STRUCTURED AND INTERPRETABLE PATIENT EMBEDDINGS FROM SINGLE-CELL FOUNDATION MODELS

Gonçalo Rei Pinto[1], **Su Han Cho**[1,3,4], **Selman Özleyen**[1], **Eva Fast**[6], **Soroor Hediyeh Zadeh**[1,2]*
Jie Quan[6]*& **Fabian Theis**[1,2,5]*

[1]Institute of Computational Biology, Computational Health Center, Helmholtz Munich, Neuherberg, Germany
[2]School of Life Sciences, Technical University of Munich, Munich, Germany
[3]Department Genes and Environment, Max Planck Institute of Psychiatry, Munich, Germany
[4]International Max Planck Research School for Translational Psychiatry (IMPRS-TP), Munich, Germany
[5]School of Computation, Information and Technology, Technical University of Munich, Munich, Germany
[6]Pfizer Worldwide Research and Development, Cambridge, MA, USA

{goncalo.pinto, soroor.zadeh, fabian.theis}@helmholtz-munich.de
jie.quan@pfizer.com

ABSTRACT

Recent advances have led to rapid proliferation of single-cell foundation models (scFM); however, methods for extracting biologically meaningful and interpretable knowledge from these large pre-trained models remain limited. We propose SCOPE (Structured Compositional Patient Embeddings) a model for learning interpretable patient representations from transcriptomic scFM models using a Concept Bottleneck Gaussian Mixture Variational Autoencoder (CB-GM-VAE). SCOPE models the distribution of cell types and a set of pre-defined concepts across patients from single-cell representations, resulting in patient representations that are both structured and interpretable. Using a single-cell RNAseq breast cancer atlas, we demonstrate that patient representations extracted from a continually pre-trained scFM by the CB-GM-VAE can outperform both the specialized patient representation learning baselines and simple pseudobulk approaches in various downstream prediction tasks. Moreover, the learned concept activities highlight biologically meaningful differences between primary and invasive tumors particularly involving $\mathrm{CD4}^+$ T cells, mast cells, and endothelial cells that are well supported by prior studies. Collectively, these findings demonstrate that SCOPE enables the extraction of human-interpretable, disease-relevant signatures from scFMs, bridging the gap between foundation models and mechanistic insight in translational genomics.

1 INTRODUCTION

Single-cell foundation models such as scGPT, TranscriptFormer, and Geneformer (Cui et al., 2024; Pearce et al., 2025; Theodoris et al., 2023) learn rich, low-dimensional cellular embeddings that capture complex gene-gene interactions by pre-training on vast, heterogeneous scRNA-seq data. However, their latent space is typically composed of entangled, non-linear representations that lack interpretability. In clinical settings, where decision-making requires human validation, robustness and reliability, lack of interpretability remains a significant barrier to the adoption of FMs for precision medicine.

With the growing availability of large-scale, high-dimensional data obtained from human subjects, patient representations are increasingly learned as compact latent embeddings derived from single-cell transcriptomics (Liu et al., 2024; Litinetskaya et al., 2024), genetics (Alsentzer et al., 2025), and Electronic Health Records (EHR) (Shmatko et al., 2025). These representations facilitate modeling patient trajectories and support the development of predictive models of human health, laying

[*] Co-corresponding authors

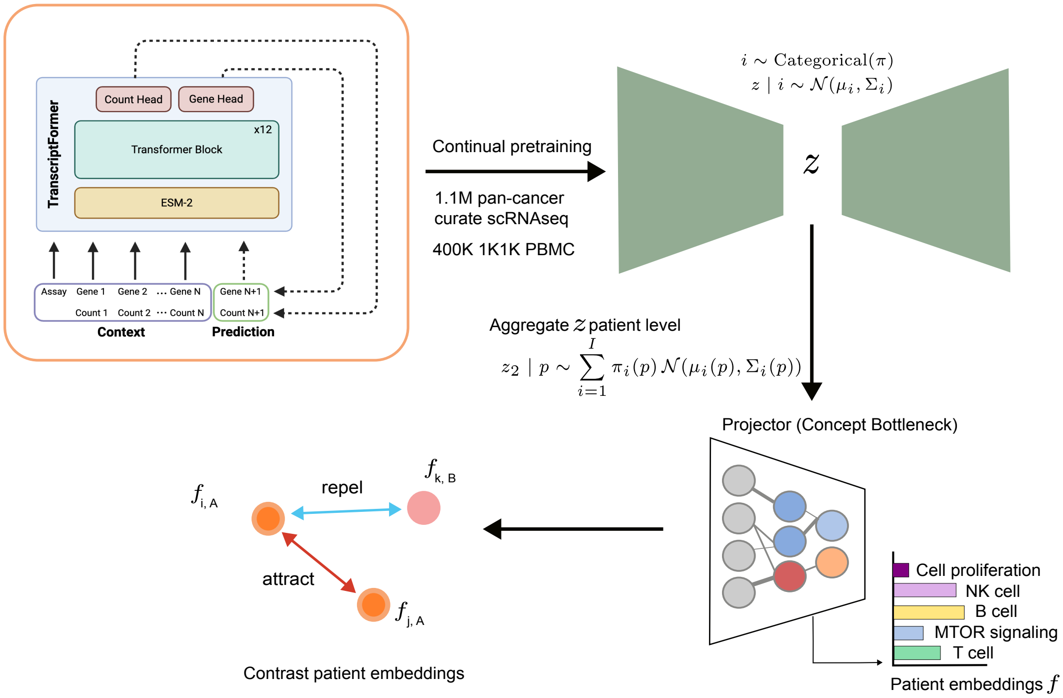

Figure 1: **Schematic representation of SCOPE**. A scFM, here the TranscriptFormer model, is continually pre-trained with a large-scale pan-cancer single-cell dataset. Single-cell representations are passed to a Gaussian Mixture VAE (GM-VAE) to model cell-type distributions across patients and aggregated at the patient level. These representations are then processed by a concept bottleneck projector to extract biological concepts that are relevant for a supervised learning task using a contrastive loss.

the foundation for patient-specific digital twins that can simulate disease dynamics and clinical outcomes.

Recent advances have seen the deployment of foundation models for learning patient representations (Pang et al., 2025; Jiang et al., 2025; Liu et al., 2024). Once learned, these representations are used for downstream tasks, such as disease classification, perturbation prediction (Pang et al., 2025), risk prediction (Shmatko et al., 2025), and therapeutic response prediction (Shen et al., 2025). These models generally leverage their internal cross-attention mechanisms to make representations interpretable. However, the interpretability enabled by these mechanisms are not comparable to human-interpretable biological concepts such as pathways.

The contributions of this work are summarized as follows:

- **CB-GM-VAE Architecture for interpretable patient representations**: We propose an architecture that extracts biologically meaningful, human-interpretable patient-level representations from single-cell foundation model representations.
- **Probing scFMs for interpretable Concepts:** We demonstrate that the internal representations of transcriptomic foundation models contain linear representations of genes that can be used to extract human-interpretable concepts such as gene modules and biological pathways that are relevant to a disease context, opening up new opportunities for application of foundation models in single cell and patient biology.

2 Methods

We propose **SCOPE** (Structured Compositional Patient embeddings) a model that learns interpretable patient representations from foundation models. Our approach seeks to bridge the single-cell-level and patient-level scales by learning embeddings where patient similarities are modeled based on similarity of cell type composition distribution, as well as transcriptional programs (i.e. gene programs).

Our methodology consists of three modules: A single-cell transcriptomic foundation model, a Gaussian Mixture Variational Autoencoder (GM-VAE) and a Concept Bottleneck projector optimized by a contrastive loss (Figure 1). We start by adapting the foundation model backbone to cancer, healthy and inflammatory specific datasets followed by the GM-VAE architecture. The GM-VAE is employed to capture the multinomial distribution of cell types across patients. Then, a projector module Shen et al. (2025) adds an interpretability layer and projects the representations generated by GM-VAE into an interpretable latent space. We leverage the pre-trained TranscriptFormer model, the TF-Sapiens, trained only on human data, and apply continual pre-training. The two other modules, that is the GM-VAE and Projector, are trained from scratch on the same training corpus. Our model attempts to disentangle the rich, complex representations from FMs into patient embeddings were each dimension is a biological pathway.

2.1 Continual Pre-training

To adapt the TranscriptFormer backbone for the disease-specific projector concepts, we employ a Continual Pre-training strategy. Unlike traditional fine-tuning, continual pre-training utilizes a balanced corpus that includes both novel disease-specific datasets and foundational references previously encountered by the model—such as the OneK1K cohort (Yazar et al., 2022) and the inflammatory precursors from the CELLxGENE Census (Chan Zuckerberg Initiative Single-Cell Biology System, 2023; Oliver et al., 2024) while adding newly seen, specific cancer datasets such as the ones seen in Tumor Immune Single-Cell Hub 2 (TISCH2) (Han et al., 2023). This approach (Li et al., 2026) allows to maintain previously learned universal gene relationships while adapting the model towards disease specific representations. The full dataset amounted to 1,495,677 cells. Due to compute resources we could only train the full model on one epoch.

2.2 Modeling cell type distribution across patients with GM-VAE

We first input raw single-cell RNA-seq counts into TranscriptFormer to generate cell embeddings. These embeddings are passed to a GM-VAE, which models the underlying cell type distributions by attempting to reconstruct the embeddings. We adopt a Gaussian Mixture prior in our VAE architecture, as it was previously shown to better capture the cell type diversity in patients (Joodaki et al., 2025; Wang et al., 2024a). Unlike unimodal VAEs that assume $p(\mathbf{z}) = \mathcal{N}(\mathbf{0}, \mathbf{I})$, the GM-VAE models the latent space as a mixture of K Gaussian components, representing distinct cell type distributions defined *a priori*. By utilizing a learnable categorical prior y and a Gumbel-Softmax (Jang et al., 2017; Maddison et al., 2017) relaxation, the model is able to approximate cell type-specific categorical distributions.

Subsequently, individual cell embeddings are aggregated into a patient-level representation using the cluster probabilities derived from the encoder. This patient vector is then transformed through an element-wise product with patient-specific gene embeddings to generate a Patient × Gene matrix. Finally, this matrix is mapped via the biological projector into a latent space of human-interpretable dimensions, where each value denotes the activation level of a specific tumor pathway.

2.3 Generation of patient embedding from GM-VAE representations

Disease subtypes are defined at the patient level, yet single-cell foundation models operate on individual cells. Bridging this resolution gap requires an aggregation scheme that is fully differentiable such that gradients from the patient-level contrastive objective (Eq. 9) can propagate back to the GM-VAE and TranscriptFormer while capturing the heterogeneous cellular composition of each patient. The aggregation consists of mainly two steps, aggregation of the cells of each patient weighted by the posteriors of the GM-VAE and scaling by the gene embeddings generated by TranscriptFormer for the cells of that specific patient.

2.3.1 Aggregation Step

A simple average over cell embeddings would discard information about the type of cells and its composition since two patients could have identical mean embeddings yet very different cell types and its proportion. To preserve this compositional structure, we aggregate through the GM-VAE's soft cluster assignments $\boldsymbol{\pi}_i \in \Delta^{K-1}$. First, for each mixture component $k = 1, \dots, K$, we compute a *patient-specific cluster centroid*—the average location of patient j's cells in embedding space, weighted by how strongly each cell is assigned to cluster k:

$$\boldsymbol{\mu}_{j,k} = \frac{\sum_{i \in \mathcal{P}_j} \pi_{i,k}\, \mathbf{h}_i}{\sum_{i \in \mathcal{P}_j} \pi_{i,k}}, \tag{1}$$

where $\mathcal{P}_j$ denotes the set of cells belonging to patient j and $\mathbf{h}_i \in \mathbb{R}^D$ is the cell embedding from TranscriptFormer. Second, we combine these K centroids into a single patient vector by weighting each centroid by its *relative abundance*—the fraction of the patient's total soft membership that falls in cluster k:

$$\mathbf{v}_j = \sum_{k=1}^{K} \alpha_{j,k}\, \boldsymbol{\mu}_{j,k}, \qquad \alpha_{j,k} = \frac{\sum_{i \in \mathcal{P}_j} \pi_{i,k}}{\sum_{k'=1}^{K} \sum_{i \in \mathcal{P}_j} \pi_{i,k'}}. \tag{2}$$

Here $\alpha_{j,k}$ sums to one over k and can be interpreted as the proportion of patient j's cells that belong to sub-population k (in the soft-assignment sense). The resulting $\mathbf{v}_j \in \mathbb{R}^D$ therefore encodes both *what* each sub-population looks like (via the centroids $\boldsymbol{\mu}_{j,k}$) and *how much* of each sub-population the patient has (via the proportions $\alpha_{j,k}$). Because every operation is differentiable, gradients from the patient-level contrastive loss (Eq. 9) can propagate through $\alpha_{j,k}$ and $\boldsymbol{\mu}_{j,k}$ back to the model.

2.3.2 Scaling Step

We combine the patient vector and the gene matrix via element-wise multiplication:

$$\mathbf{T}_j = \mathbf{v}_j \odot \mathbf{G} \; \in \; \mathbb{R}^{V \times D}, \tag{3}$$

where $\odot$ denotes the Hadamard product between $\mathbf{v}_j$ and the gene embeddings $\mathbf{G}$. This modulation scales each patient's aggregated cellular state by its gene's contextual embedding, producing a patient-specific view of the transcriptomic landscape. The tensor $\mathbf{T}_j$ serves as input to the disentangled projector described in Section 2.4, which maps it to gene-set and pathway enrichment scores via attention-based aggregation over biologically curated gene signatures.

2.4 The Projector Architecture

To align latent representations with established biological priors, we incorporate a hierarchical concept projector (Shen et al., 2025). While the GM-VAE encoder models the cell type distributions in patients, the projector facilitates a mapping from genes to 132 literature curated gene signatures, which we denominate by *Gene Scores*, into 43 high-level Tumor Immune Microenvironment (TIME) concepts, which we will denominate by *Pathway Scores*. These signatures capture diverse immune cell states—including T cell exhaustion, NK cell cytotoxicity, and activation—key signaling pathways distilled from bulk and single-cell RNA-seq studies.

The aggregated patient-level cell embeddings are projected onto the 132 signature space and subsequently aggregated into the 43 TIME-related concepts. To account for across-tissue heterogeneity, an auxiliary cancer-type token is integrated as a contextual feature. Through this module we get a 44-dimensional patient embedding that serves as the biologically interpretable bottleneck for the supervised contrastive loss defined in equation 9.

2.5 The Loss Function

The total objective function is a weighted combination of the generative evidence lower bound (ELBO) (Kingma & Welling, 2014), an entropy regularization term, and a supervised contrastive loss. Let $\mathbf{x} \in \mathbb{R}^D$ denote a cell embedding produced by the TranscriptFormer encoder, $\mathbf{z} \in \mathbb{R}^d$ the continuous latent variable, and $y \in \{1, \dots, K\}$ the discrete cluster assignment variable. The scalar weights w_{gauss}, w_{cat}, w_{ent}, and w_{proj} control the relative contribution of each term:

$$\mathcal{L}_{\text{total}} = \underbrace{\mathcal{L}_{\text{rec}} + w_{\text{gauss}}\mathcal{L}_{\text{gauss}} + w_{\text{cat}}\mathcal{L}_{\text{cat}} + \mathcal{L}_{\text{ent}}}_{\text{GMVAE Components}} + \underbrace{w_{\text{proj}}\mathcal{L}_{\text{cont}}}_{\text{Structural Component}} \tag{4}$$

The GM-VAE component maximizes the ELBO. Here $\text{Dec}(\mathbf{z})$ is the decoder reconstruction of $\mathbf{x}$ from the latent sample $\mathbf{z}$, $q(\mathbf{z}|\mathbf{x}, y)$ is the encoder posterior conditioned on both the input and the cluster assignment, $p(\mathbf{z}|y)$ is the cluster-specific Gaussian prior, $q(y|\mathbf{x})$ is the categorical posterior over K clusters, and $p(y) = \text{Uniform}(1, \ldots, K)$ is the categorical prior:

$$\mathcal{L}_{\text{GMVAE}} = w_{\text{rec}}\|\mathbf{x} - \text{Dec}(\mathbf{z})\|^2 + w_{\text{gauss}}D_{\text{KL}}(q(\mathbf{z}|\mathbf{x}, y) \parallel p(\mathbf{z}|y)) + w_{\text{cat}}D_{\text{KL}}(q(y|\mathbf{x}) \parallel p(y)) \tag{5}$$

To prevent posterior collapse and ensure distinct clustering, we implement a persistent entropy regularization term (Grandvalet & Bengio, 2004). We penalize the categorical entropy H when it exceeds a target threshold τ (set to 2.5, corresponding to $\sim$12 active clusters out of K), scaled by how close the current entropy is to the maximum possible entropy $\ln K$:

$$H(q(y \mid \mathbf{x})) = -\sum_{k=1}^{K} q(y{=}k \mid \mathbf{x}) \log q(y{=}k \mid \mathbf{x}) \tag{6}$$

$$\tag{7}$$

$$\mathcal{L}_{\text{ent}} = w_{\text{ent}} \cdot \max(0, H(q(y \mid \mathbf{x})) - \tau) \cdot \left(1 + \frac{H(q(y \mid \mathbf{x}))}{\ln K}\right) \tag{8}$$

To ground the latent representations in relevant patient conditions, we use a supervised contrastive loss (Khosla et al., 2020; Shen et al., 2025). While the GM-VAE models cellular distributions, a differentiable projector with an attention mechanism maps these distributions into high-level patient pathway scores. We optimize the model such that patients sharing the same disease or disease subtype are pulled together in the concept space, while patients with different diseases are pushed apart. To ensure robustness, we utilize data augmentation strategies including Gaussian noise injection, feature dropout, and random scaling to create multiple views of each patient profile. Let $I = \{1, \ldots, N\}$ index the set of all views in the batch (including augmented copies), $P(i) \subseteq I$ the set of views sharing the same disease label as view i (excluding i itself), $A(i) = I \setminus \{i\}$ all views excluding i, $s_{i,p}$ the cosine similarity between the pathway score vectors of views i and p, and τ a temperature scaling parameter:

$$\mathcal{L}_{\text{cont}} = \sum_{i \in I} \frac{-1}{|P(i)|} \sum_{p \in P(i)} \log \frac{\exp(s_{i,p}/\tau)}{\sum_{a \in A(i)} \exp(s_{i,a}/\tau)} \tag{9}$$

For each anchor view i, the loss encourages its similarity to positive pairs $p \in P(i)$ (same disease) to be high relative to all other views $a \in A(i)$. This framework allows the projector to discover human-interpretable biological pathways that serve as the most discriminative features for disease stratification.

3 Results

To assess the fidelity and clinical utility of our learned representations, we designed a two-fold evaluation strategy: a quantitative benchmarking on a disease classification task using a multi-output logistic regressor, and a qualitative assessment of biological interpretability through pathway analysis.

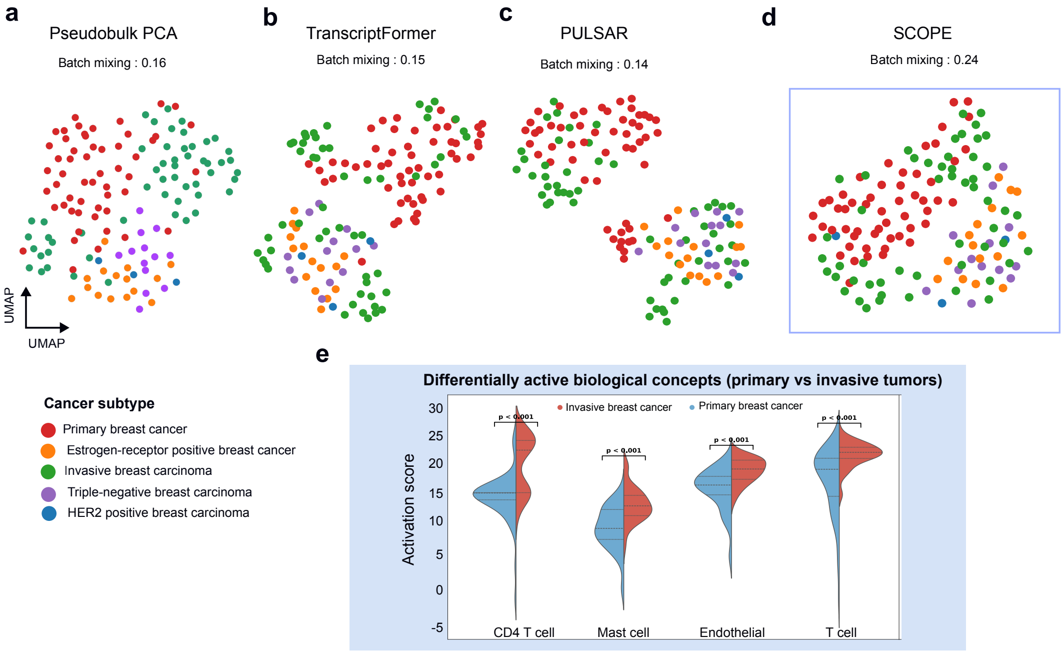

Figure 2: **Comparison of patient representations.** UMAP embeddings of patient representations generated from **(a)** PCA pseudobulked representations **(b)** pseudobulked TranscriptFormer representations **(c)** PULSAR patient representations **(d)** patient representations from our model, colored by cancer subtype. We report batch mixing using the iLISI (Luecken et al., 2022) score for assessment of integration and batch correction. The quantitative comparison of these embeddings is given in Table 1 **(e)** Differentially active concepts between primary and invasive tumors inferred by our CB-GM-VAE model.

3.1 Cancer subtype and tumor grade classification tasks

We evaluated the predictive utility of our learned representations using the scRNAseq breast cancer atlas from Chen et al. (2026) with 138 patients. Specifically we used the Immune Compartment Cells provided by CELLxGENE [1] accounting for 274,555 cells.

We benchmarked SCOPE against three distinct methodological categories: (1) Simple baselines (PCA + Pseudobulk); (2) Foundation model cell embeddings (TranscriptFormer (Pearce et al., 2025) and PULSAR (Pang et al., 2025)); and (3) VAEs or similar, including MRVI (Boyeau et al., 2025), PILOT-GM-VAE (Joodaki et al., 2025), and GLOSCOPE (Wang et al., 2024a).

The results are summarized in Table 1. For Cancer Subtype classification, our *Gene Scores* had the highest scores as the leading representation (Acc=0.68, F1=0.67), surpassing both TranscriptFormer (Acc=0.65, F1=0.66) and PULSAR (Acc=0.66, F1=0.67). This suggests that SCOPE learns biologically meaningful patient-level representations that capture subtle disease subtype distinctions and leverages the biological prior provided by the projector.

Most notably, for Tumor Grade prediction—a task requiring the detection of subtle morphological and de-differentiation signals—our *Gene Scores* representation is the top performing method (**Acc=0.63, F1=0.62**) for this classification task. This represents an improvement over the TranscriptFormer backbone and other methods relying solely on PCA representations and GM-VAEs like PILOT-GM-VAE.

While the Pathway (Concept) Scores—representing higher-level interpretable concepts—exhibit a moderate performance trade-off compared to the lower level Gene Scores, they maintain competitive results that remain superior to several black-box baselines, such as MRVI (Acc=0.64 vs. 0.49

[1] CELLxGENE Census Breast Cancer Collection: `https://cellxgene.cziscience.com/collections/9432ae97-4803-4b9f-8f64-2b41e42ad3cb`

Table 1: **Performance comparison across patient representations for cancer subtype and tumor grade prediction.** We evaluated disease classification and tumor grade prediction on the breast cancer atlas (Chen et al., 2026). For our proposed CB-GM-VAE, results are detailed for its various internal representation heads: **Gene Scores** representations are granular concepts, while **Pathway Scores** are high-level concepts[3]. Best results per target are in **bold**, and second-best results are underlined.

	Cancer Subtype		**Grade**	
Method	**Acc**	**F1**	**Acc**	**F1**
PULSAR	0.66 ± 0.06	**0.67** ± 0.06	0.48 ± 0.11	0.50 ± 0.11
PCA+Pseudobulk	0.64 ± 0.09	0.63 ± 0.09	0.52 ± 0.09	0.53 ± 0.09
PILOTGMVAE (WD)	0.59 ± 0.06	0.61 ± 0.05	0.51 ± 0.13	0.54 ± 0.11
GLOSCOPE	0.52 ± 0.11	0.54 ± 0.11	0.51 ± 0.08	0.55 ± 0.05
PILOTGMVAE	0.49 ± 0.09	0.49 ± 0.09	0.51 ± 0.11	0.53 ± 0.11
MRVI	0.49 ± 0.05	0.48 ± 0.05	0.50 ± 0.09	0.50 ± 0.06
TranscriptFormer	0.65 ± 0.05	0.66 ± 0.06	0.44 ± 0.14	0.45 ± 0.15
SCOPE (CB-GM-VAE)				
Patient Embeddings (before Projector)	0.34 ± 0.09	0.32 ± 0.11	0.30 ± 0.08	0.26 ± 0.11
Gene Scores	**0.68** ± 0.06	**0.67** ± 0.06	**0.63** ± 0.16	**0.62** ±0.14
Pathway (Concept) Scores	0.64 ± 0.08	0.64 ± 0.09	0.55 ± 0.14	0.55 ± 0.15

for Subtype). Crucially, these scores provide a transparent link between latent patient representations and human-interpretable oncologic pathways, effectively bridging the gap between high-dimensional embeddings and clinical interpretability.

3.2 Analysis of Concepts

Beyond classification accuracy, a key contribution of our model is the interpretable projector of the learned representations. By comparing concept activation scores between primary breast cancer and invasive carcinoma patients, we identified several statistically significant pathways and in 2 we show the four most significant differential biological concepts: $\mathrm{CD4}^+$ T cell , Mast cell, Endothelial cell and T cell general (FDR-corrected $p < 0.05$).

The enrichment of endothelial cell and T cell-related signatures in invasive tumor samples are corroborated by the established role of the TIME in driving metastasis via angiogenesis and immune evasion. Tumor-associated endothelial cells form aberrant, leaky vascular networks that sustain tumor growth, enable VEGF-driven vascular remodeling, and provide permissive routes for intravasation and distant colonization (Shipp et al., 2024; Riera-Domingo et al., 2023; Elayat & Selim, 2024; Yao & Zeng, 2023). Simultaneously, the T cell signature reflects massive lymphocytic infiltration characteristic of the 'inflamed' invasive stroma—a hallmark of aggressive breast cancer that correlates with poor prognosis (Binnewies et al., 2018).

Interestingly, the $\mathrm{CD4}^+$ T cell concept showed high significance, reflecting the critical functional plasticity of helper T cells in the TIME. During primary-to-invasive transition, CD4+ infiltration surges, often shifting from anti-tumor Th1 to pro-invasive Th2/Treg phenotypes that drive ECM remodeling, angiogenesis, and immune suppression (Halim et al., 2024). Similarly, Mast cell activation emerged as a key differentiator. Mast cells are known to accumulate at the invasive front of breast tumors, where they release pro-angiogenic factors(e.g., VEGF, bFGF) and matrix metalloproteinases (MMPs) that degrade the basement membrane, directly facilitating local invasion (Komi & Redegeld, 2020).

[3] `https://www.immuno-compass.com/help/index.html#concepts`, accessed February 11, 2026.

4 Conclusion

We demonstrated that the latent space of single cell foundation models contains linear subspaces that are human-interpretable and disease relevant. We developed SCOPE, a model that extracts interpretable patient representations from scFMs, an architecture that is inspired by modeling patient similarities based on cell type composition distribution and biological concepts. This approach stands in sharp contrast to existing patient representation learning models that attempt to bridge the single-cell- and patient-level scales through learning scale-specific representations (Pang et al., 2025) as opposed to generative modeling of the scales.

Using a single-cell breast cancer atlas and the *TranscriptFormer* foundation model continually pretrained on pan-cancer data, we demonstrated that the patient representations constructed from granular concepts inferred by CB-GM-VAE achieved state-of-the-art performance in Cancer Subtype prediction (Acc=0.68) and Tumor Grade prediction (Acc=0.63), outperforming established baselines and frozen foundation model backbones. In addition, our model successfully identified critical biological drivers of malignant progression, specifically highlighting the differential activity of Endothelial cells, Mast cells, and CD4$^+$ T cell functional shifts as patients transition from primary to invasive disease. Analyzing the up-regulation and down-regulation patterns of pathway nodes across patient cohorts allows us to determine whether the model captures established biological mechanisms, such as immune exhaustion or proliferative signaling, rather than relying on spurious correlations. This analysis confirms that the latent dimensions semantically align with known oncogenic pathways, bridging the gap between predictive performance and human interpretability. The identified concepts not only reflect canonical hallmarks of cancer progression but also highlight specific immune and stromal programs that may serve as potential therapeutic targets or biomarkers for invasive transition.

Despite these promising results, significant avenues for optimization remain. Due to hardware and time constraints, SCOPE was trained for only a single epoch; we anticipate that extended training schedules and the inclusion of an even larger, more heterogeneous pan-cancer pre-training corpus will further refine the sensitivity of the concept bottleneck.

There is currently significant interest in connecting biological scales to develop virtual patient models capable of forecasting disease progression and treatment response. The scarcity of disease- and context-specific patient profiles at the single-cell level, together with the high-dimensional and sparse nature of single-cell measurements, makes the zero-shot capabilities of foundation models attractive for learning patient representations. Furthermore, diversity of training data used in foundation models motivates automatic extraction of biological concepts from the internal representations of these models to enhance the interpretability of patient representations. Our results suggest that there is an opportunity in single cell foundation models for development of interpretable virtual patient models that would excel in their predictions.

References

Emily Alsentzer, Michelle M. Li, Shilpa N. Kobren, Ayush Noori, Undiagnosed Diseases Network, Isaac S. Kohane, and Marinka Zitnik. Few-shot learning for phenotype-driven diagnosis of patients with rare genetic diseases. *npj Digital Medicine*, 8(1):380, 2025. doi: 10.1038/s41746-025-01749-1. URL https://doi.org/10.1038/s41746-025-01749-1.

Ayla Bassez, Hanne Vos, Leen Van Dyck, et al. A single-cell map of intratumoral changes during anti-pd1 treatment of patients with breast cancer. *Nature Medicine*, 27(5):820–832, 2021.

M. Binnewies et al. Understanding the tumor immune microenvironment (time) for effective immunotherapy. *Nature Medicine*, 24(5):541–550, 2018.

Pierre Boyeau et al. Multi-resolution variational inference for single-cell transcriptomics. *Nature Methods*, 2025. doi: 10.1038/s41592-025-02808-x.

Chan Zuckerberg Initiative Single-Cell Biology System. The cellxgene discover census provides a cloud-native, high-performance, and standard-compliant access layer for the single-cell data corpus. *bioRxiv*, 2023. doi: 10.1101/2023.10.30.563174. URL https://www.biorxiv.org/content/10.1101/2023.10.30.563174v1.

Andrew Chen, Lina Kroehling, Christina S Ennis, Gerald V Denis, and Stefano Monti. A highly resolved integrated single-cell atlas of human breast cancers. *NAR Genomics and Bioinformatics*, 8(1):lqaf217, 2026. doi: 10.1093/nargab/lqaf217. URL `https://doi.org/10.1093/nargab/lqaf217`.

Haotian Cui, Chloe Wang, Hassal Lee Maan, Kuan Pang, Fengning Luo, Nan Duan, and Bo Wang. scgpt: toward building a foundation model for single-cell multi-omics using generative ai. *Nature Methods*, 21(7):1470–1480, 2024.

Ghada Elayat and Abdel Selim. Angiogenesis in breast cancer: insights and innovations. *Clinical and Experimental Medicine*, 24(1):178, 2024. doi: 10.1007/s10238-024-01446-5. URL `https://link.springer.com/article/10.1007/s10238-024-01446-5`. Open access, 4095 accesses, 21 citations.

Ruli Gao, Shanshan Bai, Y Henderson, et al. Delineating copy number and clonal substructure in human tumors from single-cell transcriptomes. *Nature Biotechnology*, 39:599–608, 2021.

Yves Grandvalet and Yoshua Bengio. Semi-supervised learning by entropy minimization. In *Advances in neural information processing systems*, 2004.

L. Halim et al. Cd4+ t cell plasticity in the breast tumor microenvironment: From surveillance to invasion. *Frontiers in Immunology*, 15:1023–1041, 2024.

Ya Han, Yuting Wang, Xin Dong, Dongqing Sun, Zhaoyang Liu, Jiali Yue, Haiyun Wang, Taiwen Li, and Chenfei Wang. Tisch2: expanded datasets and new tools for single-cell transcriptome analyses of the tumor microenvironment. *Nucleic Acids Research*, 51(D1):D1425–D1431, 2023. doi: 10.1093/nar/gkac959. URL `https://doi.org/10.1093/nar/gkac959`.

Lukas Heumos, Anna C Schaar, Christopher Lance, Almut Litinetskaya, Felix Drost, Luke Zappia, Malte D Lücken, Daniel C Strobl, Juan Henao, Francesca Curion, et al. Best practices for single-cell analysis across modalities. *Nature Reviews Genetics*, 24(8):550–572, 2023. doi: 10.1038/s41576-023-00586-w.

Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. In *International Conference on Learning Representations*, 2017.

Longda Jiang, Zhixin Cyrillus Tan, Isabella N. Grabski, Yuhan Hao, Nathan Nakatsuka, Sourav Sarkar, Anagha Shenoy, and Rahul Satija. Reconstructing developmental and disease progression with sample-level embeddings. *bioRxiv*, 2025. doi: 10.64898/2025.12.10.693462. URL `https://doi.org/10.64898/2025.12.10.693462`.

Mehdi Joodaki, Mina Shaigan, Samaneh Samiei, James Nagai, Tiago Maié, Christoph Kuppe, and Ivan G Costa. Pilot-gm-vae: patient-level analysis of single-cell disease atlas with optimal transport of gaussian mixture variational autoencoders. *Briefings in Bioinformatics*, 26(5):bbaf547, 2025. doi: 10.1093/bib/bbaf547. URL `https://academic.oup.com/bib/article/26/5/bbaf547/8287234`.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in Neural Information Processing Systems*, 33:18661–18673, 2020.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In *2nd International Conference on Learning Representations (ICLR)*, 2014. URL `https://arxiv.org/abs/1312.6114`.

D. E. G. Komi and F. A. Redegeld. The role of mast cells in neoplasia and angiogenesis. *Clinical Reviews in Allergy & Immunology*, 58(1):52–67, 2020.

Michelle M. Li, Ben Y. Reis, Adam Rodman, Tianxi Cai, Noa Dagan, Ran D. Balicer, Joseph Loscalzo, Isaac S. Kohane, and Marinka Zitnik. Scaling medical ai across clinical contexts. *Nature Medicine*, Feb 2026. doi: 10.1038/s41591-025-04184-7. URL `https://doi.org/10.1038/s41591-025-04184-7`.

Anastasia Litinetskaya, Maiia Shulman, Soroor Hediyeh-zadeh, Amir Ali Moinfar, Fabiola Curion, Artur Szałata, Alireza Omidi, Mohammad Lotfollahi, and Fabian J. Theis. Multimodal weakly supervised learning to identify disease-specific changes in single-cell atlases. *bioRxiv*, 2024. doi: 10.1101/2024.07.29.605625. URL https://doi.org/10.1101/2024.07.29.605625.

Tianyu Liu, Edward De Brouwer, Tony Kuo, Nathaniel Diamant, Alsu Missarova, Hanchen Wang, Minsheng Hao, Hector Corrada Bravo, Gabriele Scalia, Aviv Regev, and Graham Heimberg. Learning multi-cellular representations of single-cell transcriptomics data enables characterization of patient-level disease states. *bioRxiv*, 2024. doi: 10.1101/2024.11.18.624166. URL https://doi.org/10.1101/2024.11.18.624166.

Yang Liu et al. Combined single-cell and spatial transcriptomics reveal the metabolic evolvement of breast cancer during early dissemination. *Advanced Science*, 10:2205367, 2023.

Malte D. Luecken, Maren Büttner, Kridsadakorn Chaichoompu, Anna Danese, Marta Interlandi, Michaela F. Mueller, Daniel C. Strobl, Luke Zappia, Martin Dugas, Maria Colomé-Tatché, and Fabian J. Theis. Benchmarking atlas-level data integration in single-cell genomics. *Nature Methods*, 19(1):41–50, 2022. doi: 10.1038/s41592-021-01336-8. URL https://doi.org/10.1038/s41592-021-01336-8.

Chris J Maddison, Dieter Andriyash, and Yee Whye Teh. The concrete distribution: A continuous relaxation of discrete random variables. In *International Conference on Learning Representations*, 2017.

Alexander J Oliver, Ni Huang, Rebeca Bartolome-Casado, Manu Setty, S Saveljeva, J Knight, K R James, V Cardenas, A H Jonsson, et al. Single-cell integration reveals metaplasia in inflammatory gut diseases. *Nature*, 635(8040):699–707, 2024. doi: 10.1038/s41586-024-07571-1. URL https://doi.org/10.1038/s41586-024-07571-1.

Bhupinder Pal, Yunshun Chen, Francois Vaillant, et al. A single-cell rna expression atlas of normal, preneoplastic and tumorigenic states in the human breast. *The EMBO Journal*, 40(11):e107333, 2021.

Kuan Pang, Yanay Rosen, Kasia Kedzierska, Ziyuan He, Abhejit Rajagopal, Claire E. Gustafson, Grace Huynh, and Jure Leskovec. Pulsar: a foundation model for multi-scale and multicellular biology. *bioRxiv*, 2025. doi: 10.1101/2025.11.24.685470. URL https://www.biorxiv.org/content/10.1101/2025.11.24.685470v1.

James D Pearce, Sara E Simmonds, Gita Mahmoudabadi, Lakshmi Krishnan, Giovanni Palla, Ana-Maria Istrate, Alexander Tarashansky, Benjamin Nelson, Omar Valenzuela, Donghui Li, Stephen R Quake, and Theofanis Karaletsos. A cross-species generative cell atlas across 1.5 billion years of evolution: The transcriptformer single-cell model. *bioRxiv*, 2025. doi: 10.1101/2025.04.25.650731. URL https://www.biorxiv.org/content/10.1101/2025.04.25.650731v1.

Junbin Qian, Sarah Olbrecht, Bram Boeckx, et al. A pan-cancer blueprint of the heterogeneous tumor microenvironment revealed by single-cell profiling. *Cell Research*, 30(9):745–762, 2020.

Carla Riera-Domingo, Eduarda Leite-Gomes, Iris Charatsidou, Peihua Zhao, others, and Massimiliano Mazzone. Breast tumors interfere with endothelial trail at the premetastatic niche to promote cancer cell seeding. *Science Advances*, 9(12), 2023. doi: 10.1126/sciadv.add5028. Published 22 March 2023.

Wanxiang Shen, Thinh H Nguyen, Michelle M Li, Yepeng Huang, Intae Moon, Nitya Nair, Daniel Marbach, and Marinka Zitnik. Generalizable ai predicts immunotherapy outcomes across cancers and treatments. *medRxiv*, 2025. doi: 10.1101/2025.05.01.25326820. URL https://www.medrxiv.org/content/10.1101/2025.05.01.25326820v2.

C. Shipp et al. The role of endothelial cell heterogeneity in breast cancer neovascularization and metastasis. *Nature Reviews Cancer*, 24(2):89–105, 2024.

Artem Shmatko, Alexander Wolfgang Jung, Kumar Gaurav, Søren Brunak, Laust Hvas Mortensen, Ewan Birney, Tom Fitzgerald, and Moritz Gerstung. Learning the natural history of human disease with generative transformers. *Nature*, 647(8088):248–256, 2025. doi: 10.1038/s41586-025-09529-3. URL https://doi.org/10.1038/s41586-025-09529-3.

Christina V Theodoris, Ling Xiao, Anant Chopra, Mark D Chaffin, Ziad R Al Sayed, Matthew C Hill, Helene Mantineo, Elizabeth Brydon, Zexi Zeng, X Shirley Liu, and Patrick T Ellinor. Transfer learning enables predictions in network biology. *Nature*, 618(7965):616–624, 2023. doi: 10.1038/s41586-023-06139-9.

Sandra Tietscher, Johannes Wagner, Tobias Anzeneder, et al. A comprehensive single-cell atlas of the human breast tumor microenvironment. *Nature Communications*, 14(1):98, 2023.

Hao Wang, William Torous, Boying Gong, and Elizabeth Purdom. Visualizing scrna-seq data at population scale with gloscope. *Genome Biology*, 25(1):259, 2024a. doi: 10.1186/s13059-024-03398-1. URL https://genomebiology.biomedcentral.com/articles/10.1186/s13059-024-03398-1.

S Wang et al. Single-cell transcriptomics reveals the role of antigen presentation in liver metastatic breast cancer. *Nature Communications*, 2024b.

Sunny Z Wu, Ghamdan Al-Eryani, Daniel L Roden, et al. A single-cell and spatially resolved atlas of human breast cancers. *Nature Genetics*, 53(9):1334–1347, 2021.

Xinghong Yao and Ye Zeng. Tumour associated endothelial cells: origin, characteristics and role in metastasis and anti-angiogenic resistance. *Frontiers in Physiology*, 14:1199225, 2023. doi: 10.3389/fphys.2023.1199225. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC10301839/. Review article.

Seyhan Yazar, Jose Alquicira-Hernández, Kemp Wing, Anne Senabouth, Matthew G Gordon, Sanda Andersen, Qinyi Lu, Alex Rowson, Thomas R Taylor, Linda Clarke, et al. Single-cell eqtl mapping identifies cell type–specific genetic control of autoimmune disease. *Science*, 376(6589): eabf3041, 2022. doi: 10.1126/science.abf3041.

A Appendix

A.1 Datasets

The datasets compiled and aggregated for our study were sourced from diverse, high-quality repositories to ensure broad biological coverage. We utilized the TISCH2 database (Han et al., 2023), an invaluable resource for studying the tumor immune microenvironment as it unifies data from nearly 200 cancer datasets under a standardized hierarchical annotation framework, resolving cross-study labeling inconsistencies to enable robust, pan-cancer analysis of fine-grained immune and stromal cell states. To capture healthy baseline variability, we included OneK1K (Yazar et al., 2022), a large-scale single-cell RNA sequencing cohort profiling approximately 1.27 million peripheral blood mononuclear cells (PBMCs) from 982 donors, serving as a robust reference for characterizing inter-individual variability and genetic regulation across human immune cell types.

Additional disease-specific cohorts were sourced via the CELLxGENE Census (Chan Zuckerberg Initiative Single-Cell Biology System, 2023), including a study linking chronic inflammation to cancer (Oliver et al., 2024), which identifies a specific metaplastic stem-cell state shared between Inflammatory Bowel Disease (IBD) and Colorectal Cancer (CRC), suggesting a common inflammation-driven precursor.

Finally, we used multiple cancer datasets (Wu et al., 2021; Pal et al., 2021; Qian et al., 2020; Gao et al., 2021; Tietscher et al., 2023; Bassez et al., 2021; Liu et al., 2023; Wang et al., 2024b). This collection was selected to provide the necessary variance for learning disentangled disease concepts: it spans the full disease trajectory from pre-neoplastic states (Pal et al., 2021) and clonal tumor heterogeneity (Gao et al., 2021) to metastatic dissemination (Liu et al., 2023; Wang et al., 2024b), while simultaneously offering deep profiling of the immune microenvironment (Wu et al., 2021; Qian et al., 2020; Tietscher et al., 2023) and longitudinal responses to anti-PD1 immunotherapy (Bassez et al., 2021).

A.2 TranscriptFormer Loss

The model is trained using two distinct loss functions corresponding to the two decoder heads: a cross-entropy loss for gene prediction and a zero-truncated Poisson negative log-likelihood (NLL) loss for count prediction.

A.2.1 Gene Decoder Loss

The gene decoder predicts a probability distribution over the gene vocabulary at each sequence position. The standard cross-entropy loss is used to supervise the gene predictions. Given the predicted gene probabilities π_j at position j, the gene loss is given by:

$$\mathcal{L}_{\text{gene}} = -\sum_{j=1}^{M} \log \pi_j \tag{10}$$

This loss encourages the model to assign high probability to the correct gene identity at each position.

A.2.2 Count Decoder Loss

The count decoder predicts a set of expected expression counts per sequence position. Since gene expression counts follow a discrete, non-negative distribution, we model them using a Poisson likelihood. However, because counts of zero are excluded from the input sequence, we employ a zero-truncated Poisson negative log-likelihood (NLL) loss.

Given the predicted rate parameter λ_j for position j and the observed ground truth count c_j, the loss is computed as:

$$\mathcal{L}_{\text{count}} = -\sum_{j=1}^{M} \left(c_j \log \lambda_j - \lambda_j - \log(1 - e^{-\lambda_j}) \right) \tag{11}$$

The additional term $\log(1 - e^{-\lambda_j})$ accounts for the truncation of zero values, ensuring proper normalization.

A.2.3 Total TranscriptFormer Loss

The final loss function is the sum over gene and count losses:

$$\mathcal{L} = \mathcal{L}_{\text{gene}} + \mathcal{L}_{\text{count}} \tag{12}$$

A.3 Training details

We train the full model end-to-end on a single NVIDIA A100 80 GB GPU. Data is streamed out-of-core from a Zarr store via AnnBatch (unpublished work) to avoid loading the full dataset into memory. We use an effective batch size of 512 cells (micro-batch of 256 with 2 gradient accumulation steps). Training uses bfloat16 mixed precision and `torch.compile()` to enable fused FlexAttention kernels. The Gumbel-Softmax temperature is initialised at 1.0. Due to the computational cost of the full-attention TranscriptFormer operating over long token sequences, we restrict input to the 1,000 most highly expressed genes per cell rather than the 2,000 recommended by single-cell best practices (Heumos et al., 2023); as this reduces both peak memory. Under this configuration, a single training epoch required approximately 17 hours, and resource constraints limited the current results to one epoch.

B Benchmarking parameters

B.1 CB-GM-VAE Training Parameters

Table 2 summarises the hyperparameters used for training the CB-GM-VAE model. The model was trained end-to-end (no LoRA; all 12 TranscriptFormer layers unfrozen) on a single NVIDIA A100 80 GB GPU using bfloat16 mixed precision and `torch.compile()` for fused FlexAttention kernels.

Table 2: CB-GM-VAE training hyperparameters.

Parameter	**Value**
Architecture	
Backbone	TranscriptFormer (tf-sapiens)
Backbone embedding dim (D)	2048
Backbone layers	12
Backbone attention heads	16
GMM-VAE components (K)	50
Gaussian latent dim (d)	64
Max genes per cell (S)	1000
Gene count clipping	30
Optimisation	
Optimiser	AdamW (fused)
Backbone learning rate	1×10^{-4}
GM-VAE & Projector learning rate	1×10^{-3}
LR scheduler	CosineAnnealingLR
Effective batch size	512 (256 × 2 accumulation steps)
Mixed precision	bfloat16
Gradient clipping	max norm 1.0
Loss weights	
w_{rec} (reconstruction)	1.0
w_{gauss} (Gaussian KL)	1.0
w_{cat} (categorical KL)	1.0
w_{ent} (entropy regularisation)	1.0
w_{proj} (contrastive)	1.0
Reconstruction type	MSE
Entropy target τ	2.5
Gumbel-Softmax	
Initial temperature	1.0
Minimum temperature	0.5
Temperature decay rate	0.00693
KL warmup	
Warmup epochs	3
Minimum KL weight	0.1
Schedule	Linear ramp

Patient Representations

Patient-level representations were derived from three model-based methods—GloScope, PILOT-GMVAE, and MrVI and Pseudobulk PCA. For GloScope, the K-nearest neighbor (KNN) divergence was computed using the default parameters. For PILOT-GMVAE, two types of patient representations were evaluated: (1) the Wasserstein distance–based representation, computed following the procedure described in the original publication, and (2) the weight-based representation, defined as the proportion of component-specific cell types within each sample. For MrVI, local sample representations were first obtained at the single-cell level and averaged across all cells belonging to the same sample to yield a single, sample-level latent representation. To generate dimensionality reduction-based patient embeddings, we applied Pseudobulk PCA. Principal Component Analysis (PCA) was performed on the raw gene expression matrix using 2,000 highly variable genes (HVGs) selected per batch, with the batch identifier used as the grouping key. The PCA model was fit on the training subset of each cross-validation fold and applied to both training and test sets to prevent data leakage. For pseudobulk aggregation, cell-level PCA values corresponding to each donor were summed across all cells belonging to the same donor, producing a single pseudobulk-level representation per sample. These aggregated PCA features served as patient-level representations for downstream classification.

For TranscriptFormer (zero-shot), cell embeddings were obtained using the official `TranscriptFormer inference` CLI with the pretrained tf-sapiens checkpoint. Inference was run with a batch size of 4096, and duplicate genes removed. Cell embeddings were then aggregated to the donor level by summing all cell vectors belonging to the same donor (pseudobulk).

For UCE + PULSAR, cell-level embeddings were first generated using Universal Cell Embeddings (UCE) with default 4 layer model weights, a batch size of 100, and the species set to human. These cell-level UCE embeddings (`X_uce`) were then passed to PULSAR (`pulsar-pbmc` from HuggingFace) to obtain 512-dimensional donor-level representations. All available cells per donor were used without subsampling (`replace=False`), and the model was loaded in bfloat16 precision.

For CB-GM-VAE (ours), the model was trained end-to-end on a single NVIDIA A100 80 GB GPU using bfloat16 mixed precision and `torch.compile()` for fused FlexAttention kernels. The backbone consisted of the full 12-layer TranscriptFormer (tf-sapiens) encoder with all layers unfrozen and a learning rate of 1×10^{-4}, while the GMM-VAE head and differentiable projector were trained with a learning rate of 1×10^{-3}. Both parameter groups were optimised jointly using fused AdamW with a cosine annealing schedule and gradient clipping at max norm 1.0. The effective batch size was 512, achieved through a micro-batch size of 256 with 2 gradient accumulation steps. The GMM-VAE head modelled K=50 Gaussian mixture components with a 64-dimensional latent space, using Gumbel-Softmax sampling with an initial temperature of 1.0, decayed to a minimum of 0.5 at a rate of 0.00693 per step. KL divergence terms were warmed up linearly over the first 3 epochs from a minimum weight of 0.1. All loss terms—reconstruction (MSE), Gaussian KL, categorical KL, entropy regularization (target τ=2.5), and supervised contrastive projection—were weighted equally at 1.0. Training ran for up to 10 epochs with early stopping (patience of 5 epochs). The data was split at the donor level into 80% training and 20% test sets, stratified by disease label, and streamed out-of-core from zarr format using AnnBatch (unpublished).

Model Training and Evaluation

A logistic regression classifier was used to benchmark the performance of each set of patient representations. The model was implemented in the `scikit-learn` library (`LogisticRegression`) with parameters `max_iter=2000`, `class_weight='balanced'`, and regularization strength `C=0.01`. Model performance was assessed using five-fold stratified cross-validation to preserve class balance in each fold. Within each fold, which was fit on the training data and subsequently applied to both training and test sets.