Identifying Biological Priors and Structure in Single-Cell Foundation Models

Flavia Pedrocchi [1] **Stefan Stark** [1] **Gunnar Ratsch** [1] **Amir Joudaki** [1]

Abstract

Foundation models pre-trained on large-scale transcriptomic data are gaining popularity for generating latent representations of cells or genes for downstream analysis. While these models suggest promising results towards a better understanding of cellular behavior, their complexity and black-box nature poses a challenge in their wider adoption in computational biology. Without a clear understanding of how these models process data and make predictions, it is difficult to discern their strengths and limitations and identify areas where they can be improved. In this study, we explore approaches for uncovering structural and biological connections within foundation models, using Geneformer and UCE as case studies. Our explored approaches are straightforward to implement, are adaptable across various transformer architectures, and suggest possible strategies to interpret and optimize existing models and architectures. Our primary findings utilize attention rollout for biological interpretation of attention maps, linear probes to uncover where learned biological concepts appear as well as a comparison of hidden states to show learning progression and the emergence of token patterns.

1. Introduction

Foundation models have revolutionized natural language processing, with models like BERT (Devlin et al., 2018) and GPT-2 (Radford et al., 2019) setting new benchmarks by significantly improving performance across a variety of tasks. Central to this advancement is the attention mechanism of transformers, which allows models to focus on different parts of the input data dynamically (Vaswani et al., 2023). These models excel because they require minimal labelled data for fine-tuning, as they effectively capture patterns in raw data, such as grammar and context, enabling them to generalize well across different tasks. However, they require large amounts of data to learn these general representations.

The success of foundational models in language modelling has inspired similar advancements in biology. For instance, emerging DNA and protein language models, such as ESM (Lin et al., 2023) and DNABERT (Ji et al., 2021), leverage the principles of foundation models in complex biological sequences. Meanwhile in the field of single-cell biology, large single-cell atlases are being collected (Regev et al., 2017), in addition to large databases such as CELLXGENE (CZI Single-Cell Biology Program et al., 2023). In order to digest such large data sources, single-cell foundational models (Yang et al., 2022; Cui et al., 2024; Chen et al., 2023) have been proposed as alternatives to traditional autoencoder based models (Lopez et al., 2018).

Despite the promises of these single-cell foundational models, several significant challenges remain. Firstly, these models are notoriously difficult to interpret, and mapping the complex computations performed internally to meaningful biological concepts is an ongoing challenge. Secondly, the computational costs associated with these models are extremely high, both in terms of pre-training and inference. Additionally, the lack of standardized benchmarks makes it challenging to assess the true value and efficacy of these models. Without clear benchmarks, it's not yet clear whether the substantial computational investments translate into significant improvements in solving downstream biological tasks.

To close this gap, we introduce different approaches to analyze foundation models and apply them to two popular foundation models for single cell data, Universal Cell Embeddings (UCE) (Rosen et al., 2023) and Geneformer (Theodoris et al., 2023), which allow us to analyze and interpret the structures these models uncover. These approaches include 1) attention rollout, which helps explain transformer self-attention through a biological lens, 2) linear probes, which offer insight into where learned concepts emerge and can highlight redundancies or relationships within a model, and 3) intra- and inter-layer token comparisons, which reveal the appearance of complex patterns introduced by specific embeddings or training procedures. Our hope is that these methods can help the community

[1]Department of Computer Science, ETH Zurich, Switzerland. Correspondence to: Flavia Pedrocchi <flaviap@ethz.ch>.

Proceedings of the ICML 2024 Workshop on Accessible and Efficient Foundation Models for Biological Discovery, Vienna, Austria.

better understand the biological insights captured by foundation models and assist practitioners in training and selecting model architectures.

2. Methods

We shortly review the previously published Geneformer and UCE models before describing our approaches used to analyze these models

2.1. Models

2.1.1. Geneformer

Geneformer (Theodoris et al., 2023) is a transformer-based model composed of six layers, each featuring four attention heads, an input size of 2048, and embedding dimension of 256. During pre-training, the model masks 15% of the genes within each transcriptome, and the corresponding tokens in the last hidden layer learn to predict the decode these genes using the context provided by the remaining unmasked genes. The transcriptome of each single cell is presented to the model as a rank value encoding where genes are ranked by their expression in that cell normalized by their expression across the entire Genecorpus-30M the model was trained on. The authors also report a larger, 12-layer version of the model on Hugging Face.

2.1.2. UCE

The Universal Cell Embeddings (UCE) model (Rosen et al., 2023) features 33 layers and processes inputs of 1024 genes. Each layer includes 20 attention heads, and the embeddings are represented in 1280 dimensions. During training, a random 20% of expressed genes are masked, and these are combined with a subset of non-expressed genes to form a set of query genes. The protein embedding tokens of these query genes are then paired with the Classify (CLS) token, a special token located at the start of the transformer sequence. This combination is inputted into a neural network to predict gene expression status. Thus, the CLS token is trained to represent the entire cell expression in a low-dimensional vector. The remaining tokens of the last layer's hidden units are not directly involved in the pre-training objective. This stands in contrast to Geneformer pretraig, where each token in the output layer is used to decode the masked genes. In UCE, the transcriptome is conveyed as an expression weighted sample of its corresponding genes. The authors also report a more lightweight 4-layer version of UCE with the source code.

2.2. Analysis techniques

2.2.1. Attention rollout

The self-attention mechanism generates millions of weights between hidden states, allowing the transformer to assign importance to tokens for generating context. While the self-attention mechanism between tokens in a single layer has a clear and simple interpretation, it is much harder to extend this to a multi-layer setting. In a multi-layer architecture, attention between tokens is mixed across different layers, making it harder to discern which tokens in the input are influencing an output token. To address this issue, we need to aggregate attentions across layers into a more interpretable form. One such method is known as attention rollout (Abnar & Zuidema, 2020). This technique involves tracing paths where information related to a specific input token is propagated up to the CLS token, then aggregating the attention scores along these paths. Ultimately, this process yields individual attention scores for each input token with respect to the CLS token. For Geneformer, no CLS token was available because the cell embedding is generated by averaging the embeddings of each gene detected in that cell. Therefore, we primarily tested averaging the attention scores across these gene embeddings, but we also examined the scores for individual gene embeddings.

We perform two analyses to examine the biological relevance of these CLS token attention scores. The input genes passed to the transformer are associated with clusters to aggregate them into categories and analyze transformer behavior. These clusters generated by the Human Protein Atlas (Uhlén et al., 2015) contain genes that have similar expression patterns, and each cluster has been manually annotated to describe common features in terms of function and specificity.
We also used the attention scores to generate subsets of strongly attended genes, which were compared to hallmark gene sets via over-representation analysis. These hallmark sets summarize and represent specific well-defined biological states or processes and were taken from the Human Molecular Signatures Database (Subramanian et al., 2005; Liberzon et al., 2015).

2.2.2. Linear Probe

We use linear classifiers called 'probes' (Alain & Bengio, 2018) trained only on a hidden unit of a given intermediate layer to predict certain input cell characteristics. These probes have been used to understand the roles of intermediate layers and we expect them to offer insights into where learned concepts emerge and so to highlight redundancies or relationships within a model.

2.2.3. Intra- and inter-layer similarity

For this analysis, we extract the hidden states of the models within and across layers and examine the cosine similarity between them. Cosine similarity is often used to evaluate the similarity between two vectors. It calculates the cosine of the angle between these vectors, focusing on their orientation rather than their magnitude. This is particularly useful when vectors represent embeddings, where the emphasis is on the direction or alignment of the data rather than the magnitude.

3. Results

3.1. Quantifying attended gene clusters

Using attention rollout and gene clustering as described before, we can inspect whether the transformer pays more attention to specific gene clusters than others for a given input and whether this may have a biological reason. Indeed, the 4-layer UCE model attends strongly to genes that are specific to the input cell type and weakly to non-specific genes as seen in Figure 1. The larger 33-layer model maintains some of these tendencies to a weaker degree (Extended Data Figure A3), possibly due to more complex interactions that are harder to categorize. Surprisingly, there was also a general increased interest in Langerhans and Enterocyte cell specific genes in all inputs for this model.

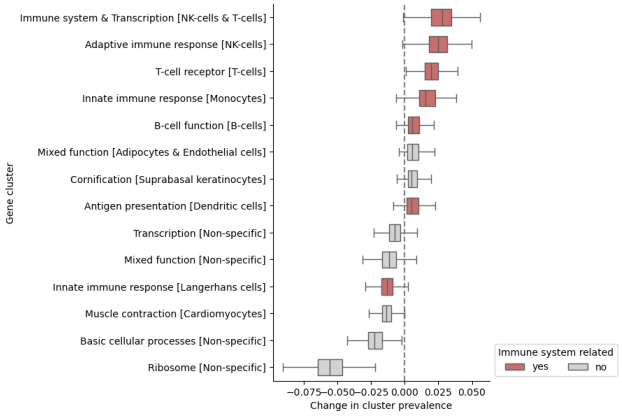

Figure 1. Difference in cluster prevalence between input sample of genes weighted by initial frequency and attention score of the 4-layer UCE model respectively. Here averaged over all CD4 T-cells. Clusters related to the immune system are shown in red

The attention pattern of the 6-layer Geneformer model is quite different to UCE and suggests a distinct approach to the task by the model. In Geneformer, there is a focus on non-specific clusters related to basic cellular processes and ribosome function, no matter the cell type of the input sample. The attention pattern of B-cells and Dendritic cells shows an increased importance of cell-specific genes, and, interestingly, Langerhans cell specific genes are also highly attended to as they were in the large UCE model.

Finally, we examine the results of the over-representation analysis of the most attended genes. In the 4-layer UCE model we found a general enrichment of the allograft rejection and inflammatory response hallmark gene sets. These sets are strongly related to the immune system, which is consistent with the nature of the PBMC dataset. Specifically T-cells show enrichment in the set comprising genes up-regulated by STAT5 in response to IL2 stimulation. Consistently, this gene set is strongly related to T-cell development.

3.2. Localizing learned concepts

We applied linear probes to the hidden states of the UCE model and explored cell type and COVID-19 presence as the classes for the linear classifier to differentiate between. The 33-layer UCE model learns these two concepts quickly, after the 10 first layers the probe on the CLS token has an F1-score of 0.9 and 0.8 for cell type and COVID-19 respectively (Figure 2). In this model, the concepts explored are learned by all hidden units simultaneously with little specialization. For example, it is possible to train a linear classifier for cell type prediction on a specific last-layer token and get good accuracy when testing that same classifier on other last-layer tokens. It may be that these units are computing nuances that are not picked up by the linear probe or that the model exhibits some redundancies that could be optimized. In general, information seems to always converge quite quickly to the CLS token, as it performs better that the average token when used as the input to the linear probe.

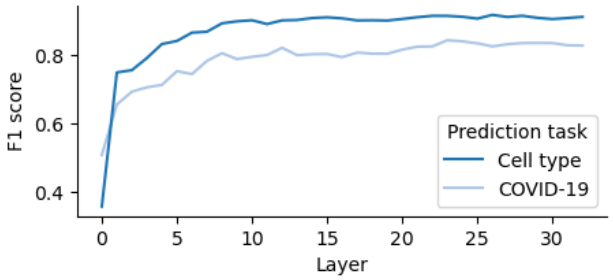

Figure 2. F1 score of the linear probe trained on the CLS token on different layers of the UCE model

3.3. Embedding structure and progression

Tracking the evolution and structure of different tokens across depth can reveal more insights about the inner workings of these models. Our results indicate pre-training procedure has a high impact how on how these token representations change through layers. While in Geneformer, each

individual token is pre-trained to decode a specific gene, in UCE the tokens don't need to adhere to a specific structure. Figure 3 shows how for UCE this produces a stark increase in cosine similarity between last layer tokens when compared to the initial embeddings. For Geneformer there is an increased spread but the mean in maintained. The group of samples with very low cosine similarity in the first UCE layer is caused by the chromosome identification tokens used, which are not generated by the protein language model.

Different gene tokenization and input procedures can introduce different latent structures. More shallow architectures maintain this structure in the last layer while larger models tend to diverge from it. We observe that the cosine similarity heatmaps and clustermaps of tokens are very different between UCE and Geneformer (Extended Data Figures A1 and A2).

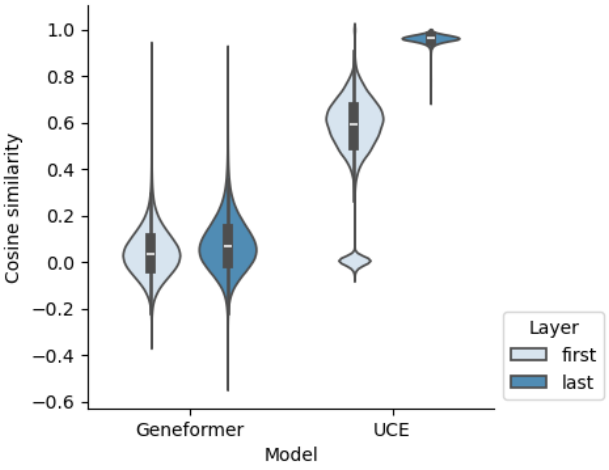

Figure 3. Comparison of token similarity within the first and last layers of Geneformer (left) and UCE (right)

We also investigate the progression of the CLS token across transformer layers and the cosine similarity between these units. In all models we can observe blocks of layers with a higher similarity that appear to represent learning phases of the model. For example, Figure 4 shows how the first 6 to 7 layers of the large UCE model form a unit and they coincide with the steepest ascent in accuracy when using linear probes. These progressions may also show when the CLS token improvement slows down and we start getting diminishing returns for transformer depth.

4. Discussion

In this work we proposed and investigated three methods to identify biological concepts and structure uncovered by

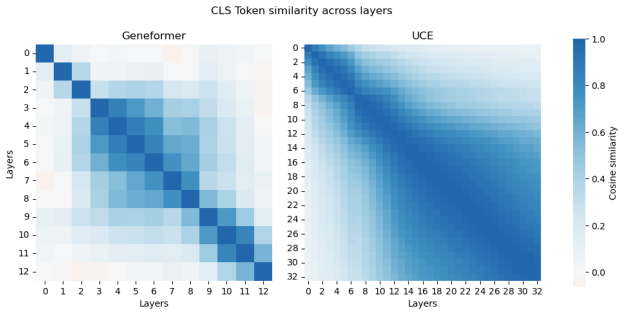

Figure 4. Cosine similarity between the CLS tokens across all layers of the corresponding models Geneformer (left) and UCE (right)

single-cell foundation models. Understanding how a model utilizes biological concepts to generate embeddings, as well as how efficiently this information is processed and how optimization procedures affect this processing, is essential to fulfilling the full potential of these emerging frameworks.

Our findings showed that the attention scores extracted via rollout were particularly biologically interpretable in the 4-layer UCE model for both gene clustering and overrepresentation analysis. In contrast, larger models may introduce too much complexity for a simple rollout towards the CLS token. The linear probes and CLS token progression indicated that at least simple concepts are learned quickly and that certain layer structures may display different learning phases of the model. Our results suggest that similarity-based approaches could identify when transformer depth becomes redundant, in which case model distillation could reduce the model complexity.

As next steps, it would be interesting to roll out the attention map only partway through the model to see whether specific parts of the model focus on different concepts. This may alleviate the higher complexity of larger models and allow for their interpretation. Additional clusters and gene sets for this analysis could also be explored. Additionally, we are interested in exploring different, more complex concepts for the linear probe as well as their exact relation to these potential learning phases.

In general, we advocate for more analysis of transformers to understand what and how they learn, particularly given the challenges in evaluating the quality of biological embeddings due to the lack of robust metrics. While some models have already been analyzed to understand their functioning, which provided useful insights, these analyses are often very architecture-specific. We believe more generalizable approaches can be a valuable addition.

References

Abnar, S. and Zuidema, W. H. Quantifying attention flow in transformers. *CoRR*, abs/2005.00928, 2020.

Alain, G. and Bengio, Y. Understanding intermediate layers using linear classifier probes, 2018.

Chen, J., Xu, H., Tao, W., Chen, Z., Zhao, Y., and Han, J.-D. J. Transformer for one stop interpretable cell type annotation. *Nature Communications*, 14(1):223, January 2023.

Cui, H., Wang, C., Maan, H., Pang, K., Luo, F., Duan, N., and Wang, B. scGPT: toward building a foundation model for single-cell multi-omics using generative AI. *Nature Methods*, February 2024.

CZI Single-Cell Biology Program, Abdulla, S., Aevermann, B., Assis, P., Badajoz, S., Bell, S. M., Bezzi, E., Cakir, B., Chaffer, J., Chambers, S., Michael Cherry, J., Chi, T., Chien, J., Dorman, L., Garcia-Nieto, P., Gloria, N., Hastie, M., Hegeman, D., Hilton, J., Huang, T., Infeld, A., Istrate, A.-M., Jelic, I., Katsuya, K., Kim, Y. J., Liang, K., Lin, M., Lombardo, M., Marshall, B., Martin, B., McDade, F., Megill, C., Patel, N., Predeus, A., Raymor, B., Robatmili, B., Rogers, D., Rutherford, E., Sadgat, D., Shin, A., Small, C., Smith, T., Sridharan, P., Tarashansky, A., Tavares, N., Thomas, H., Tolopko, A., Urisko, M., Yan, J., Yeretssian, G., Zamanian, J., Mani, A., Cool, J., and Carr, A. Cz cell×gene discover: A single-cell data platform for scalable exploration, analysis and modeling of aggregated data. *bioRxiv*, 2023. doi: 10.1101/2023.10.30.563174.

Devlin, J., Chang, M., Lee, K., and Toutanova, K. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018.

Ji, Y., Zhou, Z., Liu, H., and Davuluri, R. V. Dnabert: pre-trained bidirectional encoder representations from transformers model for dna-language in genome. *Bioinformatics*, 37(15):2112–2120, Aug 2021. ISSN 1367-4803. doi: 10.1093/bioinformatics/btab083.

Liberzon, A., Birger, C., Thorvaldsdóttir, H., Ghandi, M., Mesirov, J. P., and Tamayo, P. The molecular signatures database (MSigDB) hallmark gene set collection. *Cell Syst*, 1(6):417–425, December 2015.

Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., Shmueli, Y., dos Santos Costa, A., Fazel-Zarandi, M., Sercu, T., Candido, S., and Rives, A. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023. doi: 10.1126/science.ade2574.

Lopez, R., Regier, J., Cole, M. B., Jordan, M. I., and Yosef, N. Deep generative modeling for single-cell transcriptomics. *Nature Methods*, 15(12):1053–1058, December 2018.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. Language models are unsupervised multitask learners. 2019.

Regev, A., Teichmann, S. A., Lander, E. S., Amit, I., Benoist, C., Birney, E., Bodenmiller, B., Campbell, P., Carninci, P., Clatworthy, M., Clevers, H., Deplancke, B., Dunham, I., Eberwine, J., Eils, R., Enard, W., Farmer, A., Fugger, L., Göttgens, B., Hacohen, N., Haniffa, M., Hemberg, M., Kim, S., Klenerman, P., Kriegstein, A., Lein, E., Linnarsson, S., Lundberg, E., Lundeberg, J., Majumder, P., Marioni, J. C., Merad, M., Mhlanga, M., Nawijn, M., Netea, M., Nolan, G., Pe'er, D., Phillipakis, A., Ponting, C. P., Quake, S., Reik, W., Rozenblatt-Rosen, O., Sanes, J., Satija, R., Schumacher, T. N., Shalek, A., Shapiro, E., Sharma, P., Shin, J. W., Stegle, O., Stratton, M., Stubbington, M. J. T., Theis, F. J., Uhlen, M., van Oudenaarden, A., Wagner, A., Watt, F., Weissman, J., Wold, B., Xavier, R., Yosef, N., and Human Cell Atlas Meeting Participants. The human cell atlas. *Elife*, 6, December 2017.

Rosen, Y., Roohani, Y., Agarwal, A., Samotorčan, L., Tabula Sapiens Consortium, Quake, S. R., and Leskovec, J. Universal cell embeddings: A foundation model for cell biology. *bioRxiv*, 2023. doi: 10.1101/2023.11.28.568918.

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. 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, 2005. doi: 10.1073/pnas.0506580102.

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. Transfer learning enables predictions in network biology. *Nature*, 618(7965):616–624, Jun 2023. ISSN 1476-4687. doi: 10.1038/s41586-023-06139-9.

Uhlén, M., Fagerberg, L., Hallström, B. M., Lindskog, C., Oksvold, P., Mardinoglu, A., Åsa Sivertsson, Kampf, C., Sjöstedt, E., Asplund, A., Olsson, I., Edlund, K., Lundberg, E., Navani, S., Szigyarto, C. A.-K., Odeberg, J., Djureinovic, D., Takanen, J. O., Hober, S., Alm, T., Edqvist, P.-H., Berling, H., Tegel, H., Mulder, J., Rockberg, J., Nilsson, P., Schwenk, J. M., Hamsten, M., von Feilitzen, K., Forsberg, M., Persson, L., Johansson, F., Zwahlen, M., von Heijne, G., Nielsen, J., and Pontén, F. Tissue-based map of the human proteome. *Science*, 347