In such cases, we treat tokens as corresponding to genes when they are contained within gene names in the corpus sentences. By listing genes in descending order of activation frequency for each transcoder feature, we can identify genes that specifically activate the feature.

2. Execution of Gene Set Enrichment Analysis (GSEA) Using Identified Gene Lists

We perform GSEA using the gene lists identified through the above procedure. GSEA is a method for evaluating whether a given gene list is associated with specific biological pathways or functions. Using the enrichment scores and False Discovery Rate (FDR) obtained from GSEA, we evaluate whether each transcoder feature is biologically interpretable.

Biological Interpretation Pipeline for Extracted Circuits

While transcoders enable extraction of internal circuits that reveal how scFMs determine their outputs, we also developed a pipeline to biologically interpret these circuits.

The extracted circuits are represented as directed graphs consisting of nodes (transcoder features in each scFM layer with their corresponding token positions) and edges (contributions between nodes). We identify gene names corresponding to tokens at each high-contribution node. The resulting gene lists are considered to play particularly important roles in how scFMs determine their outputs. Furthermore, by performing GSEA on these gene lists, we can evaluate whether they are associated with known biological pathways or functions.

Experiments

Training Transcoder on C2S

Experimental Setup We trained transcoders on each MLP layer of `vandijklab/C2S-Pythia-410m-cell-type-prediction` (van Dijk Lab 2025), a model from the C2S family from Hugging Face. For training data, we used the Heart Cell Atlas v2 (Kanemaru et al. 2023) dataset, splitting it into 90% for training and 10% for validation.

Training Hyperparameters The transcoder training was conducted with the following hyperparameters:

- Maximum learning rate: 1×10^{-4}
- Number of tokens per batch: 2048
- L1 coefficient: 1.4×10^{-4}
- Hidden dimension of transcoder: 8192 (expansion factor 8)
- Number of training tokens: 60 million

Model Validation To validate the trained transcoders, we compared three models: the original model, a model with all MLPs replaced by transcoders, and a model with all MLPs removed. Table 1 shows the validation losses for each model configuration.

Model Configuration	Original	Transcoder	No MLP
Validation Loss	2.48	4.63	12.67

Table 1: Validation loss comparison across different model configurations.

While the transcoder-replaced model shows some degradation compared to the original model, it achieves substantially lower loss than the model with MLPs removed, confirming that the transcoders successfully capture significant MLP functionality.

Additionally, we computed the KL divergence between the logits of the modified models and the original model:

	Mean
KL(Original‖Transcoder)	2.406
KL(Original‖No MLP)	10.52

Table 2: KL divergence between original model and modified models.

The average number of activated transcoder features per token (L0 value) across layers is shown in Figure 2.

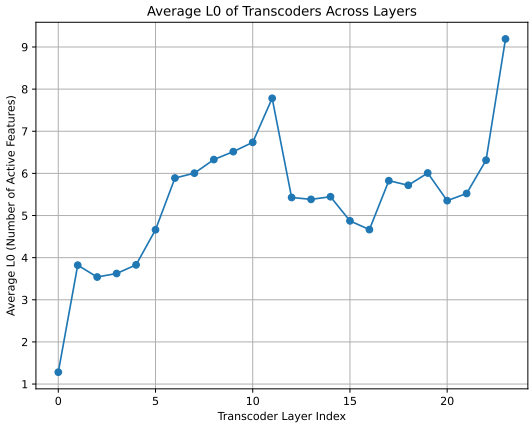

Figure 2: Average L0 values (number of active features per token) across transcoder layers.

Human Evaluation of Transcoder Features' Interpretability

To evaluate the interpretability of learned transcoder features, we conducted a human evaluation study. We focused on transcoder features from layer 12, defining "live features" as those with $\log_{10} E(f) \geq -4$, where $E(f)$ represents the probability that feature f activates per token. Figure 3 shows the distribution of live features.

From these live features, we randomly selected 20 features for detailed interpretation. Features were analyzed by investigating which tokens activate them. A feature was classified as "gene-level interpretable" if it consistently activated on tokens corresponding to specific genes or gene families.

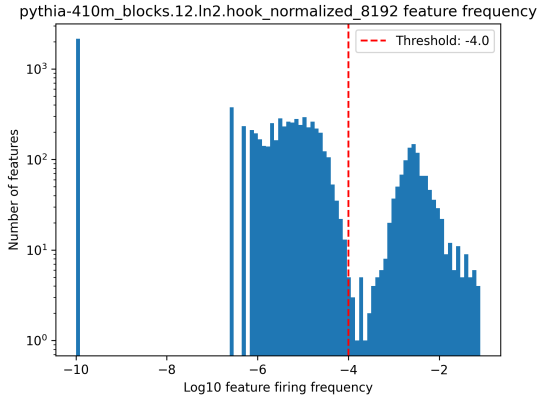

Figure 3: Distribution of live features in layer 12 transcoder with $\log_{10} E(f) \geq -4$.

Our evaluation revealed that 7 out of 20 features (35%) were gene-level interpretable. Table 3 presents the detailed analysis of each feature.

Biological Interpretation of Transcoder Features via GSEA

To interpret transcoder features from a biological perspective, we applied the previously described GSEA pipeline. In the pipeline, we used two databases for GSEA: "GO Biological Process 2025" (The Gene Ontology Consortium 2023) and "KEGG 2021 Human" (Kanehisa et al. 2021). We defined features as "biologically interpretable" if they identified at least one pathway satisfying FDR < 0.05 through this pipeline.

The C2S model we employed is a 24-layer transformer model. We applied the pipeline to "live features" with log feature activation frequency ($\log_{10} E(f)$) ≥ -3 from transcoders corresponding to each layer. The results are shown in Table 4.

As these results demonstrate, over 80% of live features in most layers were shown to be biologically interpretable.

Despite the high overall fraction of biologically interpretable features, we observed a relative dip in Layer 0. We posit two primary causes. (1) Early transformer layers preferentially encode form- and position-dependent regularities (e.g., delimiter tokens, local context, rank/positional cues), with more semantically aligned representations emerging in middle and upper layers; this layerwise progression is well documented in NLP models and plausibly carries over to scFMs operating on "cell sentences" (Tenney, Das, and Pavlick 2019; Clark et al. 2019). (2) Subword tokenization fragments gene symbols in a general-domain vocabulary, especially at low layers that behave more like form detectors; consequently, Layer-0 features often fire on token fragments (affixes, numerals) rather than whole genes, weakening the mapping from features to coherent gene sets and depressing GSEA yields. Prior evidence shows that domain-adapted vocabularies mitigate such mismatch, suggesting that biomedical tokenizers, or distillation into domain-tokenized models,

Feature ID	$\log_{10} E(f)$	Activating Tokens	Gene-Level Interpretable
6027	-2.66	_PSD token in PSD3 gene	Yes
2123	-2.60	OL token in GOLGA4, GOLGA8A, GOLPH3	No
4942	-1.25	Trailing 2 token in genes	No
2459	-3.01	_Y tokens (non-specific)	No
7892	-2.50	_ME token in MEIS2, MEF2A	No
7125	-2.53	PN token in PTPN family genes	Yes
3266	-2.90	SH token in TSHZ2, TSHZ3	No
1702	-2.43	AM token in LAM* genes (LAMC1, LAMTOR4)	No
3546	-2.34	X token in YBX1, YBX3	Yes
4319	-2.96	AA token in HSP90AA1 gene	Yes
2709	-2.57	BL token in ABLIM1, ABL1, ABLIM3	No
1283	-1.71	20 token in ZBTB20 gene	No
5085	-2.60	NA token in GNA* genes (GNAI1, GNAI2, GNA14)	No
2271	-2.67	NK token in CSNK family genes	Yes
1980	-2.91	_NA token in NAALADL2, NAIP	No
2808	-2.53	OCK token in ROCK1, ROCK2	No
4619	-2.90	NN token in TNNI3, TNNC1	No
5951	-2.41	O token in FOXO family genes	Yes
4819	-2.31	SB token in WSB family genes	Yes
5280	-1.98	3 token in RPS3, RPS3A	No

Table 3: Human interpretation of randomly selected transcoder features from layer 12. Note: In token representations, underscore (_) denotes a space character.

```
KIAA1217 VWF MT-CO1 FOXN3 MT-CO2 ENG MGLL
MT-ND4 MAGI1 MT-CO3 MT-CYB IQGAP1 SYNE1
CD36 RASAL2 SPARCL1 ST6GALNAC3 LINC00486
RAPGEF1 ID1 RBMS3 NFIB PTPRB LRMDA ARID2
MT-ATP6 SMAD2 ZBTB20 RGCC PLAA SLC48A1
TACC1 MECOM RB1 TSPAN14 FRMD4A AFDN ANO2
SHOC2 CDC42BPA RASGRF2 CCDC85A ESR2 SLC1A1
FRYL MALAT1 FAM241A DIAPH2 TSPAN15 LPAR6
HIF3A ITGA6 PARP14 NSD3 WNT2B FTX ART4
FBXW11 MTHFR AFF1 KHDRBS1 ZBTB46 ANKRD13C
RDX.

The corresponding cell type is:
```

Figure 4: Prompt for cell type classification consisting of 64 genes ordered by C2S gene rank encoding.

could improve low-layer interpretability (Beltagy, Lo, and Cohan 2019; Lee et al. 2020).

Case Study: Interpreting Cell Type Classification

To demonstrate the practical application of circuit tracing in scFMs, we analyzed how the C2S model performs cell type classification. We extracted a cell labeled as endothelial cell from Heart Cell Atlas v2 and prepared the prompt shown in Figure 4.

This prompt consists of the top 64 genes arranged according to C2S's gene rank encoding (a cell sentence), followed by a query for cell type prediction. The C2S model (`vandijklab/C2S-Pythia-410m-cell-type-prediction`) successfully predicted "endothelial cell of artery" as the cell type.

We extracted circuits for features that strongly activated on the final token (":") in the last layer. Figure 5 shows an example circuit extracted for feature ID 2692, one of the most strongly activated features.

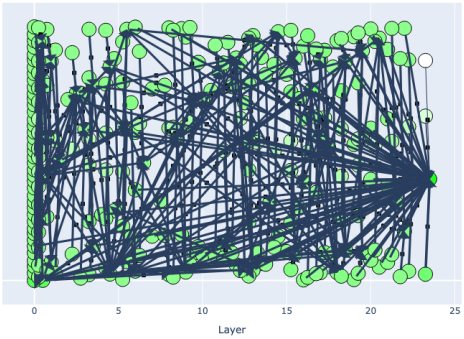

Figure 5: Extracted circuit for feature 2692 activated during endothelial cell classification. The circuit shows the computational graph tracing back from the final prediction token.

While the extracted circuit contains many activated features, many correspond to tokens in the text prompt (like "cell type:" part) rather than gene names.

When we extract only the nodes corresponding to gene name tokens, we can identify 9 gene-associated nodes: 'VWF', 'PTPRB', 'ANKRD13C', 'KHDRBS1', 'LPAR6', 'ST6GALNAC3', 'ART4', 'DIAPH2', and 'MT-ND4'. Using this gene list, we performed GSEA based on the "CellMarker Augmented 2021" (Hu et al. 2023) database, which significantly detected a pathway associated with "Endothelial cell: heart" cell type with FDR = 0.0013. This pathway achieved an enrichment score of 315.17, the highest score

Layer	Total Features	Biologically Interpretable	Fraction	Percentage (%)
0	116	78	78/116	67.24
1	179	159	159/179	88.83
2	606	529	529/606	87.29
3	665	552	552/665	83.01
4	729	614	614/729	84.22
5	919	780	780/919	84.87
6	967	854	854/967	88.31
7	921	805	805/921	87.40
8	641	563	563/641	87.83
9	542	460	460/542	84.87
10	636	540	540/636	84.91
11	625	533	533/625	85.28
12	915	808	808/915	88.31
13	865	762	762/865	88.09
14	880	802	802/880	91.14
15	868	776	776/868	89.40
16	778	682	682/778	87.66
17	766	679	679/766	88.64
18	746	664	664/746	89.01
19	791	710	710/791	89.76
20	932	827	827/932	88.73
21	938	835	835/938	89.02
22	922	804	804/922	87.20
23	269	220	220/269	81.78

Table 4: GSEA results for transcoder features across all layers, showing the number of biologically interpretable features (FDR < 0.05) among live features ($\log_{10} E(f) \geq -3$).

among all pathways satisfying FDR < 0.05. Since the input gene sequence was obtained from a cardiac endothelial cell from Heart Cell Atlas v2, this suggests that the extracted circuit is biologically plausible.

Specifically, literature review revealed that four of these genes are closely associated with cardiac endothelial cell biology:

- VWF (von Willebrand factor): a canonical endothelial marker localized to Weibel–Palade bodies (Valentijn et al. 2011).
- PTPRB (VE-PTP): an endothelial-enriched receptor-type tyrosine phosphatase that regulates junctional integrity and TIE2/EPHB4 signaling (Drexler and others 2019).
- KHDRBS1 (Sam68): an RNA-binding protein that modulates endothelial adhesion-site formation and migration (Rekad and et al. 2023).
- DIAPH2 (mDia2): a formin implicated in endothelial actin remodeling and phagocytosis-like uptake, linked to angiogenic behaviors (Rengarajan, Hayer, and Theriot 2016).

These genes are all closely associated with endothelial cell biology, suggesting that the extracted circuit captures biologically meaningful patterns. However, the circuit remains large and complex, with many features difficult to interpret biologically. This highlights the need for more refined circuit extraction methods and feature interpretation techniques in future work.

Related Work

While mechanistic interpretability of large-scale models including LLMs has attracted significant attention (Dunefsky, Chlenski, and Nanda 2025; Kamath et al. 2025; Huben et al. 2024), the field remains in its early exploratory stages. Particularly, applications of mechanistic interpretability techniques from natural language processing to bioinformatics models such as single-cell analysis models are still limited. Here we summarize the relationship between our work and these pioneering studies.

Schuster's scFeatureLens (Schuster 2025) and work by Claye et al. (Claye et al. 2025) have applied sparse autoencoders to scFMs such as Geneformer (Theodoris et al. 2023) and scGPT (Cui et al. 2024), providing frameworks to mechanistically interpret SAE features as biological concepts. Our research extends this line of work by utilizing transcoders, an advanced variant of SAEs, to extract internal decision circuits from scFMs and demonstrate their correspondence with biological concepts. The frameworks developed in these prior studies could potentially be applied to individual transcoder features, suggesting valuable directions for future research.

Additionally, mechanistic interpretability techniques have been applied to models that directly process genome sequences. For instance, Brixi et al. developed Evo 2 (Brixi et al. 2025), a genomic foundation model, and as part of their research incorporated SAE analysis to investigate the model's internal representations, revealing that it recognizes biologically important sequence patterns such as intron-exon boundaries. These studies collectively demonstrate the growing potential of mechanistic interpretability methods in understanding biological foundation models.

Conclusion and Future Work

In this work, we presented the first application of transcoders to single-cell foundation models, successfully training transcoders on all 24 layers of the cell2sentence model and developing systematic pipelines for biological interpretation. Our comprehensive analysis revealed that over 80% of transcoder features across most layers correspond to biologically interpretable concepts, as validated through Gene Set Enrichment Analysis. Notably, we observed a layer-wise interpretability pattern consistent with transformer architectural principles, where early layers focus on form and position encoding while middle and upper layers capture more semantically meaningful biological representations.

Our case study on endothelial cell classification demonstrated the practical utility of circuit extraction. The recovered circuit correctly identified endothelial-specific genes including VWF, PTPRB, KHDRBS1, and DIAPH2, and significantly enriched for "Endothelial cell: heart" pathways, highlighting the potential of extracting circuits that align with known biological pathways. This success suggests that transcoders can help analyze scFMs' internal computations

by decomposing them into interpretable biological components.

However, several challenges remain. First, our current pipeline simply extracts gene lists from circuits and applies enrichment analysis, without fully leveraging the richer information encoded in circuit structure, such as inter-node relationships, information flow patterns, and hierarchical feature dependencies. Future work should develop methods to interpret these structural properties and translate them into biological insights. Second, the impact of subword tokenization on feature interpretability, particularly evident in early layers, suggests that domain-specific tokenizers could improve mechanistic interpretability. Third, our analysis focused on cell type classification; extending to other tasks like perturbation response prediction could reveal task-specific interpretability patterns.

Looking forward, we envision several promising directions: (1) applying transcoders to other scFMs to assess generalizability, (2) developing methods to leverage interpretable features for model improvement and debugging, and (3) using mechanistic insights to guide the design of more interpretable architectures. As scFMs become central to single-cell analysis, ensuring their interpretability through techniques like transcoders will be crucial for both scientific discovery and building trust in AI-driven biological research.

References

Beltagy, I.; Lo, K.; and Cohan, A. 2019. SciBERT: A Pretrained Language Model for Scientific Text. In *Proceedings of EMNLP-IJCNLP 2019*, 3615–3620.

Biderman, S.; Schoelkopf, H.; Anthony, Q.; Bradley, H.; O'Brien, K.; Hallahan, E.; Khan, M. A.; Purohit, S.; Prashanth, U. S.; Raff, E.; Skowron, A.; Sutawika, L.; and van der Wal, O. 2023. Pythia: A Suite for Analyzing Large Language Models Across Training and Scaling. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*. PMLR.

Brixi, G.; Durrant, M. G.; Ku, J.; Poli, M.; Brockman, G.; Chang, D.; Gonzalez, G. A.; King, S. H.; Li, D. B.; Merchant, A. T.; Naghipourfar, M.; Nguyen, E.; Ricci-Tam, C.; Romero, D. W.; Sun, G.; Taghibakshi, A.; Vorontsov, A.; Yang, B.; Deng, M.; Gorton, L.; Nguyen, N.; Wang, N. K.; Adams, E.; Baccus, S. A.; Dillmann, S.; Ermon, S.; Guo, D.; Ilango, R.; Janik, K.; Lu, A. X.; Mehta, R.; Mofrad, M. R.; Ng, M. Y.; Pannu, J.; Ré, C.; Schmok, J. C.; John, J. S.; Sullivan, J.; Zhu, K.; Zynda, G.; Balsam, D.; Collison, P.; Costa, A. B.; Hernandez-Boussard, T.; Ho, E.; Liu, M.-Y.; McGrath, T.; Powell, K.; Burke, D. P.; Goodarzi, H.; Hsu, P. D.; and Hie, B. L. 2025. Genome modeling and design across all domains of life with Evo 2. *bioRxiv*.

Clark, K.; Khandelwal, U.; Levy, O.; and Manning, C. D. 2019. What Does BERT Look At? An Analysis of BERT's Attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP*, 276–286.

Claye, C.; Marschall, P.; Ouerdane, W.; Hudelot, C.; and Duquesne, J. 2025. A framework to extract and interpret biological concepts from scRNAseq generative foundation models. In *ICML 2025 Generative AI and Biology (GenBio) Workshop*.

Cui, H.; Wang, C.; Maan, H.; Pang, K.; Luo, F.; Duan, N.; and Wang, B. 2024. scGPT: toward building a foundation model for single-cell multi-omics using generative AI. *Nature Methods*, 21(8): 1470–1480.

Drexler, H. C. A.; and others. 2019. Vascular Endothelial Receptor Tyrosine Phosphatase: Identification of Novel Substrates Related to Junctions and a Ternary Complex with EPHB4 and TIE2. *Molecular & Cellular Proteomics*, 18(10): 2058–2077.

Dunefsky, J.; Chlenski, P.; and Nanda, N. 2025. Transcoders find interpretable LLM feature circuits. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24. Red Hook, NY, USA: Curran Associates Inc. ISBN 9798331314385.

Elhage, N.; Hume, T.; Olsson, C.; Schiefer, N.; Henighan, T.; Kravec, S.; Hatfield-Dodds, Z.; Lasenby, R.; Drain, D.; Chen, C.; Grosse, R.; McCandlish, S.; Kaplan, J.; Amodei, D.; Wattenberg, M.; and Olah, C. 2022. Toy Models of Superposition. *arXiv preprint arXiv:2209.10652*.

Elhage, N.; Nanda, N.; Olsson, C.; Henighan, T.; Joseph, N.; Mann, B.; Askell, A.; Bai, Y.; Chen, A.; Conerly, T.; DasSarma, N.; Drain, D.; Ganguli, D.; Hatfield-Dodds, Z.; Hernandez, D.; Jones, A.; Kernion, J.; Lovitt, L.; Ndousse, K.; Amodei, D.; Brown, T.; Clark, J.; Kaplan, J.; McCandlish, S.; and Olah, C. 2021. A Mathematical Framework for Transformer Circuits. Transformer Circuits Thread. Accessed 2025-09-08.

Hu, C.; Li, T.; Xie, Y.; and et al. 2023. CellMarker 2.0: an updated database of manually curated cell markers in human/mouse and web tools based on scRNA-seq data. *Nucleic Acids Research*, 51(D1): D870–D876.

Huben, R.; Cunningham, H.; Smith, L. R.; Ewart, A.; and Sharkey, L. 2024. Sparse Autoencoders Find Highly Interpretable Features in Language Models. In *The Twelfth International Conference on Learning Representations*.

Kamath, H.; Ameisen, E.; Kauvar, I.; Luger, R.; Gurnee, W.; Pearce, A.; Zimmerman, S.; Batson, J.; Conerly, T.; Olah, C.; and Lindsey, J. 2025. Tracing Attention Computation: Attention Connects Features, and Features Direct Attention. *Transformer Circuits Thread*.

Kanehisa, M.; Sato, Y.; Kawashima, M.; Furumichi, M.; and Tanabe, M. 2021. KEGG: integrating viruses and cellular organisms. *Nucleic Acids Research*, 49(D1): D545–D551.

Kanemaru, K.; Cranley, J.; Muraro, D.; Miranda, A. M. A.; Ho, S. Y.; Wilbrey-Clark, A.; Patrick Pett, J.; Polanski, K.; Richardson, L.; Litvinukova, M.; Kumasaka, N.; Qin, Y.; Jablonska, Z.; Semprich, C. I.; Mach, L.; Dabrowska, M.; Richoz, N.; Bolt, L.; Mamanova, L.; Kapuge, R.; Barnett, S. N.; Perera, S.; Talavera-López, C.; Mulas, I.; Mahbubani, K. T.; Tuck, L.; Wang, L.; Huang, M. M.; Prete, M.; Pritchard, S.; Dark, J.; Saeb-Parsy, K.; Patel, M.; Clatworthy, M. R.; Hübner, N.; Chowdhury, R. A.; Noseda, M.; and Teichmann, S. A. 2023. Spatially resolved multiomics of human cardiac niches. *Nature*, 619(7971): 801–810.

Lee, J.; Yoon, W.; Kim, S.; Kim, D.; Kim, S.; So, C. H.; and Kang, J. 2020. BioBERT: a Pre-trained Biomedical Language Representation Model for Biomedical Text Mining. *Bioinformatics*, 36(4): 1234–1240.

Levine, D.; Rizvi, S. A.; Lévy, S.; Pallikkavaliyaveetil, N.; Zhang, D.; Chen, X.; Ghadermarzi, S.; Wu, R.; Zheng, Z.; Vrkic, I.; Zhong, A.; Raskin, D.; Han, I.; De Oliveira Fonseca, A. H.; Ortega Caro, J.; Karbasi, A.; Dhodapkar, R. M.; and Van Dijk, D. 2024. Cell2Sentence: Teaching Large Language Models the Language of Biology. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, 27299–27325. PMLR.

Rekad, Z.; and et al. 2023. Coalescent RNA-localizing and transcriptional activities of SAM68 modulate adhesion and subendothelial basement membrane assembly. *eLife*, 12: e85165.

Rengarajan, M.; Hayer, A.; and Theriot, J. A. 2016. Endothelial Cells Use a Formin-Dependent Phagocytosis-Like Process to Internalize the Bacterium *Listeria monocytogenes*. *PLOS Pathogens*, 12(5): e1005603.

Schuster, V. 2025. Can sparse autoencoders make sense of gene expression latent variable models? arXiv:2410.11468.

Subramanian, A.; Tamayo, P.; Mootha, V. K.; Mukherjee, S.; Ebert, B. L.; Gillette, M. A.; Paulovich, A.; Pomeroy, S. L.; Golub, T. R.; Lander, E. S.; and Mesirov, J. P. 2005. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43): 15545–15550.

Tenney, I.; Das, D.; and Pavlick, E. 2019. BERT Rediscovers the Classical NLP Pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 4593–4601.

The Gene Ontology Consortium. 2023. The Gene Ontology knowledgebase in 2023. *Genetics*, 224(1): iyad031.

Theodoris, C. V.; Xiao, L.; Chopra, A.; Chaffin, M. D.; Al Sayed, Z. R.; Hill, M. C.; Mantineo, H.; Brydon, E. M.; Zeng, Z.; Liu, X. S.; and Ellinor, P. T. 2023. Transfer learning enables predictions in network biology. *Nature*, 618(7965): 616–624.

Valentijn, K. M.; Sadler, J. E.; Valentijn, J. A.; Voorberg, J.; and Eikenboom, J. 2011. Functional architecture of Weibel-Palade bodies. *Blood*, 117(19): 5033–5043.

van Dijk Lab. 2025. vandijklab/C2S-Pythia-410m-cell-type-prediction. Hugging Face. Model card. Trained on $\sim$57M single cells from CellxGene and HCA. Accessed 2025-09-08.