A framework to extract and interpret biological concepts from scRNAseq generative foundation models

Charlotte Claye [1] [2]   Pierre Marschall [1]   Wassila Ouerdane [2]   Céline Hudelot [2]   Julien Duquesne [1]

Abstract

Transcriptomic foundation models recently demonstrated strong performances on downstream tasks but remain poorly understood due to their high complexity. There is thus a growing need for post-hoc interpretability at the intersection of deep learning and biology. Sparse auto-encoders have recently been used to identify millions of meaningful concepts encoded in the latent space of large language models and were successfully applied to protein language models. A main challenge is the interpretation of these concepts, which should both reflect the internal mechanisms of the model and be comprehensible to domain experts. We introduce two novel approaches to interpret latent concepts from single-cell RNAseq models. First, we identify a set of genes that contribute to the concept activation, leveraging counterfactual perturbations of gene expressions. Second, we interpret the set of genes using textual gene descriptions from ontologies. We apply our interpretability framework to the cell embedding space of scGPT (Cui et al., 2024), focusing on immune cells. The methodology shows great promise in bridging the gap between deep learning experts and biology specialists.

1. Introduction

The development of high-throughput genomic technologies has significantly increased the availability of large-scale biological datasets (Barrett et al., 2005; Regev et al., 2017). This surge in data availability has enabled the application of recent advances in deep learning, notably the Transformer architecture (Vaswani et al., 2017) and unsupervised training strategies, to the field of biology (Consens et al., 2025; Wang et al., 2025). However, the inherent complexity of these large models poses significant challenges for interpretability, which limits their use in real-life and critical settings. Several solutions have been proposed to better understand deep learning models in general and in biology (Zhou et al., 2023; Conard et al., 2023; Treppner et al., 2022). In biology, some ante-hoc methods have been proposed where prior knowledge available in ontologies (Fabregat et al., 2016; Ashburner et al., 2000) is used at training time to obtain interpretable-by-design models (Bourgeais et al., 2022; Zarlenga et al., 2024). Alternatively, post-hoc explainability aims to explain a model once it is trained. For example, attribution methods identify the most influential genes for a prediction (Yap et al., 2021; Usman et al., 2025).

Recently, a large part of the community has focused on concept-based explainability with a human knowledge perspective (Poeta et al., 2023). Among post-hoc concept-based approaches, sparse dictionary learning demonstrated great potential for identifying concepts in the latent space of deep learning models (Sharkey et al., 2022; Huben et al., 2023; Fel et al., 2023). By decomposing internal embeddings into sparse representations, these approaches uncover concepts encoded by the model and interpretable by users. One such decomposition method, sparse auto-encoder (SAE), has been successfully applied to deep models in bioinformatics. Adams et al. (2025) demonstrated that the ESM-2 protein language model represents proteins using a combination of generic and family-specific concepts. Similarly, Schuster (2024) extracted and biologically interpreted hundreds of concepts from a generative transcriptomic model.

In this paper, we investigate the latent space of scGPT (Cui et al., 2024), a transformer-based generative model for single-cell RNAseq data commonly used in downstream tasks and performing well in cell-type classification. Single-cell RNAseq captures information about gene expression within individual cells, providing detailed insights into cell states and biological functions. A key question is what biological knowledge scGPT has encoded through its training. To address this, we train a sparse auto-encoder on the cell embeddings of 330k immune cells (Domínguez Conde et al., 2022).

A principal challenge lies in interpreting the concepts ex-

[*] Equal contribution [1] Scienta Lab, Paris, France [2] MICS laboratory, CentraleSupélec, Université Paris-Saclay, France. Correspondence to: Charlotte Claye <charlotte.claye@scientalab.com>.

Proceedings of the Workshop on Generative AI for Biology at the 42nd International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025. Copyright 2025 by the author(s).

tracted by SAEs. In particular, the interpretation should align with expert knowledge, but also reflect the mechanism of the model. We propose two novel approaches to address this. (1) For each concept, we identify the set of genes contributing to its activation using **attribution and counterfactual gene expression perturbations**. Compared to traditional differential gene expression (DEG) analysis, this method extracts genes that have an effect on the activation of the concept. (2) To further interpret these gene sets, we leverage **textual gene descriptions** from biological ontologies, extracting frequently occurring terms and visualizing them via word clouds. This method complements traditional gene set enrichment analysis (GSEA), which assumes that the gene set underlying a concept aligns with known biological pathways.

2. Methods

In this section, we introduce the methods used to extract and interpret concepts from cell embeddings, as illustrated in Figure 1.

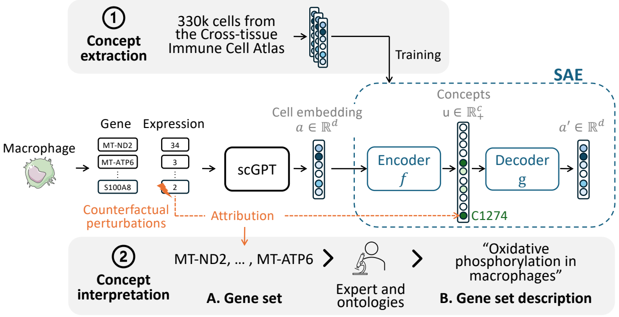

Figure 1. Methodology to extract and interpret biological concepts. (1) In a first step, a SAE is trained on 330k cell embeddings from the scGPT model to extract concepts. (2) In a second step, we interpret a concept by (A) leveraging attribution and counterfactual perturbations to identify a set of genes having an effect on the activation of the concept, and (B) including an expert in the loop and ontologies to biologically describe the set of genes. Macrophage icon provided by Servier Medical Art.

2.1. Model and dataset

We aim to analyze the cell embeddings produced by scGPT (Cui et al., 2024) using the Cross-tissue Immune Cell Atlas dataset (Domínguez Conde et al., 2022), which includes 330K immune cells from 16 tissues and 12 patients, with available cell type annotations. We follow the preprocessing described in Cui et al. (2024): we bin the gene expressions per sample and select 2000 highly variable genes. The model input only contains genes with non-zero expression. Given this preprocessed input cell, scGPT outputs a cell embedding $a \in \mathbb{R}^d$. Cell embeddings are shown in Appendix A.2.

2.2. Concept extraction

SAE architecture We rely on an SAE to learn a sparse and interpretable representation $u \in \mathbb{R}_+^c$ of the cell embedding $a \in \mathbb{R}^d$. As usual, the SAE is composed of an encoder f and a decoder g. The encoder f maps the cell embedding a to the concepts activation u with $u = f(a) = ReLU((a - b_d)W_e + b_e)$ (with $b_d \in \mathbb{R}^d, b_e \in \mathbb{R}^c, W_e \in \mathbb{R}^{d \times c}$). The decoder g then reconstructs the cell embedding via a linear combination of the concept vectors in $W_d \in \mathbb{R}^{c \times d}$ with $a' = g(u) = uW_d + b_d$. The decoder weights W_d are constrained to the unit norm. The training loss comprises a reconstruction loss l_r and a sparsity loss l_s with $l = l_r(a, a') + \lambda_s l_s(u) = MSE(a, a') + \lambda_s ||u||_1$ where MSE is the mean squared error. Hyperparameters and metrics are provided in the Appendix A.3. In particular, we use $c = 10000$ concepts.

Active concepts Due to the sparsity constraint, some concepts never activate. This means that their activation is equal to zero for all cells in the dataset. After training the SAE, 5559 out of the 10000 concepts were active for at least one cell. However, we observed that a large portion of these concepts were active for a very small number of cells and exhibited very low activation magnitude. Biologically, this signal is most likely noise. Hence, we decided to filter out concepts that activate for less than 0.01% of the cells of the dataset, which corresponds to approximately 33 cells. This post-processing has a very limited impact on the metrics. Further details are provided in Appendix A.3.

2.3. Concept interpretation

A main challenge with SAEs is to link the extracted concepts to their semantic meaning for human experts. In RNAseq data, biological signals are typically defined by sets of genes involved in the same biological pathways. A natural approach to interpret a concept is to first identify the set of genes related to it, and then link this gene set to a biological description.

Set of genes related to the concept Differential gene expression analysis (DGE) is a classic approach to identify genes that are differentially expressed between two conditions. Schuster (2024) leverage this method to identify genes that are differentially expressed between samples that highly activate a concept and those that weakly activate it. Although differentially expressed genes provide a good estimate of the signal encoded by a concept, the method does not ensure that these genes have an effect on concept activation.

We propose instead to use **attribution** between a gene expression and the activation of the concept. For each cell, an attribution score is computed for each gene, reflecting the importance of that gene's expression to the concept activation. We use Occlusion (Zeiler & Fergus, 2014), where the attribution score corresponds to the difference in concept activation before and after a gene expression perturbation. We further divide the score by the original concept activation to compare between cells. Equations are given in Appendix A.5.

To generate perturbations, we define two sets of cells. The **prototype** cells are the cells that highly activate the concept. The **counterfactual** cells are the closest cells to the prototypes, given the Euclidean distance in the cell embedding space, but having a concept activation equal to zero. Then, for each gene, its expression is replaced by its mean expression in counterfactual cells. This method avoids too big perturbations that could completely alter the cell and thus attribute high scores to genes that are not specific to the concept activation.

In order to identify a set of important genes for the concept activation, we compute attribution scores for 20 prototype cells. We then select the genes that have an attribution score higher than 0.05 (in absolute value) for at least 5 of the prototypes.

Biological interpretation of the set of genes The set of genes is most often not interpretable as is and should be further analyzed to describe the underlying biological processes. In the literature, this is usually done with gene set enrichment analysis (GSEA), as in Schuster (2024). By leveraging gene ontologies, in which gene sets have been grouped together by their involvement in the same biological pathway, one can automatically perform statistical tests to identify which biological processes the genes are involved in. However, this method relies on the assumption that the genes identified as important to the model align with biological processes already known and informed in the ontology.

In this work, we propose an alternative to enrichment analysis and interpret the genes by leveraging their **textual gene descriptions**. We collect textual descriptions of genes from the NCBI (Maglott et al., 2005) and process them to obtain a list of unique words for each gene. We remove the most common words that appear for at least 10% of the genes, after verifying that they do not convey important biological information. We then interpret a list of genes by identifying the most frequent words in their descriptions. We display them in a word cloud visualization.

3. Results

3.1. Interpretation of a megakaryocyte concept

To illustrate a concept interpretation, we analyzed the concept 676 (Fig. 2a-b), which is specifically activated in megakaryocytes (81% of the cells activating the concept are megakaryocytes, and 86% of megakaryocytes activate the concept).

We first performed a differential gene expression analysis (Appendix A.7) between the prototype and counterfactual cells, identifying numerous up- and down-regulated genes between the two populations (Fig. 2c). We used word clouds depicting frequent words in the gene descriptions of the up-regulated genes as a biological description of the concept (Fig. 2d). Because platelets are produced by megakaryocytes, the word "platelet" refers directly to this cell type. Additional words, including "alpha", "beta", "actin", "subunit", and "chain" likely refer to the cell's actin cytoskeleton. Prior research has shown that mutations in genes encoding myosin, a motor protein interacting the actin cytoskeleton, could affect platelet production (Johnson et al., 2007), suggesting a link between myosin and megakaryocytes.

Next, using the attribution method, we identified the genes that participate in the activation of concept 676 (Fig. 2e). We identified PPBP, which encodes a platelet-derived growth factor in accordance with megakaryocytes biological function, and MYL9 which codes for a light chain of myosin, a signal already identified in wordclouds from DGE.

We further analyzed the difference between attribution, which selected a few genes, and DEG, which selected hundreds of genes. We computed the variation of concept activation while progressively removing differentially expressed genes from the cell sequences. The genes are deleted from the most important to the least according to the attribution scores, and from the least to the most important. The resulting curves (Fig. 2f) show that a large number of DEGs do not have an effect on the activation of the concept.

3.2. Interpretation of macrophages and monocytes concepts

Macrophages are tissue-resident immune cells from the myeloid lineage, derived from monocytes, with improved phagocytic and antigen-presenting capabilities. We identified several concepts that are specific to monocytes, macrophages or shared by both. The visualizations of these concepts in the cell embeddings space are in Appendix A.6. Using the attribution method, we identified the genes affecting the activation of concepts. The selected genes for each concept, as well as their average attribution for 20 prototypes, are displayed in Figure 3. Despite some overlap, concepts are mostly linked to different sets of genes, suggesting that they may represent independent biological

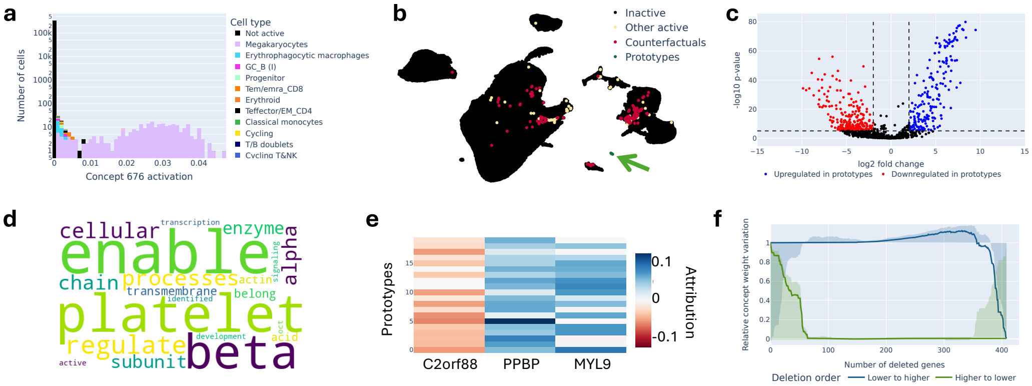

Figure 2. Illustration of a concept interpretation with concept 676 which is specifically activated in megakaryocytes. (a) Distribution of the concept 676 activations in the different cell types. (b) Prototypes and counterfactuals of concept 676 shown in the cell embedding space (UMAP). (c) Genes differentially expressed in prototypes of concept 676 compared to counterfactuals. The fold change compares the average expression of a gene in the two groups. The thresholds are adjusted p-value lower than 1e-5 and log2 fold change higher than 2 in absolute value. (d) Frequent words in the NCBI gene descriptions of up-regulated genes in prototypes. (e) Attribution scores of the genes having an effect on the activation of concept 676 for 30 prototypes. (f) Deletion curves showing the effect of removing differentially expressed genes on the activation of concept 676, sorted by average attribution score.

processes.

As an example, CTSD and CD163 genes, which are linked to phagocytosis, and HLA-DRA, HLA-DPA1, HLA-DPB1, which encode proteins crucially involved in antigen presentation on the class II major histocompatibility complex, are contained in concept 2942, which is thus related to the presentation of exogenous antigens by macrophages. The increased transcriptional activity of mitochondria-encoded genes, especially MT-ND2, MT-ND4, MT-ATP6, could underline an enhanced oxidative phosphorylation in macrophages in the concept 1274. In contrast, the concept 2337, which is rather activated in monocytes, is related to the alarmins S100A8 and S100A9 whose expression is down-regulated upon differentiation into macrophages. Finally, concept 2337 is activated by alarmins S100A8 and S100A9, but also by LYZ, which encodes an antimicrobial peptide. Although this concept could be related with immune cell activation by microbes, it is representative of a more vague concept activated in both monocytes and macrophages.

Altogether, the concepts observed between monocytes and macrophages appear relevant to well-known biological processes.

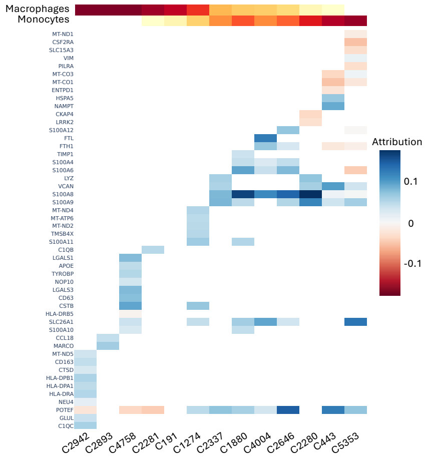

Figure 3. Genes identified by the attribution method for each monocyte/macrophage concept. The first two lines are the ratio of macrophages and monocytes among the samples that activate the concept, from 0 (white) to 1 (dark red). The attribution score is the average attribution among 20 prototypes of the concept.

4. Conclusion

We introduced a methodology to extract and interpret biological concepts encoded in the internal representations of transcriptomic foundation models. By applying this methodology to cell embeddings from scGPT we discovered several concepts interpretable in terms of established biological processes, demonstrating the potential of post-hoc concept-based explainability to explain deep learning models from a user perspective. Our analysis also showed that other concept interpretation methods, such as differential gene expression analysis, do not always align with the internal mechanisms of the model. While interpreting concepts is still challenging and time-consuming for experts, we believe that more methods, such as the ones we proposed, will unlock faster and more accurate interpretations. Such advances could enable the use of concepts in downstream tasks such as generation conditioned on concepts, analogous to steering approaches in large language models.

References

Adams, E., Bai, L., Lee, M., Yu, Y., and Alquraishi, M. From mechanistic interpretability to mechanistic biology: Training, evaluating, and interpreting sparse autoencoders on protein language models. *bioRxiv*, 2025. URL https://api.semanticscholar.org/CorpusID:276259557.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., et al. Gene ontology: tool for the unification of biology. *Nature genetics*, 25(1):25–29, 2000.

Barrett, T., Suzek, T. O., Troup, D. B., Wilhite, S. E., Ngau, W.-C., Ledoux, P., Rudnev, D., Lash, A. E., Fujibuchi, W., and Edgar, R. Ncbi geo: mining millions of expression profiles—database and tools. *Nucleic acids research*, 33 (suppl_1):D562–D566, 2005.

Bourgeais, V., Zehraoui, F., and Hanczar, B. Graphgonet: a self-explaining neural network encapsulating the gene ontology graph for phenotype prediction on gene expression. *Bioinformatics*, 38(9):2504–2511, 2022.

Conard, A. M., DenAdel, A., and Crawford, L. A spectrum of explainable and interpretable machine learning approaches for genomic studies. *Wiley Interdisciplinary Reviews: Computational Statistics*, 15(5):e1617, 2023.

Consens, M. E., Dufault, C., Wainberg, M., Forster, D., Karimzadeh, M., Goodarzi, H., Theis, F. J., Moses, A., and Wang, B. Transformers and genome language models. *Nature Machine Intelligence*, pp. 1–17, 2025.

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*, 21(8):1470–1480, 2024.

Domínguez Conde, C., Howlett, S., Suchanek, O., Polanski, K., King, H., et al. Cross-tissue immune cell analysis reveals tissue-specific features in humans. *Science*, 376 (6594):eabl5197, 2022.

Fabregat, A., Sidiropoulos, K., Garapati, P., Gillespie, M., Hausmann, K., Haw, R., Jassal, B., Jupe, S., Korninger, F., McKay, S., et al. The reactome pathway knowledgebase. *Nucleic acids research*, 44(D1):D481–D487, 2016.

Fel, T., Picard, A., Bethune, L., Boissin, T., Vigouroux, D., Colin, J., Cadène, R., and Serre, T. Craft: Concept recursive activation factorization for explainability. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2711–2721, 2023.

Huben, R., Cunningham, H., Smith, L. R., Ewart, A., and Sharkey, L. Sparse autoencoders find highly interpretable features in language models. In *The Twelfth International Conference on Learning Representations*, 2023.

Johnson, G. J., Leis, L., Krumwiede, M., and White, J. The critical role of myosin iia in platelet internal contraction. *Journal of Thrombosis and Haemostasis*, 5(7):1516–1529, 2007.

Maglott, D., Ostell, J., Pruitt, K. D., and Tatusova, T. Entrez gene: gene-centered information at ncbi. *Nucleic acids research*, 33(suppl_1):D54–D58, 2005.

Poeta, E., Ciravegna, G., Pastor, E., Cerquitelli, T., and Baralis, E. Concept-based explainable artificial intelligence: A survey. *CoRR*, 2023.

Regev, A., Teichmann, S. A., Lander, E. S., Amit, I., Benoist, C., Birney, E., Bodenmiller, B., Campbell, P., Carninci, P., Clatworthy, M., et al. The human cell atlas. *elife*, 6: e27041, 2017.

Schuster, V. Can sparse autoencoders make sense of latent representations? *arXiv preprint arXiv:2410.11468*, 2024.

Sharkey, L., Braun, D., and Millidge, B. Taking features out of superposition with sparse autoencoders. In *AI Alignment Forum*, volume 8, pp. 15–16, 2022.

Treppner, M., Binder, H., and Hess, M. Interpretable generative deep learning: an illustration with single cell gene expression data. *Human genetics*, 141(9):1481–1498, 2022.

Usman, M., Varea, O., Radeva, P., Canals, J., Abante, J., and Ortiz, D. Explainable ai model reveals disease-related mechanisms in single-cell rna-seq data. *arXiv preprint arXiv:2501.03923*, 2025.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Wang, Z., Wang, Z., Jiang, J., Chen, P., Shi, X., and Li, Y. Large language models in bioinformatics: A survey. *arXiv preprint arXiv:2503.04490*, 2025.

Wolf, F. A., Angerer, P., and Theis, F. J. Scanpy: large-scale single-cell gene expression data analysis. *Genome biology*, 19:1–5, 2018.

Yap, M., Johnston, R. L., Foley, H., MacDonald, S., Kondrashova, O., Tran, K. A., Nones, K., Koufariotis, L. T., Bean, C., Pearson, J. V., et al. Verifying explainability of a deep learning tissue classifier trained on rna-seq data. *Scientific reports*, 11(1):2641, 2021.

Zarlenga, M. E., Shams, Z., Nelson, M. E., Kim, B., and Jamnik, M. Tabcbm: Concept-based interpretable neural networks for tabular data. *Transactions on Machine Learning Research*, 2024.

Zeiler, M. D. and Fergus, R. Visualizing and understanding convolutional networks. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13*, pp. 818–833. Springer, 2014.

Zhou, Z., Hu, M., Salcedo, M., Gravel, N., Yeung, W., Venkat, A., Guo, D., Zhang, J., Kannan, N., and Li, S. Xai meets biology: A comprehensive review of explainable ai in bioinformatics applications. *arXiv preprint arXiv:2312.06082*, 2023.

A. Appendix

A.1. Cell types

Cell types used in this work are given in Table A.1.

Cell type (annotations)	Cell type (mid-level)	Cell type (high-level)	Count
Progenitor	Progenitor	Progenitor	1518
Cycling	Cycling	Cycling	1161
Cycling T&NK	Cycling	Cycling	2126
MNP/T doublets	Doublets	Doublets	2508
T/B doublets	Doublets	Doublets	1458
MNP/B doublets	Doublets	Doublets	744
NK_CD16+	ILC	ILC	20591
NK_CD56bright_CD16-	ILC	ILC	8902
ILC3	ILC	ILC	1312
MAIT	T_misc	T	4849
T_CD4/CD8	T_misc	T	5631
Tgd_CRTAM+	T_gamma_delta	T	4690
Trm_Tgd	T_gamma_delta	T	6887
Trm_Th1/Th17	T_CD4	T	16099
Tfh	T_CD4	T	15293
Teffector/EM_CD4	T_CD4	T	19869
Tnaive/CM_CD4	T_CD4	T	33865
Tnaive/CM_CD4_activated	T_CD4	T	3748
Tregs	T_CD4	T	12143
Trm_gut_CD8	T_CD8	T	25519
Tem/emra_CD8	T_CD8	T	14612
Tnaive/CM_CD8	T_CD8	T	7801
Trm/em_CD8	T_CD8	T	12674
Naive B cells	B_lymphocytes	B	13998
ABCs	B_lymphocytes	B	1209
GC_B (I)	B_lymphocytes	B	369
GC_B (II)	B_lymphocytes	B	203
Memory B cells	B_lymphocytes	B	28915
Pre-B	B_lymphocytes	B	75
Pro-B	B_lymphocytes	B	39
Plasma cells	Plasma	B	6270
Plasmablasts	Plasma	B	1710
migDC	Dendritic	Myeloblast	262
DC1	Dendritic	Myeloblast	356
DC2	Dendritic	Myeloblast	1147
pDC	Dendritic	Myeloblast	713
Intestinal macrophages	Macrophages	Myeloblast	599
Intermediate macrophages	Macrophages	Myeloblast	2236
Erythrophagocytic macrophages	Macrophages	Myeloblast	2103
Alveolar macrophages	Macrophages	Myeloblast	17238
Nonclassical monocytes	Monocytes	Myeloblast	2420
Classical monocytes	Monocytes	Myeloblast	21847
Erythroid	Erythrocytes	Erythrocytes	445
Megakaryocytes	Megakaryocytes	Megakaryocytes	317
Mast cells	Mast cells	Mast cells	3291

Table 1. Cell types and their count in the Cross-tissue Immune Cell Atlas (Domínguez Conde et al., 2022). As detailed in Appendix A.4, we group cell types into mid-level categories and high-level categories.

A.2. scGPT cell embeddings

Cell embeddings from scGPT are plotted in Figures 4 and 5.

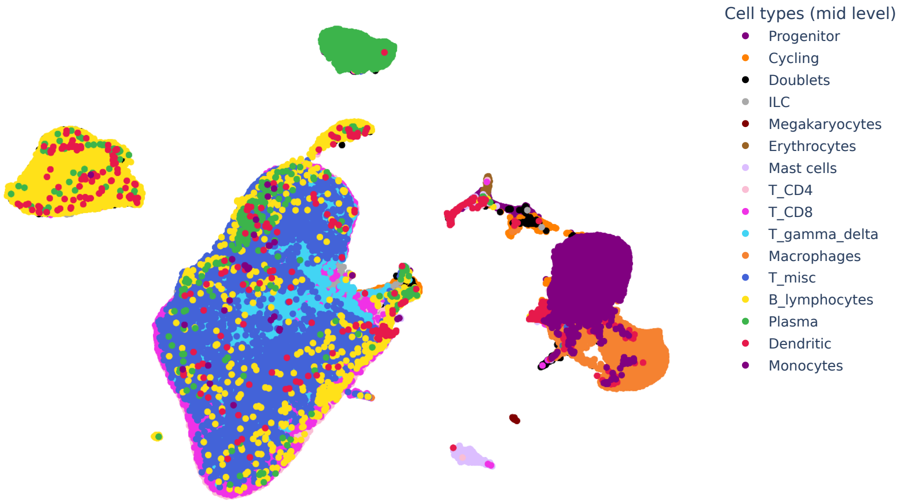

Figure 4. UMAP of scGPT cell embeddings, colored by cell types (mid-level aggregation)

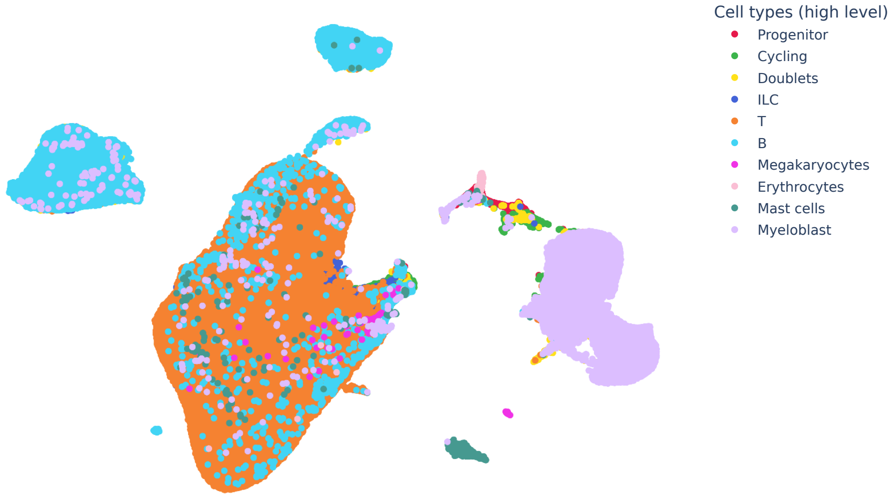

Figure 5. UMAP of scGPT cell embeddings, colored by cell types (high-level aggregation).

A.3. SAE

Hyperparameters We use $c = 10000$, $batch\_size = 1024$, $learning\_rate = 1e - 6$, $\lambda_s = 5e - 5$, $epochs = 1000$.

Evaluation Let $U \in \mathbb{R}_+^{n \times c}$ be the concept activations for n samples. We evaluate the sparse auto-encoder using classic metrics from the SAE literature. The explained variance evaluates the ability to recover the original model's activation a from the concepts u. The number of active concepts provides insight into the diversity of concepts we can expect. A concept i is active if $||U_i^T||_0 > 0$. The concept activation sparsity indicates how specific to a few samples the concept is. For a concept i, it is defined as $1 - \frac{1}{n}||U_i^T||_0$. Finally, the number of concepts per sample indicates how many concepts are needed to explain a sample and should be low to remain interpretable. For a sample j, it is defined as: $1 - \frac{1}{c}||U_j||_0$.

The metrics are provided in Table 2. Detailed metrics per sample are given in Figures 6 and 7.

Selection of concepts As detailed in Section 2.2, many concepts have a very low frequency. These concepts also exhibit very low activation magnitude, as presented in Figure 8.

Table 2. SAE metrics on the training and validation sets.

Metric	Training	Validation	Post-processed
Variance explained	0.86	0.86	0.87
Active concepts	5500	2361	304
Concepts per sample	35.7	35.7	35.5
Concept activation frequency	0.006	0.015	0.006

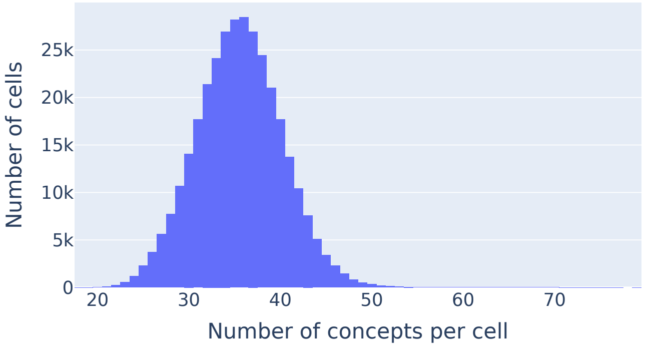

Figure 6. Distribution of the number of active concepts per cell.

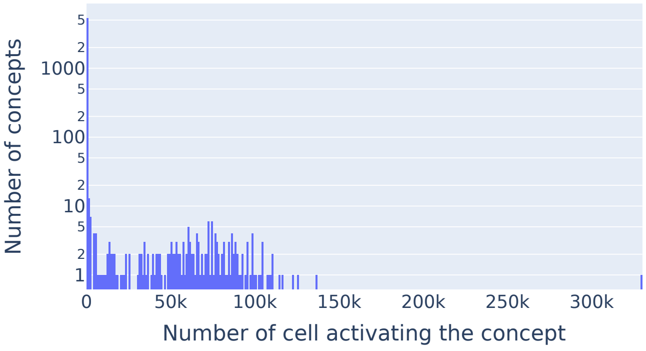

Figure 7. Distribution of the number of samples activating a concept.

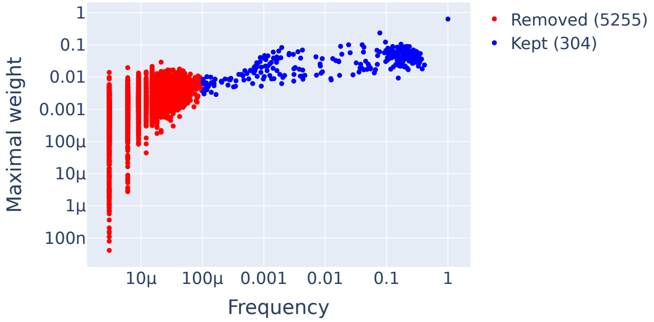

Figure 8. Frequency of activation of each concept and its maximal activation. Concepts with very low frequency ($frequency < 0.0001$) are removed from the analysis.

A.4. Concepts and cell types

To further understand how concepts related to cell types, we analyzed the ratio of cell types among samples activating a concept. For each concept, we computed the ratio of each cell type among the samples that activate the concept. We observed that some concepts are activated by several cell types that could be grouped into meaningful categories. To easily detect such patterns, we designed *mid-level* and *high-level* cell categories. For instance, the *Myelobloid* high-level category contains *Dendritic*, *Macrophages*, and *Monocytes* mid-level categories. Then the *Monocytes* mid-level category contains *Nonclassical monocytes* and *Classical monocytes* cell types. The cell types and categories are detailed in Appendix A.1. We refer to the different annotations as *low*, *mid*, and *high*.

For each concept, we determined the main cell type or category. We first looked for a specific (ratio> 0.8) cell type at the low level, then at the mid level, if not found, then at the high level. In Figure 9, we show the specificity and the sensitivity of the main cell type for each concept. Additional figures are provided in Appendix A.4. From this analysis, we identified three types of concepts. (1) **13 whole-cell-type concepts**, for which the samples that activate them are from the same cell type ($specificity > 80\%$), and samples from this cell type almost always activate them ($sensitivity > 80\%$). (2) **74 sub-cell-type concepts**, for which samples that activate it are from the same cell type ($specificity > 80\%$), but samples from this cell type do not always activate it ($sensitivity \leq 80\%$). (3) **217 cross-cell-type concepts** that are shared across several cell types.

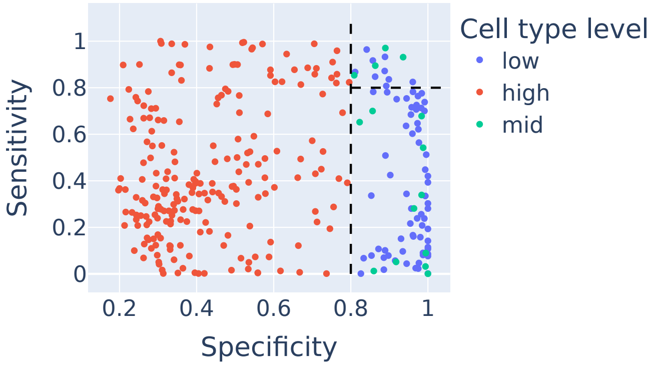

Figure 9. Specificity and sensitivity to the most common cell types among samples that activate a concept.

In Figure 10, we show the proportion of the main cell types among the cells activating the concept. Figures 11 and 12 show the ratio of each cell type among the samples activating a concept. Figure 13 displays cell type distributions for 4 concepts as well as their localization in the cell embedding space.

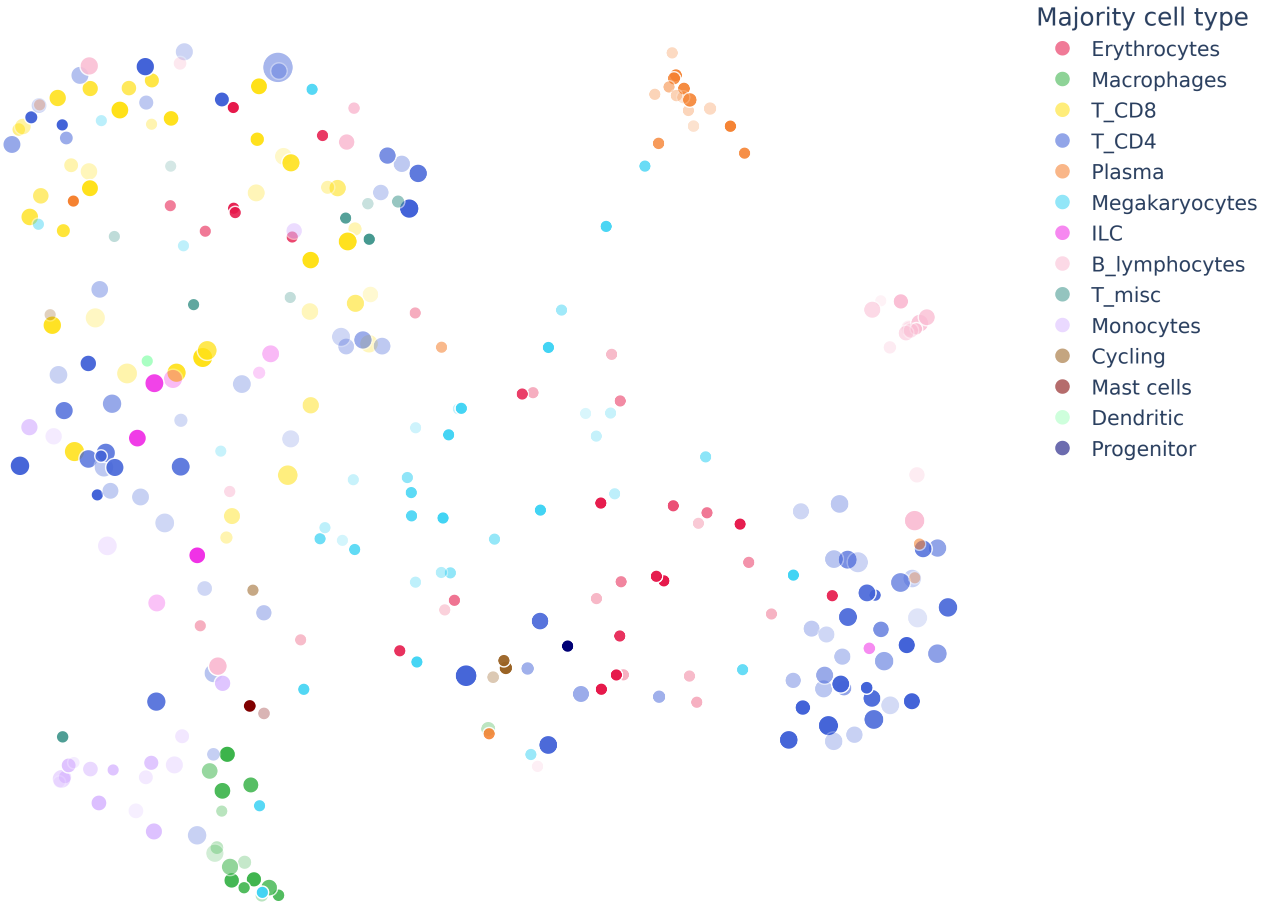

Figure 10. Each dot is a concept. The position is determined by the concept vectors using UMAP. The size corresponds to the activation frequency. The color corresponds to the main cell type among the cells activating the concept, and the transparency corresponds to the ratio of this cell type.

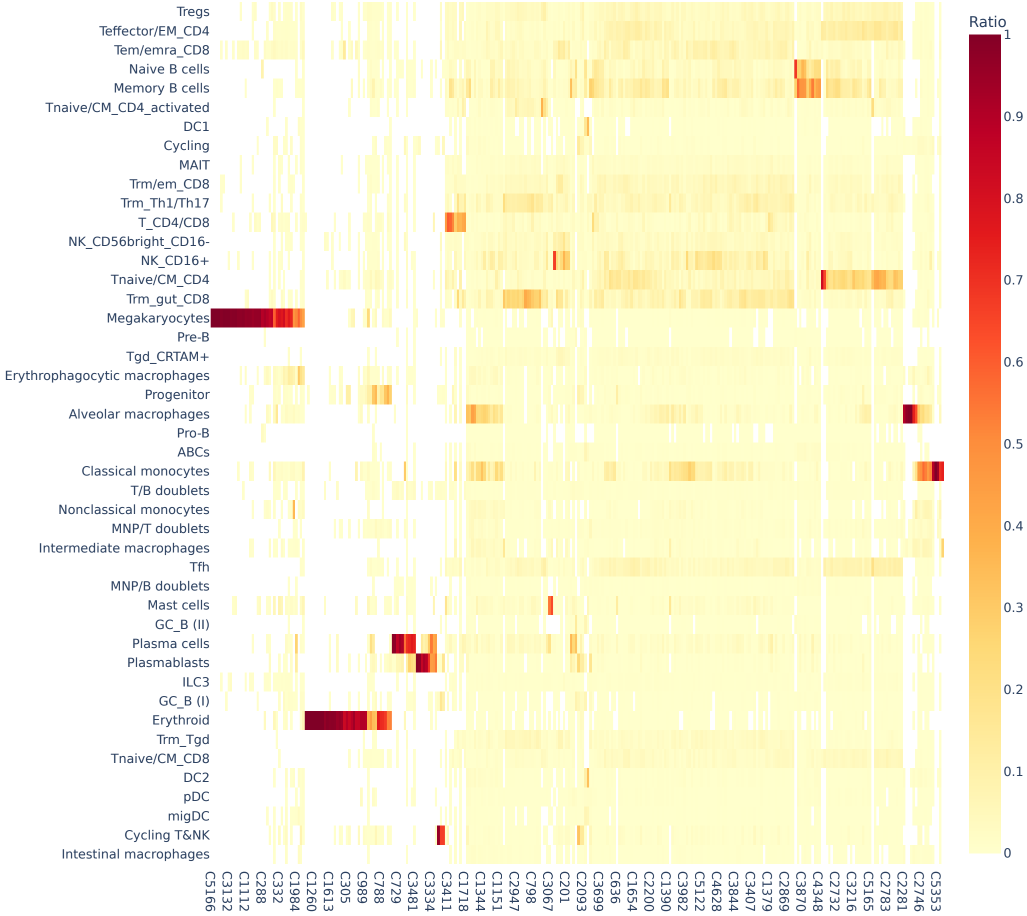

Figure 11. Ratio of a cell type (low-level) among the samples activating a concept.

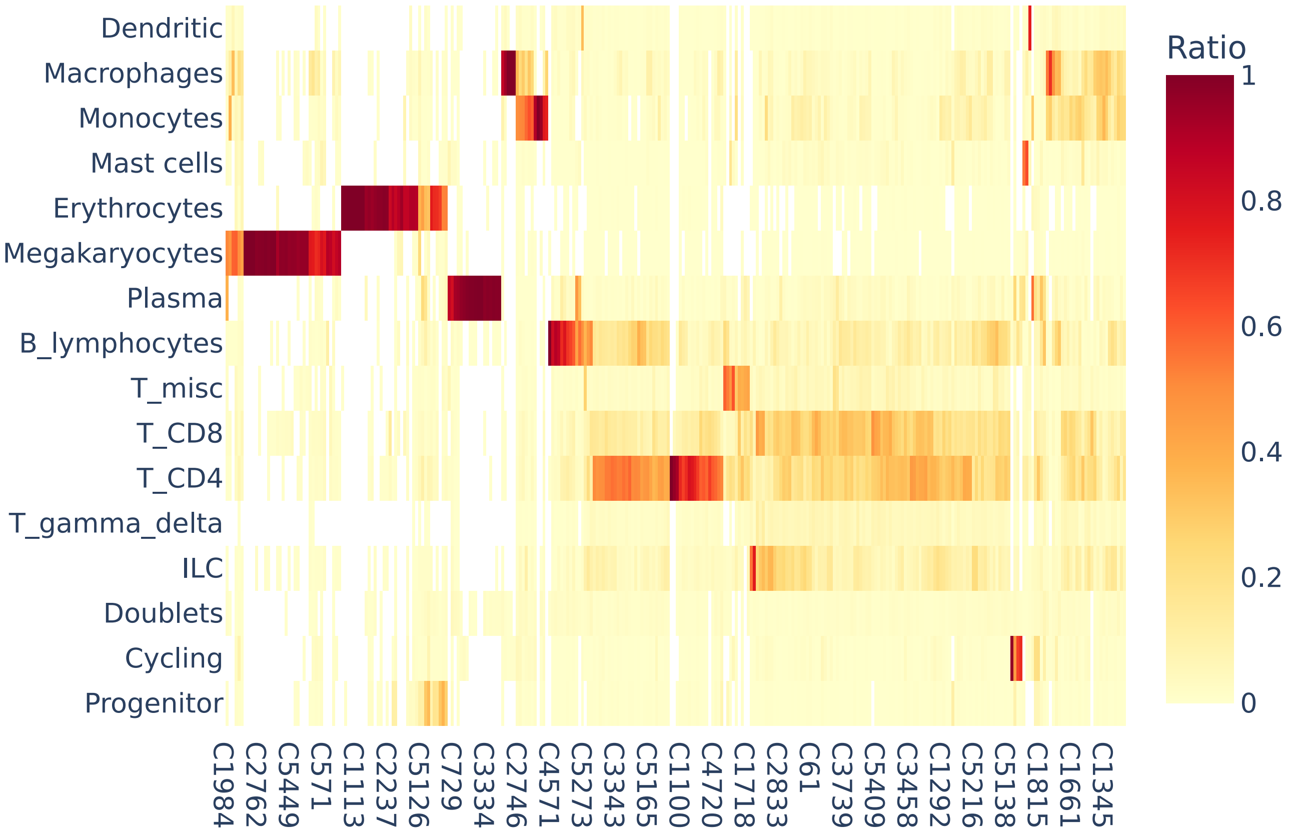

Figure 12. Ratio of a cell type (mid-level) among the samples activating a concept.

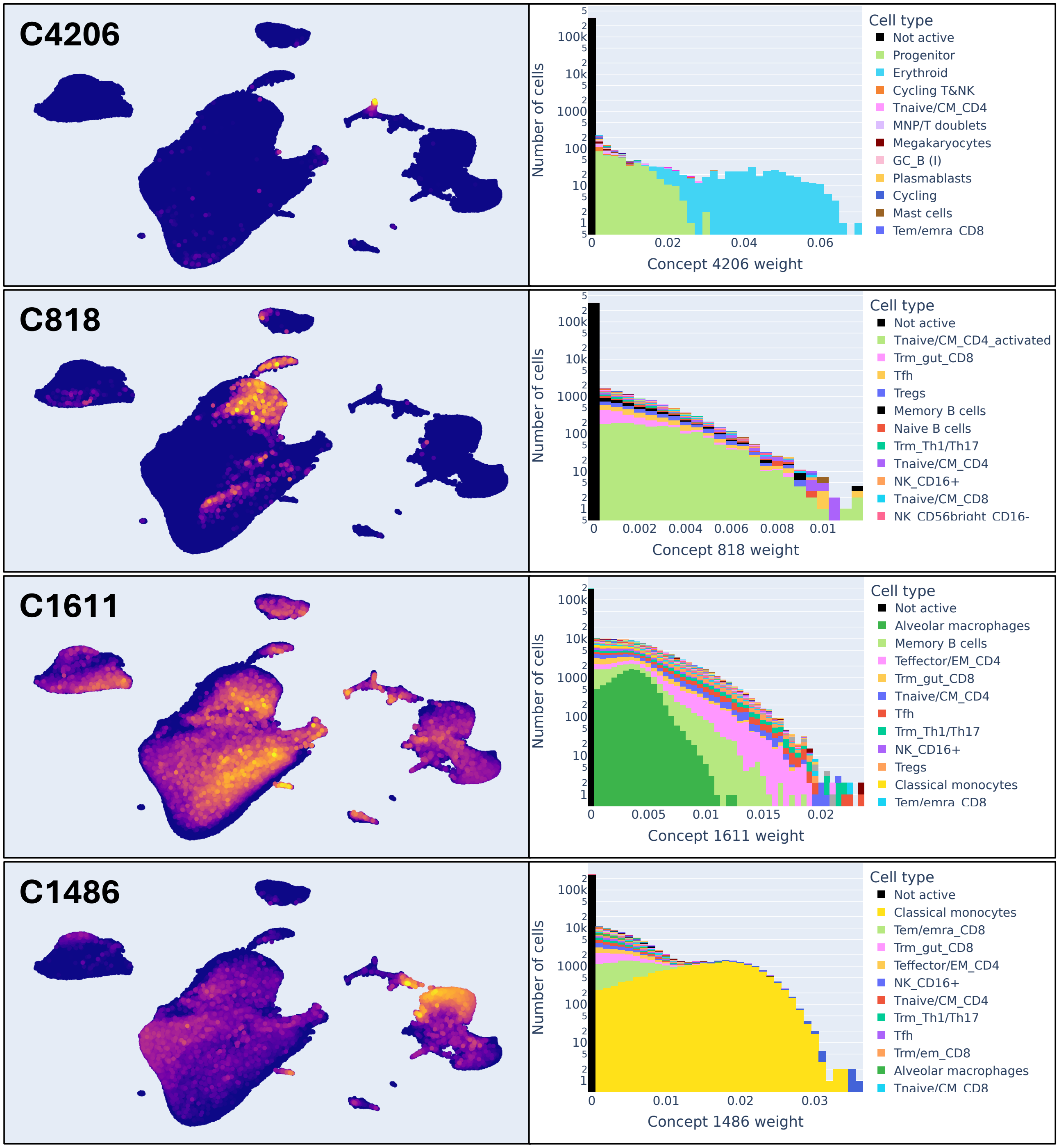

Figure 13. Examples of four concepts with different cell type distribution. (Left) localization in the cell embedding space (UMAPs) of the cells activating the concept, yellow color indicates high concept activation while dark blue indicates no concept activation. (Right) the corresponding cell type distribution

A.5. Occlusion

In this section, we define the attribution score of a gene i to a concept j.

Let ϕ be a single-cell RNAseq model which, given a sequence of N gene expressions $x = [x_1, x_2, ..., x_N] \in \mathbb{R}_+^N$, outputs a cell embedding $a \in \mathbb{R}^d$. Let f be the encoder of the trained SAE, which, given a cell embedding, outputs the concept activations. Let $x_i' \in \mathbb{R}_+$ be the average expression of gene i in counterfactual cells of concept j. The attribution score l_{ij} of gene i to concept j is given in Equation 1.

$$l_{ij} = \frac{f(\phi(x))_j - f(phi(\tilde{x}^i)_j}{f(\phi(x))_j} \tag{1}$$

$$\tilde{x}_k^i = \begin{cases} x_k & \text{if } k \neq i \\ x_i' & \text{if } k = i \end{cases}$$

A.6. Monocytes and macrophages concepts

Figure 14 shows the localization in the cell embedding space of the samples activating monocyte and macrophage concepts.

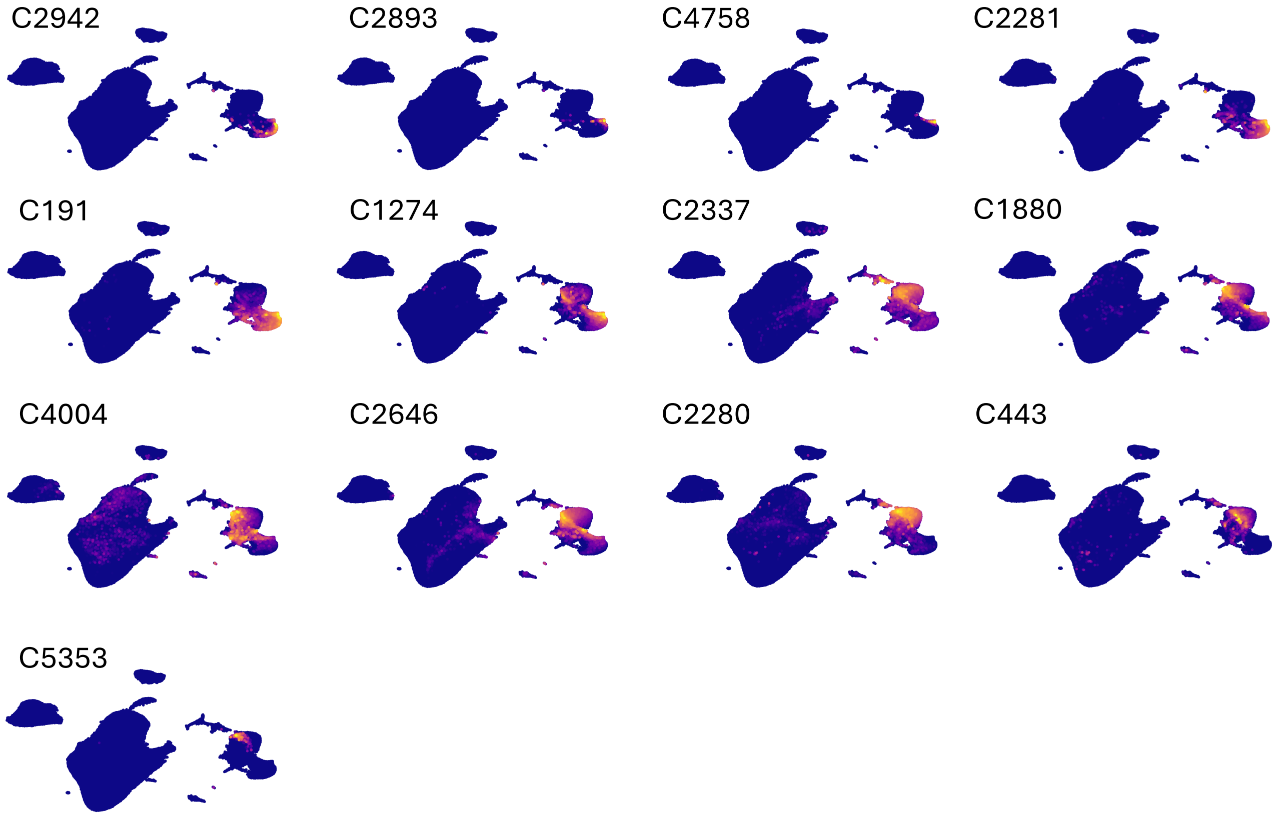

Figure 14. UMAP visualizations of monocyte and macrophage concepts in the cell embedding space. Yellow color indicates high concept activation while dark blue indicates no concept activation.

A.7. Differential gene expression analysis

The objective of differential expression analysis is to perform statistical analysis to discover changes in expression levels of genes between groups. In this work, we consider two groups: cells that activate a concept and cells that do not activate it. The fold change (Equation 2) compares the average expression of a gene in the first group with the average expression of this gene in the second group.

$$foldchange = \frac{\frac{1}{N_1}\sum_{i=1}^{N_1} e_i^1 + \epsilon}{\frac{1}{N_2}\sum_{i=1}^{N_2} e_i^2 + \epsilon} \tag{2}$$

We use the scanpy implementation (Wolf et al., 2018) with the Wilcoxon non-parametric statistical test and the Benjamini-Hochberg p-value correction method.

Due to the definition of fold change and the ϵ in the implementation, we observed very large log2 fold changes for the genes that are never expressed in one of the two groups. We removed them from the analysis after verifying that the mean expression in the other group is low.